Lip Synchronization Model for Sinhala Language Using Machine Learning

A dissertation submitted for the Degree of Master of Science in Computer Science

P.D.C. Ranaweera
University of Colombo School of Computing
2023

DECLARATION

I hereby declare that the thesis is my original work, and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Student Name: P. Dilani Chamarika Ranaweera

Registration Number: 2018/MCS/073

Index Number: 18440733

____

Signature of the Student & Date

This is to certify that this thesis is based on the work of Mr. /Ms. ________________________ under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by,

Supervisor Name: Dr. A.R. Weerasinghe

____

Signature of the Supervisor & Date

I would like to dedicate this thesis to my family and all my friends whose inspiration have been the motivation behind my academic successes.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my research supervisor, Dr. A.R. Weerasinghe, senior lecturer of University of Colombo School of Computing and my research co-supervisor, Dr. D.M.R. Dinalankara, Head of the Department of Computer Engineering, Faculty of Engineering, University of Sri Jayewardenepura for providing me with the ongoing guidance and supervision throughout the research.

Additionally, I would want to express my deep gratitude to every one of the panelists who participated in the interim defense and proposal defense for their insightful criticisms of my research, which helped me to strengthen it.

This thesis is also dedicated to my devoted husband, who has been a tremendous help to me during this research project by offering guidance on technical matters. I also take this chance to thank my friends for their encouragement and criticism while I worked toward my research objectives.

Last but not least, it gives me great pleasure to thank everyone who supported me, both directly and indirectly, to successfully complete my research.

ABSTRACT

Currently, a lot of nations produce cartoon characters for a variety of purposes, including the animation film industry, the gaming industry, live broadcast television programs, etc. These characters are made available so that users can interact more with the films, video games, or television shows. For such cartoon figures to appear more alive while speaking a language, lip synchronization is crucial. Lip synchronization is the process of synchronizing speech to a synthetic facial model's lip movement. To create realistic lip-synchronization animation, the voice and lip motions in this procedure must be appropriately timed. Building a talking face utilizing various methods for languages including English, Korean, and Portuguese has been the subject of numerous studies. Compared to other languages, Sinhala has less resources due to less contribution in the researches. The interaction between the synthetic mouth and the Sinhala sounds will be especially interesting to observe. This model can be used to create cartoon characters that speak Sinhala smoothly instead of opening and closing their mouths a lot.

The most difficult challenge is to match the "phonemes," which are the fundamental sounds formed in any language, with the "visemes," a visual representation of lip movement. There are three main methods for lip synchronization: the static viseme approach, which uses the viseme alphabet to derive the language's phonemes, the dynamic approach, which employs visual cues from speech in real time, and the deep learning technique, which makes use of a vast visual data set. Because the letters in the Sinhala language indicate the language's phonemes, the viseme classification in this study is based on a variety of letter pairings. Overall, 23 viseme classes have been found. Finally, a deep learning model was produced utilizing a multiclass classification method.

In the final system implementation, text input is provided first, after which the system will produce audio and the deep learning model will produce a collection of visemes based on the provided text. The system interface then offers three options for playing the vesmes at various speeds, including rapid, normal, and slow. The user interface was created in Python, and the deep learning model was integrated into the system. The deep learning model for the viseme classification is created using Google collabs. This model will be very helpful in the future when the Sinhala alphabet gets a new character. This approach can also be used to train deaf persons to read lips.

Keywords: lip synchronization, Sinhala, Static viseme, multi class classification

LIST OF PUBLICATIONS

Ranaweera P.D.C, Gunasekara R.P.T.H , 2014 , Investigation of Improving user Interaction on Websites, 1st Wayamba International Conference,Sri Lanka, ISBN 978-955-4709-18-8 pp 39.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 01 INTRODUCTION

This chapter describes about what is a lip synchronization and its background in the field of natural language processing. Furthermore, this gives explanations for the motivation to the research work to develop lip synchronization model to a native language used in Sri Lanka, problem statement, aims, objectives and the scope of the study. The chapter also concludes the structure of the thesis at last.

1.1 Background to the research

Nowadays, a lot of other countries produce cartoon characters for a variety of purposes, including the live-action and animated film and television industries as well as the gaming sector. These characters are made available so that users can interact more with the films, video games, or television shows. One of the key elements that creates a powerful engagement between the user and the cartoon character is communication.

The speech signals, which are bimodal channels that incorporate both audio and visual representation, make up a significant portion of human communication. (Bear and Harvey, 2019). Because it gives the user a realistic experience, visual speech animation, which matches speech with lip movements, has a significant impact on the gaming and animation film industries. This method is often referred to as lip synchronization. (Weerathunga et al., 2020).

Several studies have been done to build a talking face using different approaches for other languages such as English, Korean, Portuguese, etc.(Bear and Harvey, 2019) . As a results of different researches done based on different languages for visual speech animation, cartoon characters are talking different languages in a smooth manner by showing proper synchronization with the audio and lip movements.

Lip synchronization is the process of synchronizing speech to a synthetic facial model's lip movement. To create realistic lip-synchronization animation, the voice and lip motions in this procedure must be appropriately timed.

Currently, there are two major types of developments available as a development of lip synchronization model.

- The first one is creating a talking artificial face or mouth for a specific language. For instance, Google has integrated a 2D talking face into its browser to teach users how to pronounce a word at various speeds using British and American pronunciations. These artificial talking faces are used by the cartoon and animation industries to create animated films.

- The second form involves creating talking mouths for characters in real life. As an illustration, Prajwal K R created a face-to-face translation for Hindi to English using five well-known actors (Chris Anderson, Andrew Ng, Obama, Modi, and Elon Musk). This can translate a video of a person speaking in English into target language Hindi with realistic lip synchronization (K R et al., 2019). This type of implementations are widely used in movie dubbing, educational videos and television news and interviews.

In addition, some researchers have been developed a model for real time synchronization with cartoon character which appeared in live broadcasting of a television program. Deepali Aneja and Wilmot Li implemented an interactive system which automatically generates live lip synchronization for 2D character by taking streaming audio as input and producing visual representation (Aneja and Li, 2019) . To produce synchronized face motions with audio produced by natural speech or a Text to Speech tool, Yuyu Xu and his colleagues showed a lip animation algorithm for real-time applications(Xu et al., 2013).

Lip reading is a technique which is widely used by hearing impaired people to avoid social isolation while communicating with others. Lip synchronization technique uses to practice lip reading for hearing impaired people to understanding lectures or public speeches and some detective purposes like reading lips in noisy environments etc. Joon Son Chung and Andrew Zisserman has done a research to find an answer to read lip in profile in a same standard (Son and Zisserman, 2017).

Furthermore, another system developed by D.Ivanko , D,Ryumin and A. karpov to solve the problem of inability to use speech interfaces for deaf and hearing impaired people who has the limitation for human machine interaction. They suggested using lip-reading in addition to hand gesture identification to boost the precision and dependability of automatic Russian sign language recognition. They proposed to use lip-reading in addition to hand gestures recognition to increase the accuracy and reliability of the automatic Russian sign language recognition (Ivanko et al., 2019).

1.2 Motivation

While considering background of the research study, many researches have been presented lip synchronized models to several applications using different technologies for many languages such as English, Korean, Portuguese, Arabic, Indonesian, Chinese and Hindi etc..

Comparatively Sinhala language is a low resourced native language due to less contribution in the researches (Weerathunga et al., 2020) . It would be of special interest to see how the synthetic mouth will act along with the Sinhala sounds. This model can be used for build more realistic cartoon characters which speaks Sinhala in a smooth manner rather than frequently open and close the mouth.

However, with the aid of technology and computing power, visualizing speech with lip synchronization is playing a significant role in a number of areas. This will pave the way for new developments in the fields of education, transportation, cognitive education, the entertainment industry, and the film industry, among others (Weerathunga et al., 2020). Therefore, each field requires a new strategy to improve user interaction.

For this study, it is of interest to build a visual speech animation which can synchronize lip movements with the sounds in Sinhala Language that can apply for different fields. This model can be used as a solution for deaf people to understand the audio or text by converting it into the visual representation. Also, interactive learning features can be added to virtual assistants or virtual tutors in e learning and web navigation applications in education field.

1.3 Statement of the problem

Developing a talking synthetic face model for a particular language is not new to the other foreign languages like English, Korean, Portuguese, and Chinese etc. When considering Sinhala language, development based on artificial Sinhala talking face is uncommon due to low resources and less contribution to researches.

For lip synchronization with other languages, a variety of alternative methods have been developed over the past few decades, including the static viseme technique, deep learning approach, and dynamic approach, among others.

Without utilizing machine learning or deep learning concepts, Chashika Weerathunga's research has presented a lip synchronization model for the Sinhala language based on a static viseme approach. That implemented model performs admirably for single words and short sentences, but it keeps failing for long sentences with multiple transitions and sentences spoken at various speeds.(Weerathunga et al., 2020). In this research there is a problem that not yet solved using static viseme approach is the developed model was not well synchronized with long Sinhala sentences. Therefore, find a solution to improve the synchronization for long Sinhala sentences are having higher value in a different approach and machine learning has been selected to solve the arisen problem.

The following are the research problems going to be addressed through the research study.

i. What is an appropriate way to synchronize lip with text input given in different lengths in machine learning approach?
   - The research study going to find an answer for appropriately synchronize the lip movements with audio when the text input is given. The input text may be in different lengths and the developing model should be properly matched the lip movements even though the input text length is vary.

ii. How to create training dataset by mapping phoneme to viseme?
   - Phonemes are the smallest unit of speech sounds in a language and it can be generated by analyzing text or audio. Visemes are the visual representation of phoneme. Phonemes and visemes having one to many association. Each phoneme has only one viseme representation, but each viseme has multiple phonemes attached (Serra et al., 2012). The way of mapping between phonemes and viseme is most important part of a lip synchronization. For instance, phonemes /m/, /p/, /b/ are generating same viseme representation in Sinhala language.

iii. What is the latency that can synchronize the synthetic mouth according to the phoneme generated to the given input?
   - Every phoneme has a time duration, starting and ending of each phoneme is affecting to the synchronization of the face model. Once the phoneme started, viseme related to the selected phoneme should be start animating and when the phoneme is finishing, viseme also should be finished. Each sentence input to the system will extract the phonemes attached with it and set of phonemes are feeding in to the model to find the set of visemes associates with each phonemes.
   - Correct frame rate should need to be identify and apply to have a proper synchronization with the audio and the visual components.

- Space should need to be properly handle in the text input that is generating silent mode of viseme in the animation.

1.4 Research Aims and Objectives

1.4.1 Aim

This research aims at finding a solution using machine learning approach for the challenging problem of lip synchronization for long Sinhala sentences and to develop a Sinhala talking synthetic mouth part when the Sinhala text input is given.

1.4.2 Objectives

Main objectives are listed as below.

i. Develop a machine learning algorithm for generating set of viseme according to given digit, words, short and long sentences.

ii. Identify correct time frequency rate to synchronize the face model

iii. Develop a frontal view of a synthetic mouth model including tongue, lips and teeth to synchronize lip based on digit, word and long sentences in Sinhala language.

1.5 Scope

In this study, a talking synthetic mouth for long Sinhala sentences of eight to fifteen words will be created using a machine learning algorithm and a set of data. Overall study has been going through main four areas such as;

- viseme alphabet for the Sinhala language
- build a machine learning model to generate set of visemes for the given text input
- audio generation for the input Sinhala text
- method to synchronize the visemes and audio.

Generation of different visemes should has to be created for the Sinhala language based on the approach we are selecting. In the static viseme approach viseme alphabet needs to be created by analyzing the Sinhala language structure and rules.

Additionally, it is important to identify the factors that influence the phoneme to viseme mapping to build the machine learning model. The dataset for the text inputs must therefore be created by determining the connection between the phonemes and the text. This requires identifying the graphemes, which are the fundamental units of sound used to represent text. The system is designed to operate with any sentence without relying on the two primary spoken and written sentence structures.

Then, the Sinhala letters provided as the system input must be translated into audio for the text input. The system must be built to accept Sinhala letters, and audio should be produced for the specified set of Sinhala letters. This needs to be translated from text to speech in Sinhala.

Another important task for this research is to play a visual representation alongside the audio that was generated. The timing of that must be correctly synchronized. All of the Sinhala letters, words, short sentences, and long sentences have audio output and synchronization set up in viseme.

In this artificial mouth model, only the frontal view of the mouth portion of the face has been taken into account. The rest of the face, as well as facial expressions, emotions, and head movements, have not been taken. This model should be implemented using single-digit, single-word, and long sentences which are presented in text format.

Furthermore, due to the dearth of research based on the Sinhala language, this will be another contribution to the fields of NLP, machine learning, and deep learning globally. For the development of a machine learning model for the mapping between visemes and Sinhala sounds, there seems to be no research on the lip synchronization of Sinhala. In addition, processing Sinhala text rather than audio will contribute to improving the contribution to lip synchronization.

1.6 Structure of the Thesis

The thesis outline describes an overview of the key areas in the research work as described in following sections given below.

Chapter 01 – Introduction

- Includes introduction to the research study by including its background with aims, objectives and scope of the research study.

Chapter 02 – Literature Review

- Describes the literature related to area of study with relevant work and different approaches used in different research studies for other languages done by previous researchers

Chapter 03 – Methodology

- Includes process flow diagrams, constraints and design assumptions and algorithmic design details.

Chapter 04 – Evaluation and Results

- Includes research findings and the evaluation results of the research.

Chapter 05 – Conclusion and Future Work

- Summarize overall work of the research study and describes further improvements with alternative solutions.

CHAPTER 02 LITERATURE REVIEW

This chapter includes background information regarding different research studies in similar areas which are published in research papers, web articles, journals etc... Also, this chapter gives some knowledge in technologies and different approaches used in previous studies.

2.1 A literature review for Lip synchronization

Early days two dimensional animations has been popular in many fields such as entertainment, advertising and education. The creation of animations in those days are done by using hand drawings for each frames, then the key frames and motion curves are specified manually to have the movements of the character and objects(Aneja and Li, 2019). In past few decades' animation industries highly focused to build lip synchronized animations with more realistic experience to the user for different applications. As a result, nowadays more realistic cartoon characters can be seen in animation movies and also in live broadcast television programs with smooth synchronization rather than opening and closing mouth movements.

Human communication is doing using speech which is one of the most accepted mode of carrying the ideas and thoughts of a virtual character's personality. Articulatory movements which are actions necessary to vocalize language and facial expressions are highly affected to speech. Naturalness and believability of virtual character is highly impacted by possess of lip movements and the expressions of the face (Serra et al., 2012).

Lip synchronization is a method of matching speech with a synthetic face's lip movement that is most often used in animated movies. Animating cartoon characters which need lip synchronization is a challenging task in mapping lip animation movement with sounds produced by different languages (Loh, 2014). Basic sound units in a language is called as "phonemes" (Serra et al., 2012) and visual representation of mouth movement for each phoneme is called as "visemes" (Bear and Harvey, 2019).

Furthermore, visemes are the mouth shape corresponds when user pronounce phonemes. Visemes and phonemes are having one to many association, because the same viseme represents many phonemes in a particular language (Britto Mattos et al., 2018). For instance, bilabial sounds such as /p/, /b/, and /m/ which are usually grouped into one viseme (Bear and Harvey, 2019).

Several researches are done to systemized lip synchronization process for different languages which are producing several sounds using different approaches to avoid the difficulties arisen in animating with traditional techniques like hand drawings and manual specifications of key frames for movements . The next section will be discussed different approaches use to build lip synchronization process.

2.2 Different approaches for lip synchronization

Various approaches are used to create lip synchronization animation models, including the static visemes method (Weerathunga et al., 2020), the dynamic method (Thangthai et al., 2019), and the deep learning method (Britto Mattos et al., 2018).

The main component of static viseme approach is that it requires the viseme alphabet for the language in order to derive the speech animation sequence. Large data sets are typically used in deep learning approaches, and datasets for the English language are typically already available; however, Sinhala visual speech datasets are not readily available (Weerathunga et al., 2020). Ausdang Thangthai, Ben Milner and Sarah Taylor are proposing to increase naturalness of a visual speech using dynamic visemes with deep learning framework. They have considered Feed forward deep neural network and recurrent neural network using LSTM (Long short term memory) with many to one and many to many architecture (Thangthai et al., 2019).

At the moment, models are created using machine learning and deep learning techniques to have smooth lip synchronization between phoneme and viseme mapping. When it comes to more difficult tasks like word or sentence recognition, deep learning architectures have significantly outperformed earlier traditional methods, with word recognition rates increasing by 40%. The results of a survey on automatic lip reading in the era of deep learning techniques were presented by Adriana Fernandez and Federico Sukno (Fernandez-Lopez and Sukno, 2018).

One such supervised machine learning algorithm Support Vector Machine (SVM) is widely implemented in classification issues. Phoneme classification of lip synchronization also can be classified using SVM. Hanseoko and David has contributed to a research to reveal a live lip-sync using SVM method for phoneme classification to reduce the computational load. They have noticed that using SVM rather than the method created using the Hidden Markov Model (HMM), phoneme merging and recognition speed have both increased by two times (Ko and Han, 2006). HMM is a statistical model and that is also used in machine learning to describe the events which is depend on internal factors and those are not directly observable.

Convolution Neural Networks (CNN) is used to recognize viseme sequence in the synthetic data. Andrea Britto has done a research work using deep learning-based method to obtain high accuracy for the Visual Speech Recognition by improving CNN architecture for viseme recognition using deferent data set. It shows good results for words and sentences. (Britto Mattos et al., 2018). Sliding window regression approach is also another deep learning predictor that has been used to produce speech animation that synchronizes with input speech and looks natural. Sarah Taylor and the team could able to achieve minimal parameter tuning, generalizing the model for novel input sequence, real time execution and compatibility for existing animation retargeting approach by using sliding window predictor to produced speech animation (Taylor et al., 2017).

Lip reading models also can be produced using CNN architecture for real time processing. The study done by Karan Shrestha has been used two separate CNN architecture to train the subset of the dataset to devise an automated lip-reading system. Furthermore, model was implemented in a web application for real time word predictions(Shrestha, 2019).

To achieve lip synchronization by retaining information for a long time, Recurrent Neural Networks (RNN) have such a specific technique called Long Short Term Memory (LSTM). LSTM is also broadly used in deep learning approaches. Deepali Aneja and Wilmot Li presented a research work by implementing an interactive system to appear two dimensional character in live broadcast and streaming platforms with minimum 200ms latency using LSTM architecture in deep learning. They could able to generates live lip synchronization for 2D character by taking streaming audio as an input and producing viseme sequence as output using deep learning with LSTM architecture. (Aneja and Li, 2019).

The LSTM architecture also can be used to develop lip reading models. Reading lip in profile to the same standard by using Multi-view Watch architecture which is built using LSTM approach with five categories of cropped images by Joon Son Chung and Andrew Zisserman. They could able to find a proper solution for the question they have targeted at last with the limitation; that is the standard is inferior to reading frontal faces. Furthermore, they have mentioned this is an interesting to investigation of how the deep learning has learnt to select relevant information for each view and different architectures (Son and Zisserman, 2017).

Speaker independent video creation from audio input is another application that has been implemented using LSTM - RNN architecture in deep learning approach to generate proper mouth texture (Bhuiyan, 2021). Suwajanakorn could able to obtain remarkable results in the combination of former president Barack Obama speaking with accurate lip-sync matching an arbitrary input audio track using LSTM technique with video footage. Recurrent neural networks are used in this study to figure out exactly how to map raw audio features to mouth shapes. (Suwajanakorn et al., 2017)

Bear and Harvey introduced the ground-breaking two-pass training method for phoneme classifiers to develop a system for lip reading. The classification of visual units, which are derived from phoneme confusions, was done using a high-level process with three steps.(Bear and Harvey, 2019). High level three steps viseme unit classification is shown in Figure1.

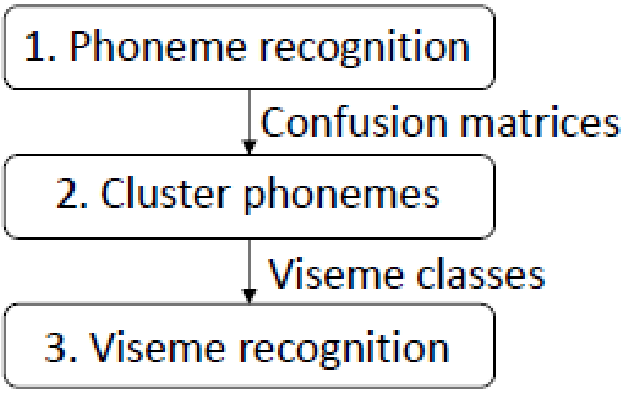


Figure 1: High level three steps for visual unit classification

By considering above facts, machine learning and deep learning approaches are highly used to develop talking speech models in different applications. Next section will be directed to explore existing speech models for Sinhala language.

2.3 A literature review for Sinhala language

Sinhala is the official language in Sri Lanka (Nadungodage et al., 2018) with 40 distinct sounds, including 14 vowel sounds and 26 constant sounds (Weerathunga et al., 2020).

A research study has been done by Chashika Weerathunga, is presenting talking synthetic face model for Sinhala language based on static viseme approach with rule-based algorithm. Fifteen (15) viseme groups are discovered using clustering and subjective analysis for Sinhala language. According to his results and evaluation, that model has shown 70.8% ranking accuracy and 68.8% rating accuracy. The synthetic face that was implemented tends to work well for single words and short sentences, but it does not demonstrate better synchronization for long sentences or sentences spoken in the various speeds. (Weerathunga et al., 2020).

CHAPTER 03 METHODOLOGY

This chapter includes research methodology for whole study, system design which is used to find the solution for the research problems, background of study and system implementation for the research.

3.1 Research Methodology

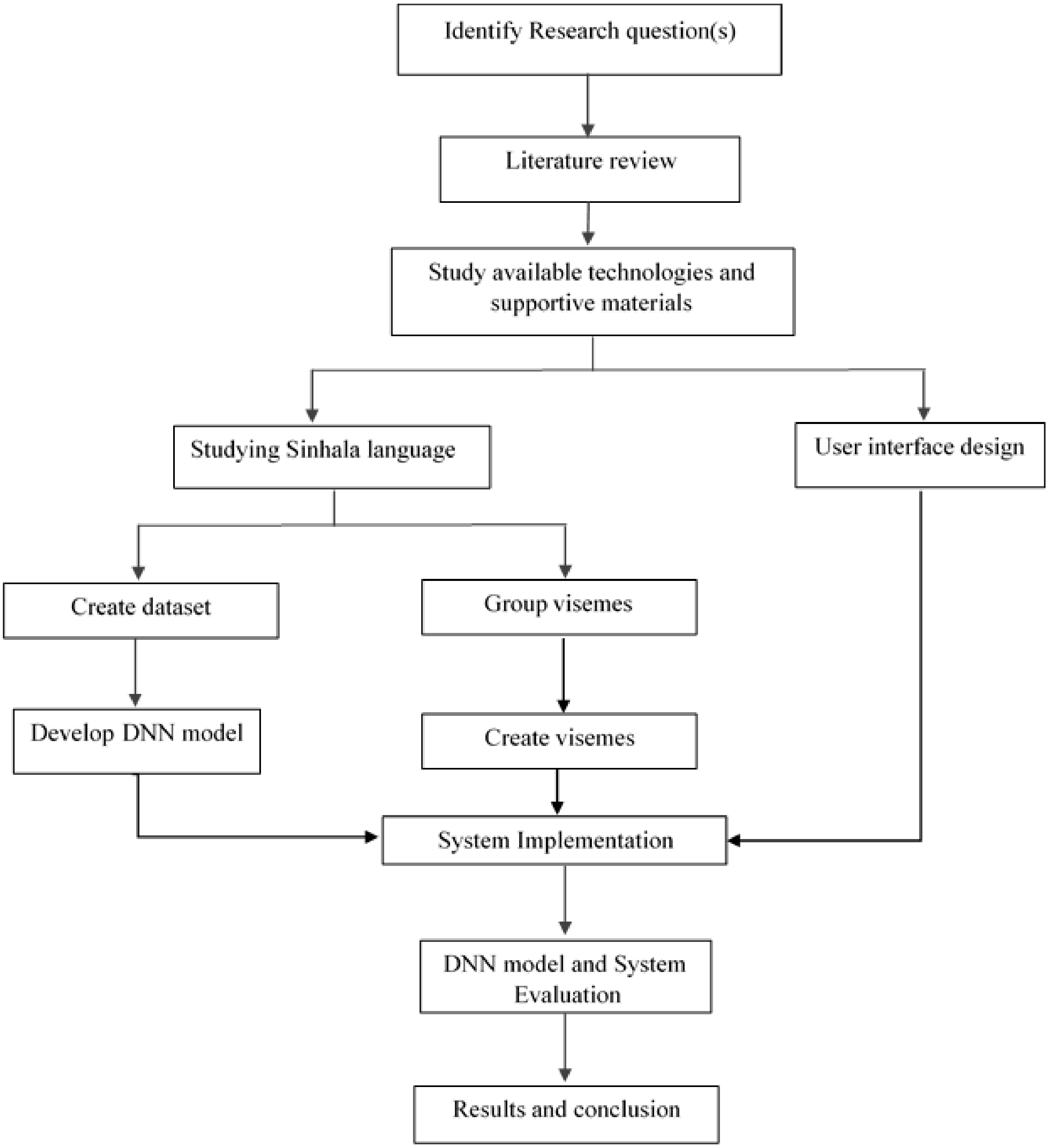


Figure 2: Block diagram for research methodology

Research methodology is the systematic way of research study to solve the research problem and achieve the objectives. The above Figure 2 shows the complete research design for the whole study by starting at **identifying research question(s)**.

In this research lip synchronized synthetic mouth is needed to build for the Sinhala language using machine learning technology. Therefore **literature review** has been done to understand the background of this study and identify the different approaches used in other researches. As a next step **supportive technologies and materials** has been studied. This study uses Python language for the coding for user interface designing and build model in 'Google colabs' which is a writable and executable platform in web browser allows to build the machine learning algorithm.

3.1.1 Studying the Sinhala language

In Sri Lanka Sinhala is the official language which is used by majority of population in the country. Literary and spoken are the two major varieties in Sinhala language. Sinhala text is the main input in this study and that has to be processed to build the visemes by identifying different phonemes as well as to create the dataset for the Deep Neural Network (DNN) model. That input text can be sentences given either literary format or spoken format.

Studying the Sinhala language is most essential in this study to create dataset for the DNN model and also to create static visemes by identifying the different sounds with visemes. Therefore, GCE O/L Sinhala reference book has been referred to find out the language structure and its rules. There are two major categories of Sinhala letters called vowels and consonants. Sinhala language consist with 18 vowels and 42 consonants in the alphabet.

Vowels in Sinhala build based only on one letter and representing one sounds such as; අ,ආ,ඇ, ඈ,ඉ,ඊ,උ,ඌ,එ,ඒ,ඔ,ඕ except ඓ and ඖ. Both are built as a combinations of two vowels as follows.

- ඓ = අ + එ
- ඖ = අ + ඔ

Consonants are the other category of Sinhala letters which are having base letter such as ක්,ඛ,ග්, ඝ්, ඞ්, ච්, ඡ්, ජ්, ඣ, ඤ්, ෂ්, ශ්, ඥ්, ට, ඨ්, ඩ, ඪ්, ණ්, ඬ, ත්, ථ්, ද්, ධ, න්, ප්, ඵ්, බ, භ්, ම්, ඹ, ය්, ර්, ල්, ළ්, ව, ෆ්, ස්, හ් . Sinhala alphabet will create another set of letters by combining base letters with vowels. For instance, letter ක is a combination of ක් and අ. Following Table: 1 shows the variations of letter ක with vowels. Other consonants also creates different types of letters in the similar way shows in the Table: 1.

Table 1: Variations of letter ක් with vowels

Consonants	Combination
ක	ක්+අ
කා	ක්+ආ
කැ	ක්+ඇ
කෑ	ක්+ඈ
කි	ක්+ඉ
කී	ක්+ඊ
කු	ක්+උ
කූ	ක්+ඌ
කෙ	ක්+එ
කේ	ක්+ඒ
කො	ක්+ඔ
කෝ	ක්+ඕ

Combining consonants with two vowels ඓ and ඖ will create another letter set with different sounds. As given in Table 2 by combining different letters with ඓ and ඖ will create different letters and sound for other consonants letters.

Table 2: Variation of letter ක් with ඓ and ඖ

Consonants	Combination
කෛ	ක් +අ+ එ
කෞ	ක්+අ+ ඔ

Another set of letters available with combining three letters such as ක්‍ර is a combination of ක්+ර්+අ which indicates by “රකාරාංශය” and ක්‍ය is a combination of ක්+ය්+ අ which indicates by “යංශය”. Similarly remaining consonants such as ක්‍ර, ග්‍ර, ට්‍ර, ත්‍ර, ද්‍ර, ප්‍ර, බ්‍ර, භ්‍ර, ව්‍ර, ශ්‍ර etc... also creates new letters as given in Table 2.

Table 3: Variations of letter ක් with ර් and vowels

Consonants	Combination
ක්‍ර	ක්+ර්+අ
ක්‍රා	ක්+ර්+ආ
ක්‍රැ	ක්+ර්+ඇ
ක්‍රෑ	ක්+ර්+ඈ
ක්‍රි	ක්+ර්+ඉ
ක්‍රී	ක්+ර්+ඊ
කෘ	ක්+ර්+උ
කෲ	ක්+ර්+ඌ
ක්‍රෙ	ක්+ර්+එ
ක්‍රේ	ක්+ර්+ඒ
ක්‍රො	ක්+ර්+ඔ
ක්‍රෝ	ක්+ර්+ඕ

Table 4: Variations of consonant letters with ය් and vowel අ

Consonants	Combination	Consonants	Combination
ක්‍ය	ක්+ය්+ අ	ඵ්‍ය	ඵ්+ය්+ අ
ඛ්‍ය	ඛ්+ය්+ අ	බ්‍ය	බ්+ය්+ අ
ච්‍ය	ච්+ය්+ අ	හ්‍ය	හ්+ය්+ අ
ජ්‍ය	ජ්+ය්+ අ	ම්‍ය	ම්+ය්+ අ
ට්‍ය	ට්+ය්+ අ	ල්‍ය	ල්+ය්+ අ
ඩ්‍ය	ඩ්+ය්+ අ	ව්‍ය	ව්+ය්+ අ
ත්‍ය	ත්+ය්+ අ	ළ්‍ය	ළ්+ය්+ අ
ථ්‍ය	ථ්+ය්+ අ	ෂ්‍ය	ෂ්+ය්+ අ
ධ්‍ය	ධ්+ය්+ අ	ශ්‍ය	ශ්+ය්+ අ
න්‍ය	න්+ය්+ අ	ඥ්‍ය	ඥ්+ය්+ අ
ප්‍ය	ප්+ය්+ අ		

Sinhala alphabetical letters are directly representing the sounds in the language. Also those sounds representing by the letters are not changed depending on the place where it is used. For Instance, letter අ represents the phoneme ‘a’. This sound is not changed based on the place where the letter is used. Anywhere that letter අ will give the sound ‘a’ without depending on the place where it is used. But in English language letter ‘u’ is representing sound ‘a’ at the word ‘run’, but it gives different sound at word ‘put’.

By studying Sinhala language following properties are identified. Those are;

i. Each letter in Sinhala alphabet represents sounds called phonemes.
ii. Combination of vowels and consonants produces different letters with different sounds
iii. Phonemes and visemes has the association while the Sinhala letters associates with phonemes. Therefore, visemes has an association directly with Sinhala letters.

After studying the language properties, structures and rules viseme creation and dataset creation for the DNN model training has been done. Considering the third factor (iii) stated above dataset has been created by mapping visemes with Sinhala letters. ‘Static viseme approach’ is selected to find the solution for the research study by considering the feasibility and the other facts.

3.1.2 Static viseme approach

Static viseme approach uses viseme alphabet which is derived as a sequence of speech animation for the different phonemes in the language. (Weerathunga et al., 2020). In this study frontal view of mouth movements which is showing lips, tongue and teeth for phonemes has been taken to create the viseme for each phonemes. Figure 3 shows the frontal view which is used to create visemes for the phonemes.

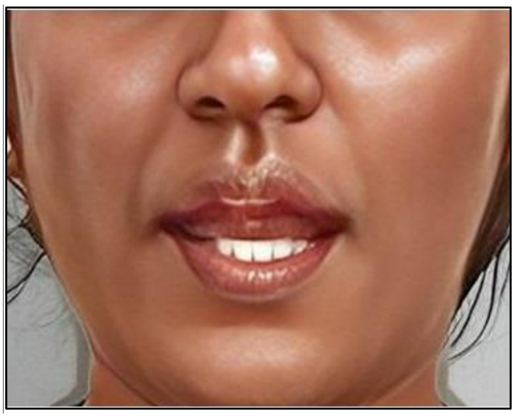


Figure 3: Frontal view of a viseme

Viseme alphabet has been created by saving the images according to the label given in ‘.png’ format. Following Figure 4 shows labeled visemes for letter අ and ආ in the png format.

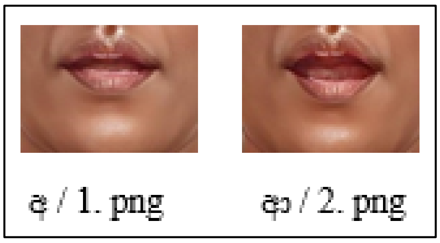


Figure 4: Visemes for අ and ආ

Visemes for vowels: Vowels which are a syllabic speech sound pronounced without any limit in the vocal tract. Our mouth is open freely when vowel sound is generated. There are twelve (12) vowels are basically selected by considering the language usage to create the visemes for vowels. Following figure is showing the vowels and its created visemes for each and number is the label given to the each viseme.

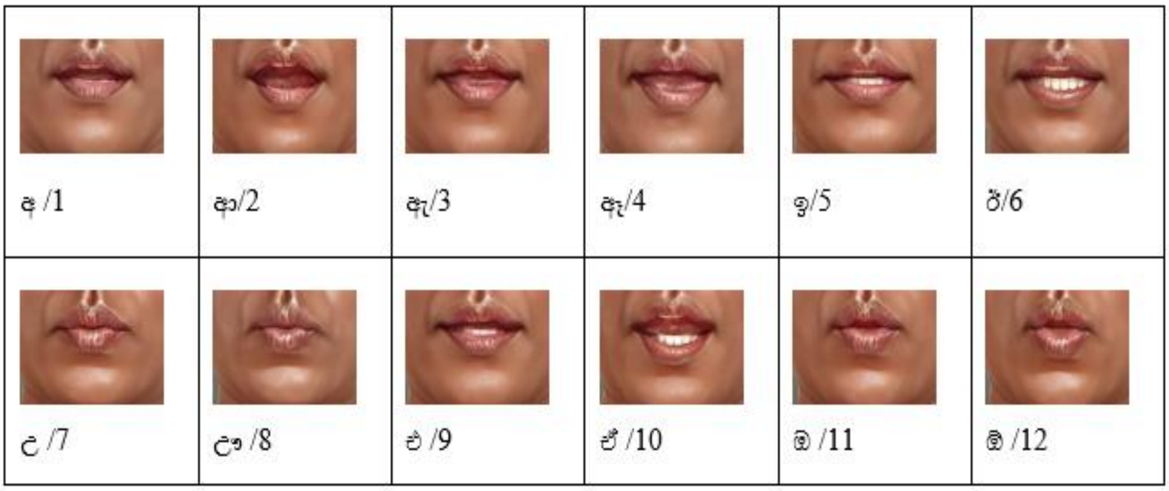


Figure 5: Visemes for vowels

Visemes for consonants: Consonants which produce speech sound by fairly closing the mouth. Vocal tract has a limit in the mouth when producing the sounds. Sinhala visemes of consonants are categorized based on the similarity features of the physical representation. For an example, frontal view of while saying the ප්, ඵ්, බ, භ්, ම, ඹ is same. The Figure 6 shows the grouping of consonants for visimes.

There are twenty three (23) visemes are identified in this study to create the static viseme alphabet including vowels and consonants. All the viseme images are saved according to the viseme number given in Figure 5 and Figure 6.

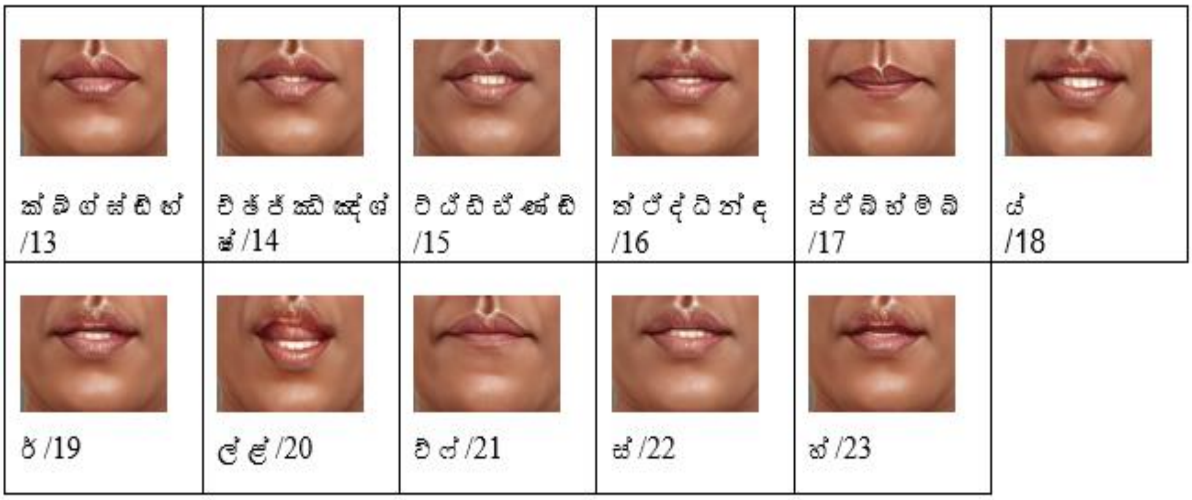


Figure 6: Visemes for consonants

Silent mode of the viseme labeled also included in the static viseme alphabet, that is saved with ‘0.png’ format and it is shown as following Figure 7.

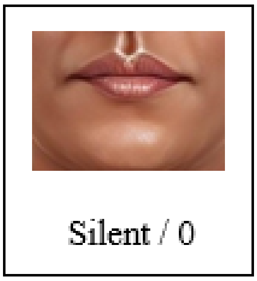


Figure 7: Silent mode

With parallel to the viseme creation data set is created by considering the language properties.

3.1.3 Data set creation

In this study viseme selection for the given text input should needs to be done using machine learning model. Text input can be a Sinhala digit, words, short sentences and long sentences. Therefore data set is created using the combination of letters in the Sinhala alphabet where the each letter representing a sound or combination of sounds.

Data set has been created by considering the combinations of letters which can be generated by mixing with vowels and consonants. In created data set only one feature vector used, that is the String type Sinhala letter and the target label is the viseme number assigned to the viseme image set as shown in Figure 4 and Figure 5.

Visemes to Sinhala letters has been mapped after identifying twenty three (23) visemes associates with different vowels and consonants. All the letters are producing the different sounds by combining or without combining letters which are listed in Table 3 where it is showing how the 23 visemes are mapped with Sinhala vowels and the consonants.

Visemes are labeled for 1 to 12 for the first twelve vowels which are producing the sound by opening vocal tract without limitations. Then for the consonants which are producing sounds from our vocal tract opening with limitation are labeled from 13 to 23. Each letter has unique viseme label but for a viseme label can be assigned for multiple letters.

Table 5: Grouping of viseme to vowels and visemes to consonants

Viseme number	Sinhala vowels
1	අ
2	ආ
3	ඇ
4	ඈ
5	ඉ
6	ඊ
7	උ
8	ඌ
9	එ
10	ඒ
11	ඔ
12	ඕ

Viseme number	Sinhala consonants
13	ක් ඛ ග් ඝ් ඞ ඟ්
14	ච ඡ් ජ් ඣ ඦ් ශ් ෂ්
15	ට ඨ් ඩ ඪ් ණ් ඬ
16	ත් ථ් ද් ධ න් ඳ
17	ප් ඵ් බ භ් ම ඹ
18	ය්
19	ර්
20	ල් ළ්
21	ව ෆ්
22	ස්
23	හ්

Four (4) major types of Sinhala letters which are showing different sounds in the alphabet has been identified and labeled as follows.

1. Vowels with a one sound

The given Table 4 is representing mapping between twelve vowels with labels of visemes. Here only one label is added as the target because those 12 vowels produced only one specific sound and also one specific viseme representation. Those phonemes are having the association with one to one relation with visemes. For the easiness of the coding purpose target labels are added within the "[]" square brackets.

Table 6: Vowels with one sound mapping to the visemes

Letter	Target
අ	[1]
ආ	[2]
ඇ	[3]
ඈ	[4]
ඉ	[5]
ඊ	[6]
උ	[7]
ඌ	[8]
එ	[9]
ඒ	[10]
ඔ	[11]
ඕ	[12]

2. Vowels with a two sounds

The following table contains the two vowels which are a combination only of the other two vowels. No consonants letters are mixed here. Therefore, while labeling the target two viseme labels are included.

Table 7: Vowels with two sounds mapping with visemes

Letter	Target
ඓ	[1,10]
ඖ	[1,11]

3. Consonants with initial sounds

According to the grouping of Table:3 consonants with base sounds are mapped with visemes by considering two major associations such as; some letters are associates with one to one and some letters are with one to many with visemes and letters.

Table 8: Consonants with base sounds mapping with visemes

Letter	Target
ක්	[13]
ඛ	[13]
ග්	[13]
ඝ්	[13]
ඞ්	[13]
ච	[14]
ඡ්	[14]
ජ්	[14]
ඣ	[14]
ඤ්	[14]
ෂ්	[14]
ශ්	[14]
ඦ්	[14]

Letter	Target
ට	[15]
ඨ්	[15]
ඩ	[15]
ඪ්	[15]
ණ්	[15]
ඬ	[15]
ත්	[16]
ථ්	[16]
ද්	[16]
ධ	[16]
න්	[16]
ප්	[17]
ඵ්	[17]

Letter	Target
බ	[17]
භ්	[17]
ම	[17]
ඹ	[17]
ය්	[18]
ර්	[19]
ල්	[20]
ව	[21]
ෆ්	[21]
ස්	[22]
හ්	[23]

4. Consonants with two sounds

Consonants letters are produced as a combination of two sounds are labeled using two viseme labels. For instance, letter කො is a combination of ක්+ඔ both are labeled in the viseme alphabet 13 and 11 respectively. Therefore, label for letter කො is [13, 11]. The given Table 7 shows labeling by referring the Table 1 for variations of letter ක් with 12 different vowels producing different letters and sounds. In similar way other consonants also labeled in the dataset.

Table 9: Labeling for consonants with two sounds for letter ක

Letter	Target	Letter	Target
ක	[13,1]	කු	[13,7]
කා	[13,2]	කූ	[13,8]
කැ	[13,3]	කෙ	[13,9]
කෑ	[13,4]	කේ	[13,10]
කි	[13,5]	කො	[13,11]
කී	[13,6]	කෝ	[13,12]

5. Consonants with three sounds

There are some letters which are a combination of three sounds can be represented using letters. Basically there are three major types of letters produces letters by combining three letters.

i. Letters combined with two vowels ඓ and ඖ

Following table represents the labeling for letters build using two vowels ඓ and ඖ by referring Table 7 and Table 8. Similarly other consonants such as; ගෛ, ගෞ, මෛ, මෞ, සෛ, සෞ etc.… which are available with both vowels as a combination has been labeled accordingly.

Table 10: Labeling letters combined with two vowels

Letter	Target
කෛ	[13,1,10]
කෞ	[13,1,11]

ii. Base letters with ර and vowels

With the help of Table 3 which is showing variations of ක් with ර and vowels could be able to label in the data set. The given table shows the mapping for selected consonants ක් with ර and vowels. For an example ක්‍ර is a combination of ක්+ර+අ which are mapped with viseme labels as 13, 19 and 1 respectively. In the dataset letter ක්‍ර has the target as [13, 19, 1].

Table 11: Labeling letters combined with ර and vowels

Letter	Target
ක්‍ර	[13,19,1]
ක්‍රා	[13,19,2]
ක්‍රැ	[13,19,3]
ක්‍රෑ	[13,19,4]
ක්‍රි	[13,19,5]
ක්‍රී	[13,19,6]
කa	[13,19,7]
කaa	[13,19,8]
ක්‍රෙ	[13,19,9]
ක්‍රේ	[13,19,10]
ක්‍රො	[13,19,11]
ක්‍රෝ	[13,19,12]

iii. Letters with “යංශය”

According to the Table 4 which is representing variations of consonant letters with ය් has the viseme label 18 and vowel ඇ has the viseme label 1 labeling for letters works with has been mapped.

Table 12: Labeling letters combined with ය් and vowel ඇ

Letter	Target	Letter	Target	Letter	Target	Letter	Target
ක්‍යා	[13,18,1]	ත්‍යා	[16,18,1]	බ්‍යා	[17,18,1]	ෂ්‍යා	[14,18,1]
ඛ්‍යා	[13,18,1]	ථ්‍යා	[16,18,1]	භ්‍යා	[17,18,1]	ශ්‍යා	[14,18,1]
ච්‍යා	[14,18,1]	ධ්‍යා	[16,18,1]	ම්‍යා	[17,18,1]	ඥ්‍යා	[14,18,1]
ජ්‍යා	[14,18,1]	න්‍යා	[16,18,1]	ල්‍යා	[20,18,1]		
ට්‍යා	[15,18,1]	ප්‍යා	[17,18,1]	ව්‍යා	[21,18,1]		
ඩ්‍යා	[15,18,1]	ඵ්‍යා	[17,18,1]	ළ්‍යා	[20,18,1]		

In the dataset each letter has one unique viseme label but one viseme label has many letters associated with it. Those labels are the target label given for each letter in the dataset. Finally six hundred and one (601) data samples could be able to record in the data set with two hundred and fifty six (256) viseme labels over the entire dataset.

3.1.4 Deep Neural Network (DNN) Model

Dataset is showing multiple classes among 601 data samples, where each data sample is labeled to one target among 256 multiple viseme classes. Therefore the multi class classification model where the classification work done with more than two classes has been implemented. Also dataset is an imbalanced dataset because 256 classes are not represented equally among 601 data samples. Then our model should needed to perform multiclass classification with imbalanced data.

There are two major types of multiclass classification in supervised machine learning technology called as “one vs all” and “one vs one”. Both methods are describes with the usage of different classifiers based on the number of classes available in the dataset. Our data set contain 256 classes, therefore that is difficult to do the viseme classification by using both methods. Because, one versus all multiclass classification needs 256 classifiers which is decided based on the number of output classes in the data set. On the other hand for one versus one classification number of classifiers are decided based on the mathematical formula as follows.

$$\boldsymbol{Number\ of\ classifiers\ =\ n\ x\frac{n-1}{2}}\ where\ \boldsymbol{n}\ \text{is}\ number\ of\ output\ classes$$

According to the above equation one versus one multiclass classification needs 30640 classifiers where n is equals to 256.

Then, as a solution to avoid the complexity of using machine learning classification development the model has been focused to build using neural network technique which is a part of machine learning. Neural networks are behave as human brain functionality Deep learning is a subset of machine learning and it also uses neural networks which is use weights by adjusting via training the model to find the correct output for the new inputs.

Multi-layer perceptron (MLP): This is a fully connected neural network including three layers which are input layer, one hidden layer and output layer. When MLP has more than one hidden layer that is called as Deep Artificial Neural Network (DANN) or Deep Neural Network (DNN). These are feed forward artificial neural network where all the nodes in each layer is fully connected called as “dense layers”.

By considering all the facts mentioned above , DNN has been selected to build a model for the viseme selection by including two hidden layers which uses 30 nodes for one layer and 10 nodes for the next hidden layer. Input layer consists 601 nodes taken as one node per input value in the dataset. Output layer consists with 256 nodes because this is a multiclass classification build using deep neural network.

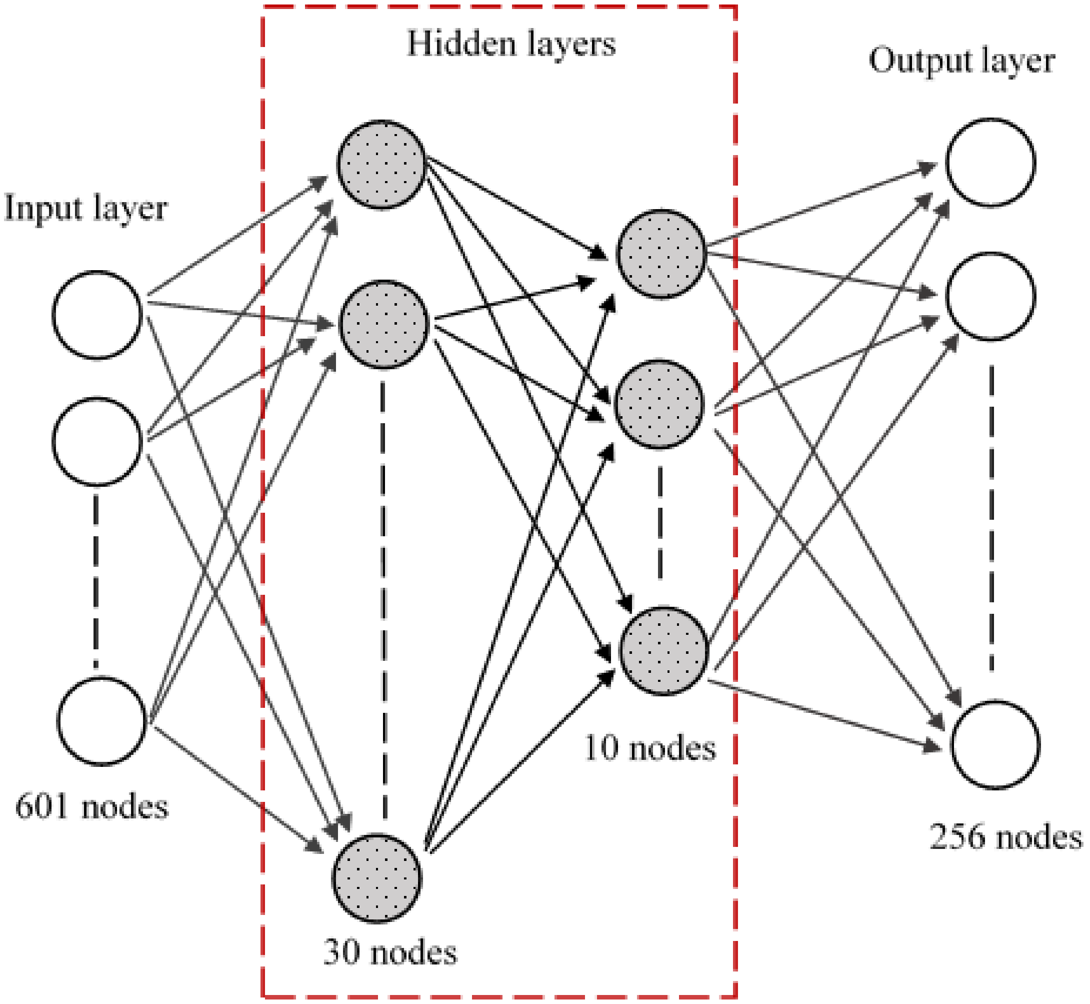


Figure 8: Neural Network model

Activation function: Neural networks uses activation function to determine a node in a layer should need to be activated or not. There are variations of activation functions such as linear function, sigmoid, tanh etc. As an activation function of each layer 'Rectified Linear Unit (ReLU)' has been used in the network to proceed output from nodes.

ReLU is less expensive in computationally with comparing to the sigmoid and tanh. Also this is widely used in multiclass classification problems. This is nonlinear activation function which can activate the multiple layers of neurons in the two hidden layers and output layer.

Activation function for the ReLU activator is given as below.

$$\boldsymbol{ReLU\ activation\ Function\ =\ max\{0, x\}}\ where\ \boldsymbol{x > 0}$$

If the input is a positive then the function will give the output else the output is zero for all the negative inputs in the network. The Figure: 9 represents the drawn graph for ReLU activation function.

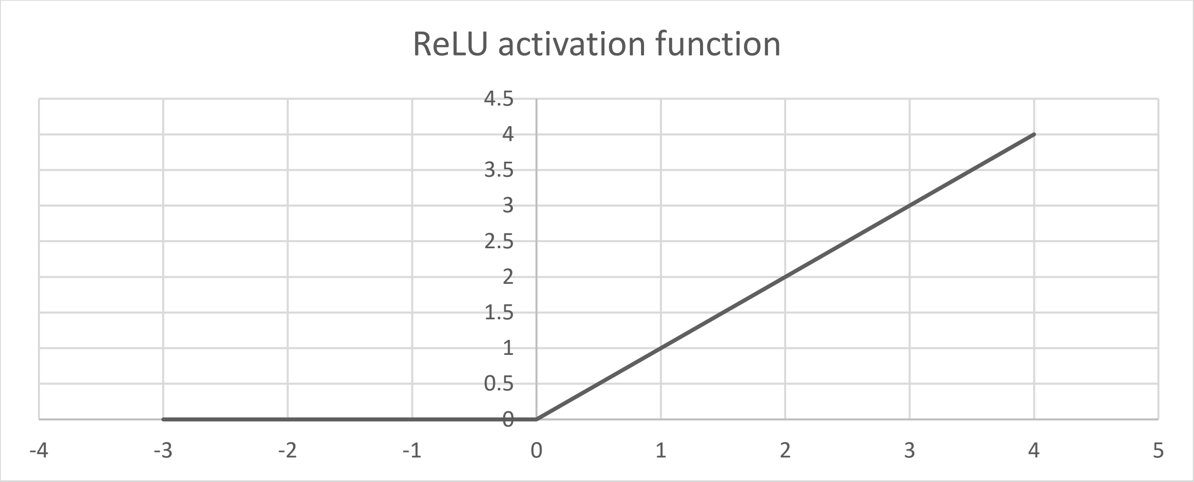


Figure 9: Graph for ReLU activation function

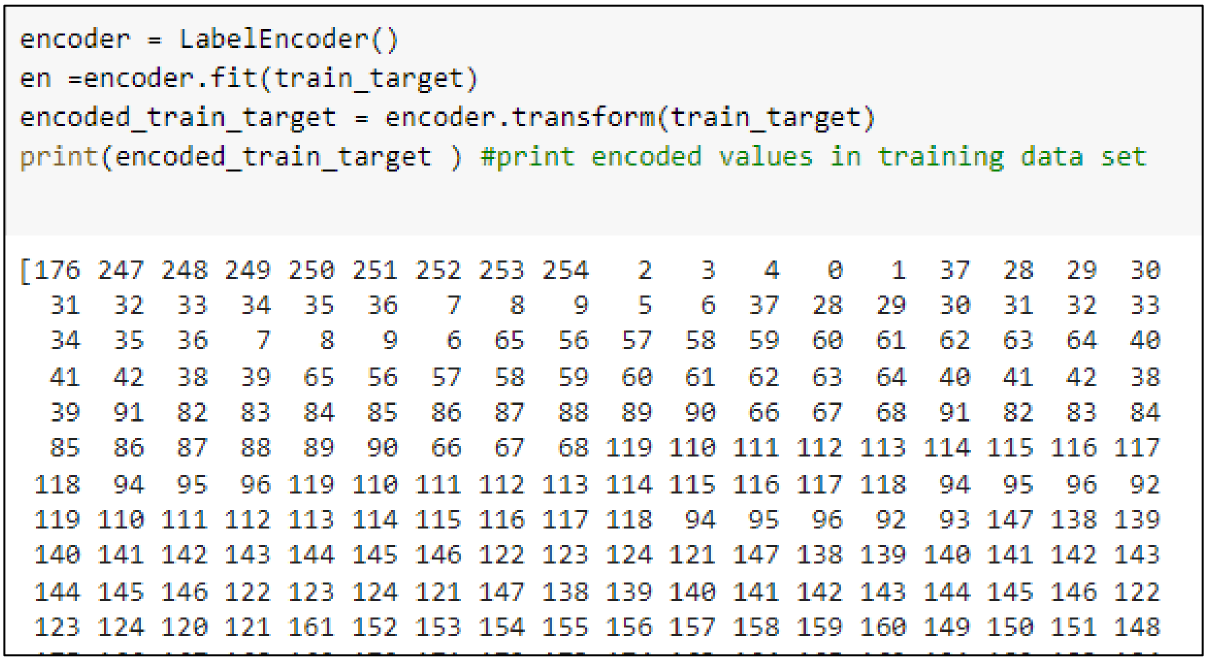
Loss function: Loss is the different between ground truth value which is actual output and the predicted output. Neural networks use different loss function to compute the loss happens and it will help to train the neural network. Classification loss function has been used in the DNN model that is “cross entropy” classification loss function.

Optimizer: Neural network has different parameters to train the model such as weights and learning rates. Those attributes has to update by fine tune the network to train the model to generate correct output for the new data. Optimizers are used to produce accurate results by reducing different between target value and the predicted value.

In the DNN model which have been developed fort the viseme classification uses ‘Adaptive Moment Estimation (Adam)’ as the optimizer for the network which is built for the gradient decent. “Adam” is a one of the best optimizer can be used in DNN to train the model within very less time.

In the implementation of the DNN algorithm which has been coded using Python language in the ‘Google colabs’ platform which is built to write and execute Python codes, that is more suitable for machine learning. Deep learning also comes under the supervised machine learning category but it is reducing feature extractions and feature presenting data steps in machine learning methods.

Training and testing data are created in the ‘.csv’ format. The testing data set has been created by shuffling the data. The testing and training data set is encoded to a numerical format because the feature vector and labels are given in string format. Following figure shows the coding and encoded values of training data set of the model



```
encoder = LabelEncoder()
en =encoder.fit(train_target)
encoded_train_target = encoder.transform(train_target)
print(encoded_train_target ) #print encoded values in training data set
```

```
[176 247 248 249 250 251 252 253 254   2   3   4   0   1  37  28  29  30
  31  32  33  34  35  36   7   8   9   5   6  37  28  29  30  31  32  33
  34  35  36   7   8   9   6  65  56  57  58  59  60  61  62  63  64  40
  41  42  38  39  65  56  57  58  59  60  61  62  63  64  40  41  42  38
  39  91  82  83  84  85  86  87  88  89  90  66  67  68  91  82  83  84
  85  86  87  88  89  90  66  67  68 119 110 111 112 113 114 115 116 117
 118  94  95  96 119 110 111 112 113 114 115 116 117 118  94  95  96  92
 119 110 111 112 113 114 115 116 117 118  94  95  96  92  93 147 138 139
 140 141 142 143 144 145 146 122 123 124 121 147 138 139 140 141 142 143
 144 145 146 122 123 124 121 147 138 139 140 141 142 143 144 145 146 122
 123 124 120 121 161 152 153 154 155 156 157 158 159 160 149 150 151 148
```

Figure 10: A part of encoded data set of training data

Then the model has been trained and tested, then the model evaluation has been done by writing a lambda function in Python code. Finally the model has been trained until it reaches more accuracy and minimum losses. All the evaluation scenarios of the DNN model has been checked.

As a next step **system implementation** has been done by creating user interfaces using Python language. System has been integrated by including DNN model in the system architecture and giving the reference path to the static viseme alphabet directory in the coding.

Last two steps of the research methodology is **evaluation and results**. Evaluation has been done under two major areas.

i. Evaluation for DNN model

   This evaluation has been done by using performance matrices, accuracy and losses calculations

ii. Evaluation for entire system

   The completed system is evaluated in three major areas; rating evaluation for checking the user satisfaction, evaluation for checking the system produces correct viseme and another evaluation for checking whether the visemes representation gives help to memories more words rather than hearing only the sounds.

Results and discussion is done by considering the evaluation results obtained in the evaluation process.

3.2 System Design and Implementation

This section will discuss about system design use to address the research questions to find solutions and the system implementation details for the research study. Following Figure 9 describes about the main components and the relationship among them to perform the lip synchronization for the given text input.

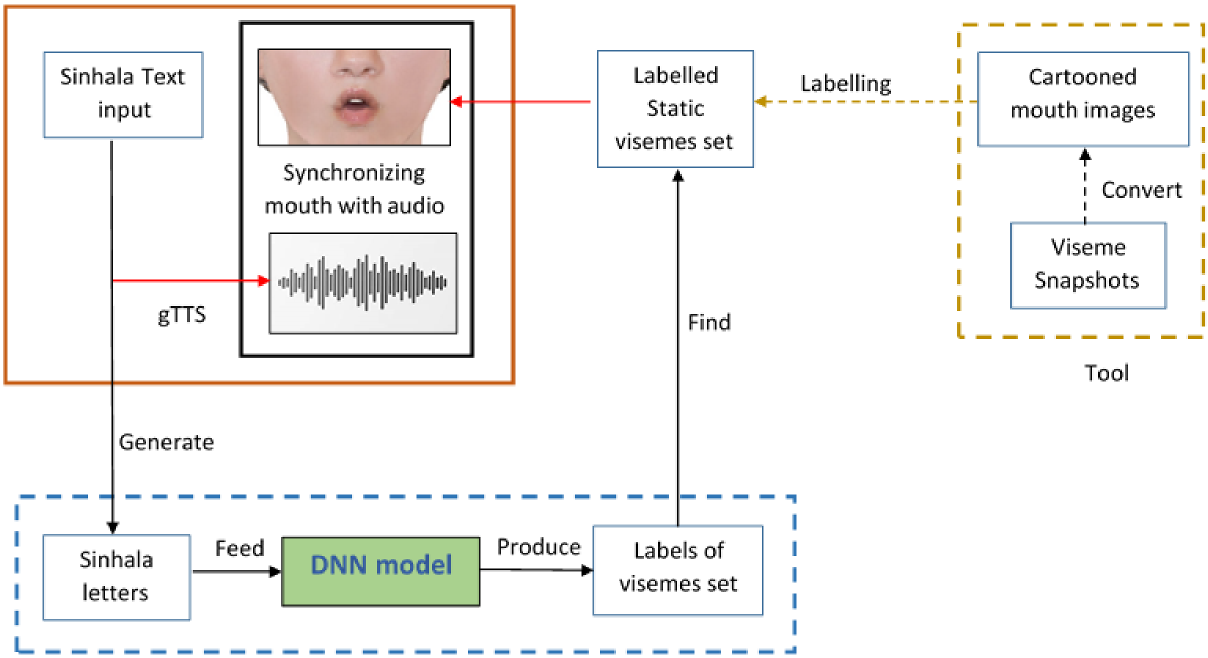


Figure 11: Block diagram for system design

Mainly the system is working with Sinhala text which can be a Sinhala digit, word, short sentence or a long sentences. That input is taken to generate audio and the desired visemes using the gTTS and DNN model. The Sinhala letters extracted from the input text is feed to the DNN model to generate viseme labels. Then the produced viseme labels are used to find viseme images stored in "image" folder using an algorithm written in the main system. Then those images are taken to an array to call until the audio start playing.

As shown in above figure the entire system is consisted with three main units. Those are;

1. Viseme creation

   Visemes are created by taking snapshots for different viseme representations for Sinhala alphabet. Then it has been converted to the cartoon format using "TOONME.COM" mobile application. Finally, cartooned images has been cropped with same dimensions using online tool. Those cropped images are saved in an "image" folder to refer it to the main system.

2. Audio generation based on the text input

   Sinhala text is processed by Google Text to Speech (gTTS) which is a Python library use to mediate with Google Translates text to speech API.

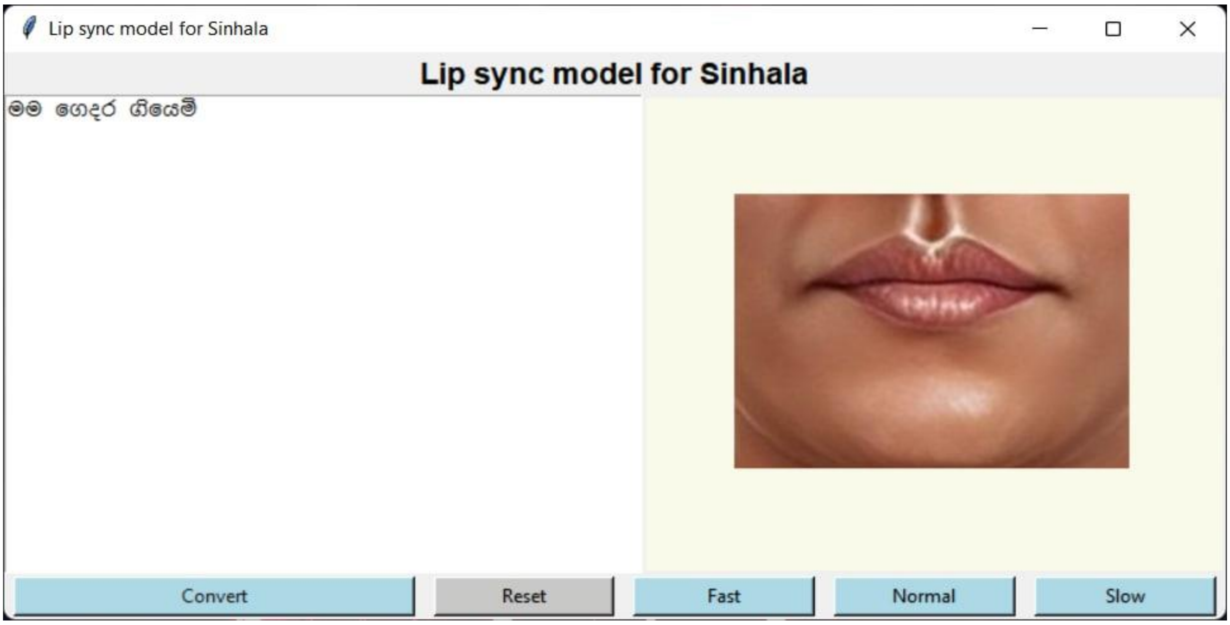
3. DNN model

   DNN model built using ‘Google colabs’ platform using ‘TensorFlow’ API to generate the correct set of visemes for given text input. It was evaluated after training and testing the model and generated model is saved in the google drive. Then it will be downloaded and merged with the main setup to generate the viseme labels once the Sinhala letters feed to the model.

3.2.1 Implementation

In the system implementation user interface designing and other internal functions are built using “Python” programing language. The main interface is developed as shown in Figure 3 that is including text input area, viseme displaying canvas and five buttons used to control the system.

Sinhala text input is converting to the audio once click on the “Convert” button and also it will calling internal function to split the Sinhala letters in input text. Then letters will feed into the DNN model to find the viseme labels along with letters. After that, system will find viseme images by referring the produced viseme labels and selected visemes are taken to the display canvas in the user interface.



මම ගෙදර ගියෙමි

Figure 12: User Interface for Lip synchronization model

There are three (3) buttons included to control the speed of audio and viseme while the synchronization happening. “Fast” button will play with speed rate 2, “Normal” button speed rate is 1 and “Slow” button speed rate is 0.5.

In addition to that “Reset” button has been put to the interface to clear the text area and function it again if needed.

3.2.2 Time for Synchronization

Synchronization is a process where the identified viseme set and generated audio playing together during a time period. Following figure shows the identify the parameters for calculation.

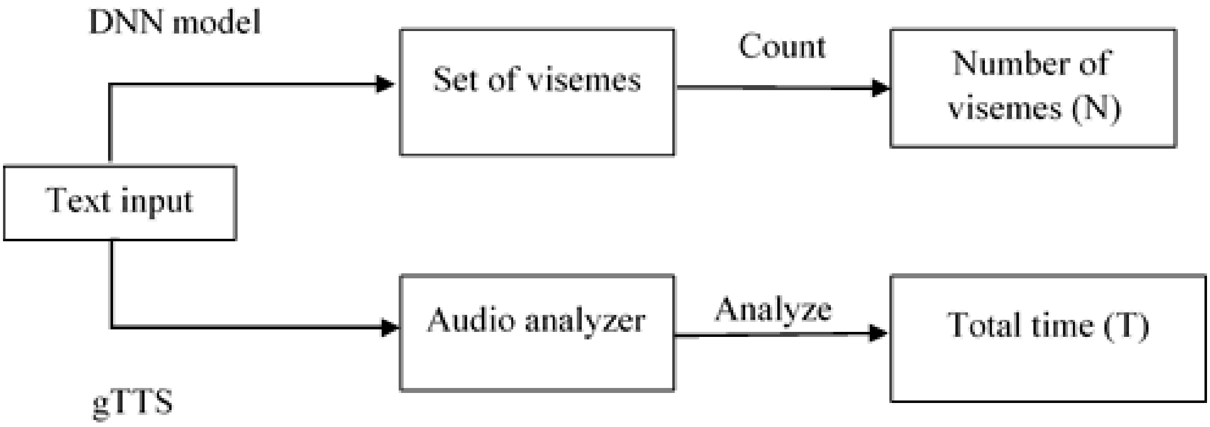


Figure 13: Calculation of a time for viseme

Audio is generated by gTTS module and viseme set will be identified by machine learning model. Total time (T) for the audio can be identified by analyzing the audio using audio analyzer while the number of visemes (N) can be counted to calculate the time duration for each viseme to play by dividing the total time from number of visemes.

$$\boldsymbol{Time\ for\ each\ viseme\ =\ Total\ time\ (T)\ /\ number\ of\ visemes\ (N)}$$

The system will give the calculated time durations for each visemes to play with the audio.

CHAPTER 04 EVALUATION AND RESULTS

In this chapter describes the evaluation methods of the developed model and its results. The previous study of the lip synchronization using rule based algorithm is showing the subjective analysis for the model but no other research can be found for the evaluation for the machine learning with DNN model built to the lip synchronization of Sinhala language.

Basically the deep learning model and entire system has been evaluated separately. Following sections will be described the evaluation methods of deep learning model and system evaluations done based on rating and other two methods.

4.1 Deep learning model evaluation

The deep learning model has been evaluated in the different scenarios such as accuracy, average loss and loss has been evaluated. “TensorBoard” is a graphical representing tool build for the “Tensorflow” to visualize the model accuracy, average loss and hidden layer zero activations. When building the model batch size is given as 32 which is the number of processing samples. Step size has been calculated by dividing the total data samples by batch size.

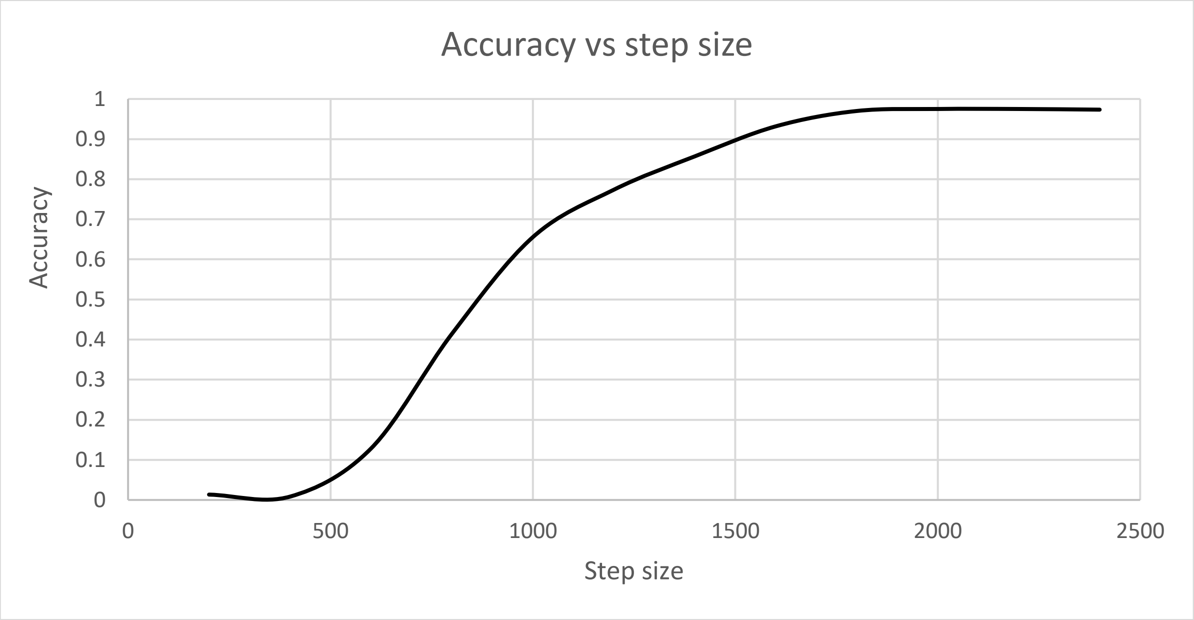


Figure 14: Graph for accuracy changes in step size

The following Figure: 14 shows how the accuracy has been changed when the step size is increased. Accuracy is increased with the step size and the highest accuracy for the model taken is 0.975 in the model evaluation. That value is obtained at step size is 2000 while the model is trained. In order to accuracy graph that model is showing better accuracy for the data we have trained.

Average loss is another parameter of evaluation of the DNN model. The Figure: 15 is showing the average loss versus step size.

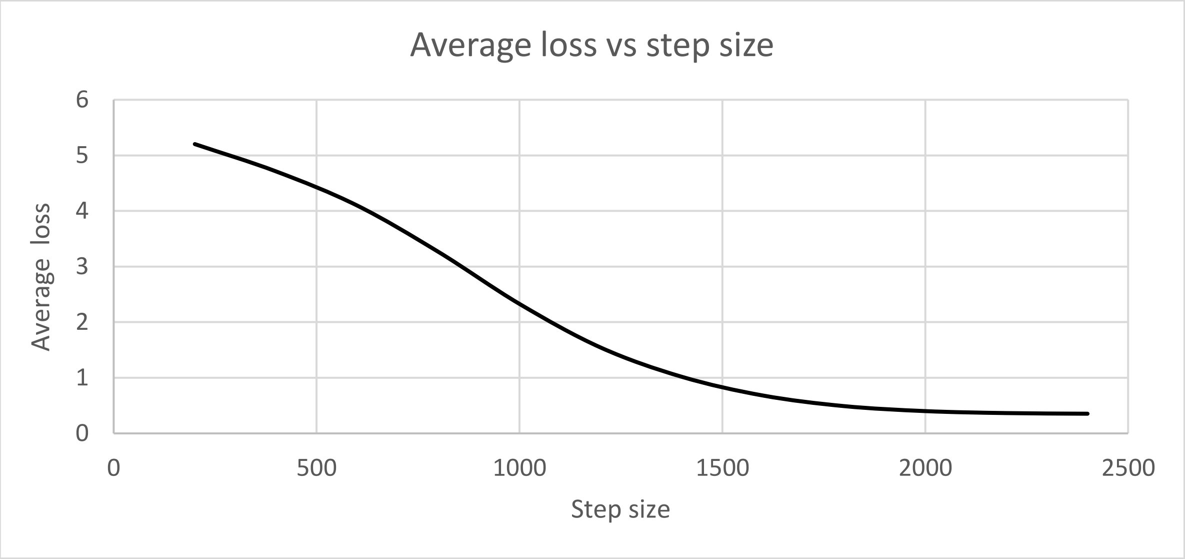


Figure 15: Graph for Average loss vs Step size

At the beginning of this graph shown above the average loss is having higher rate 5.202 at the 500 steps. Average loss has been gradually decreased with step size while the model is training. That value is at 0.3542 at the step size is 2400. With the step size the average loss has been decreased in the model we have trained.

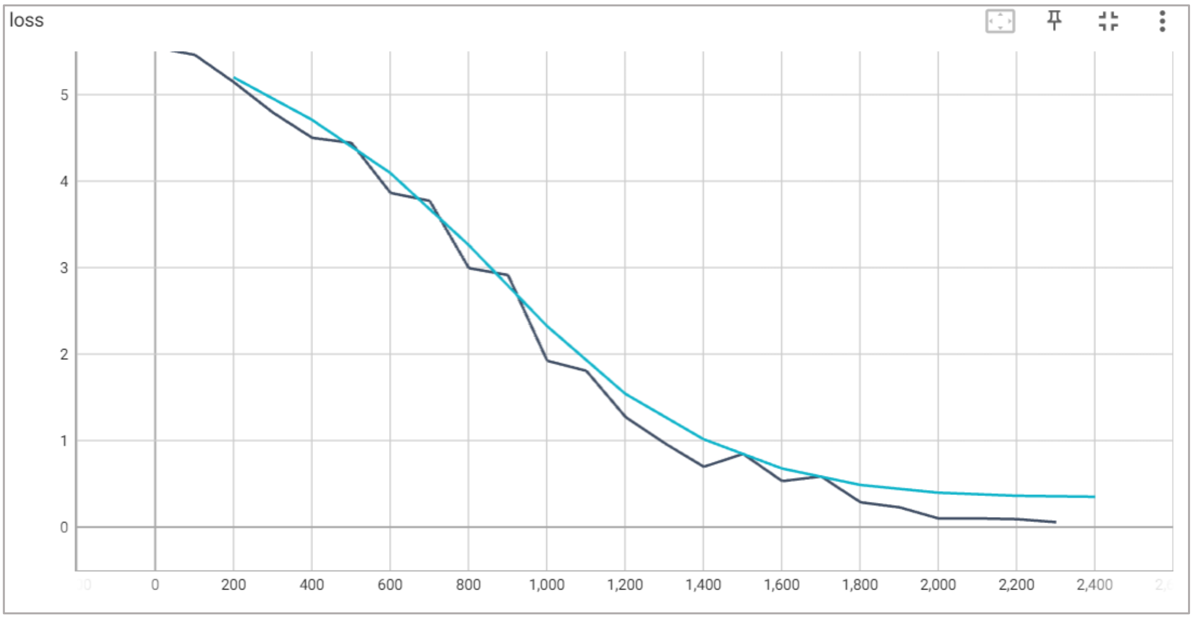


Figure 16: Loss in training and evaluation

The Figure: 16 given below is comparing the loss when the model is training and in the evaluation. Blue colored line is the loss at evaluation and black colored graph is the loss at the training process of model. Both the graphs are going along with each other and that is not showing any of the either model overfitting or under fitting.

Considering the behaviors of both the graphs our model is good fit at training and evaluation process because the training and validation loss is going along with each other.

Batch size of the DNN model is 32 which is number of samples per one step and our dataset contain 601 data samples. One epoch which is a one complete cycle of the training data. Epoch value is calculated by dividing the total number of data samples from batch size, that is 18.78125 ~ 19 steps. That is one epoch has 19 steps for one cycle of training data.

Global steps are the number of total steps achieved until the model reach to better accuracy.

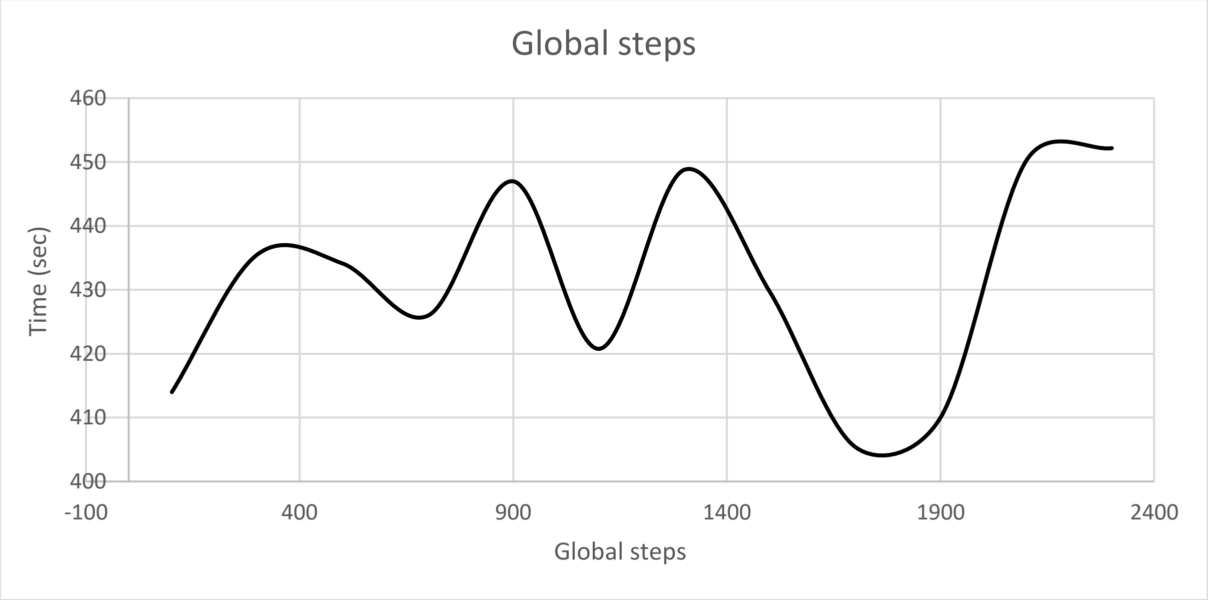


Figure 17: Graph for Global steps

The above graph is showing how the global steps varied over the time in seconds with the 19 step size. It is showing sudden fall at global step 1701 in to the time 405.4 seconds. After that the time has been gradually increased when the global steps reached to 2301. There are three peaks at global step 30, 901 and 1301 with the time 435.5s, 447s and 448.1.

"ReLU" is one of a best activation function used in DNN and it will make the dense layers into sparse by producing outputs from some nodes to maintain the efficiency of the network. Nodes which are not producing the output is called as the dropouts in the network. Every nodes in the hidden layer will not produce the output every time. When the output is produced that will indicate by 1 otherwise that value is 0.

Fraction of zero values are calculated by "TensorBoard" to show the behavior of nodes which are activated and not activated. It is providing two other graphs regarding DNN model that describes the hidden layer activations during the training phase. The fraction of zero is calculating by considering the number of nodes not produce the output as a fraction of the total number of nodes in the layer. Our DNN model has two hidden layers with 30 and 10 nodes for each. Fraction of zero values are calculated and it is showing in Graph: 18 and Graph: 19.

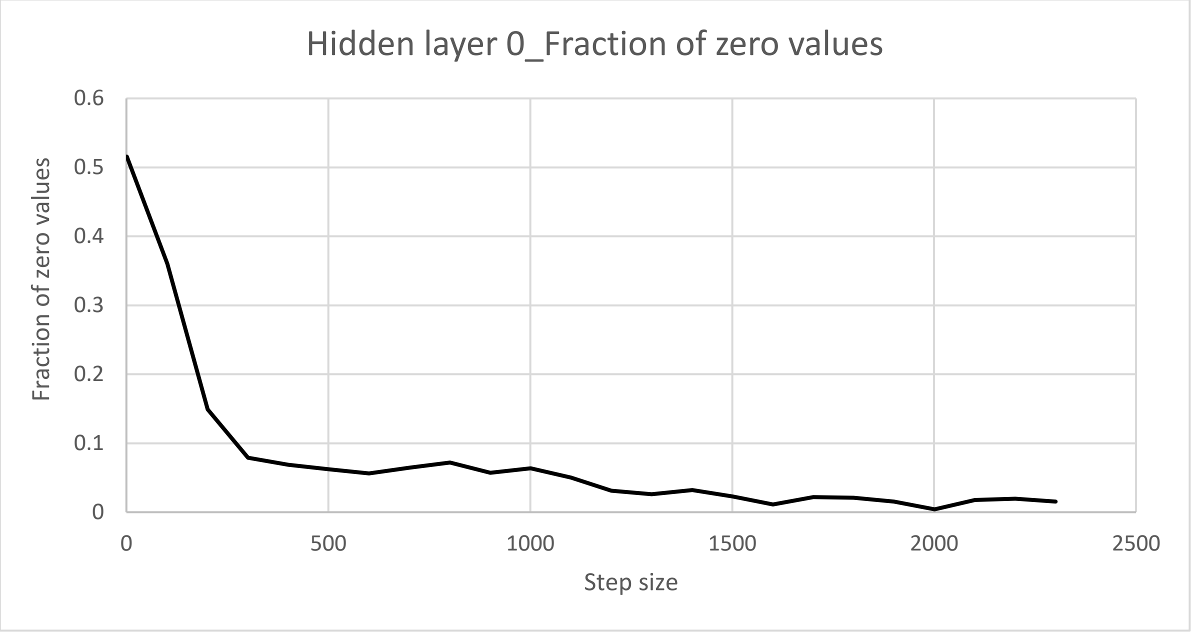


Figure 18: Graph for Fraction of zero in Hidden layer 0

The above Graph: 18 is showing the fraction of zero along with the step size. According to the graph fraction of zero value at higher rate in the beginning but it is sudden fall into 0.05625 at step size is 301. After the 500 step size that value is gradually decreased and closer to zero. In hidden layer zero very few number of nodes are not producing outputs after the 500 steps. In Graph: 19 is representing fraction of zero for hidden layer1.

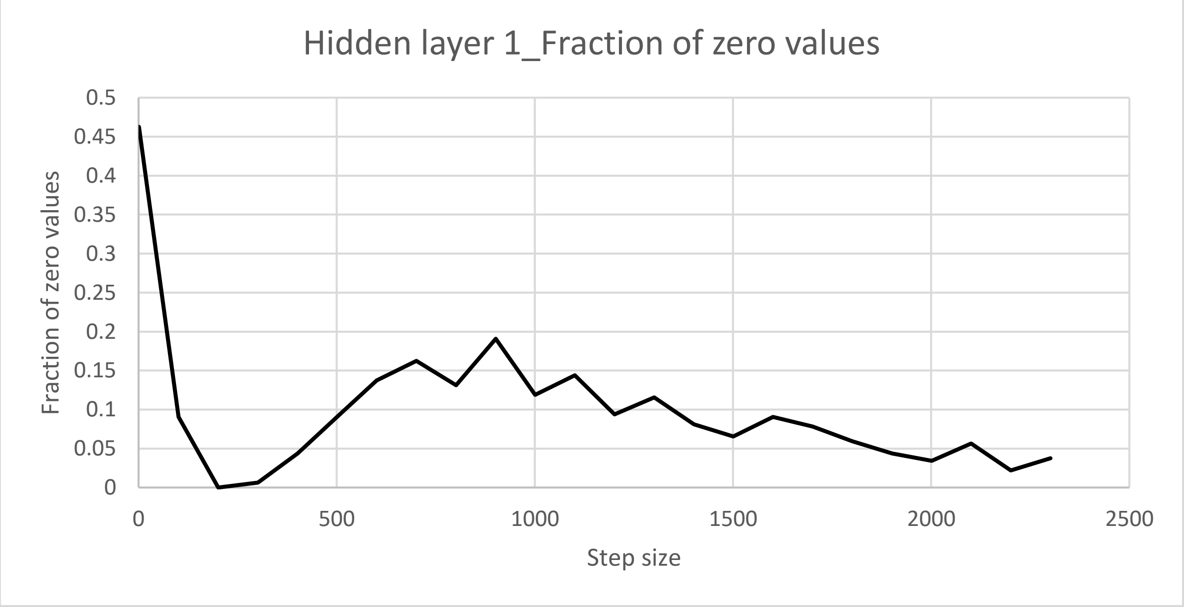


Figure 19: Graph for fraction of zero values in Hidden layer 1

At the step size 201 the fraction of zero value is at zero that means at that step all the nodes in hidden layer 1 is activated and produces the output. Then the fraction values are gradually increase up to 0.1906 at the step size 901. Then it is gradually decreased. According to the graph some nodes are not producing the outputs at the beginning but most of the nodes produces the output at the after step size is 1000.

Classification matrices: This metrics is used to know the performance of the DNN model. Following Table: 13 is showing classification matrices, those are macro average precision, macro average recall, macro average F1 score and macro average support as accuracy.

Macro average is the arithmetic mean of the each single class for precision, recall and F1 score. This is used in multiclass classification with imbalanced data which is different classes are assigned with different classes..

$$\boldsymbol{Macro\ average\ precision} = \frac{(\boldsymbol{prec1} + \boldsymbol{prec2} + \ \dots + \ \boldsymbol{precn})}{\boldsymbol{n}}$$

$$where\ n\ is\ number\ of\ classes$$

This model designed for the multi class classification and accuracy, precision which describes how many positive identifications are actually correct in the data set and recall is the how many actual positives are correctly identified in the dataset has been calculated for the each class

Above equation shows the macro average precision. In Table: 13 macro average precision is 0.99. The predicted value is correct around 99% while the model is predicting a viseme label. The macro average is calculated as shown in the following equation.

$$\boldsymbol{Macro\ average\ recall} = \frac{(\boldsymbol{rec1} + \boldsymbol{rec2} + \ \dots + \ \boldsymbol{recn})}{\boldsymbol{n}}$$

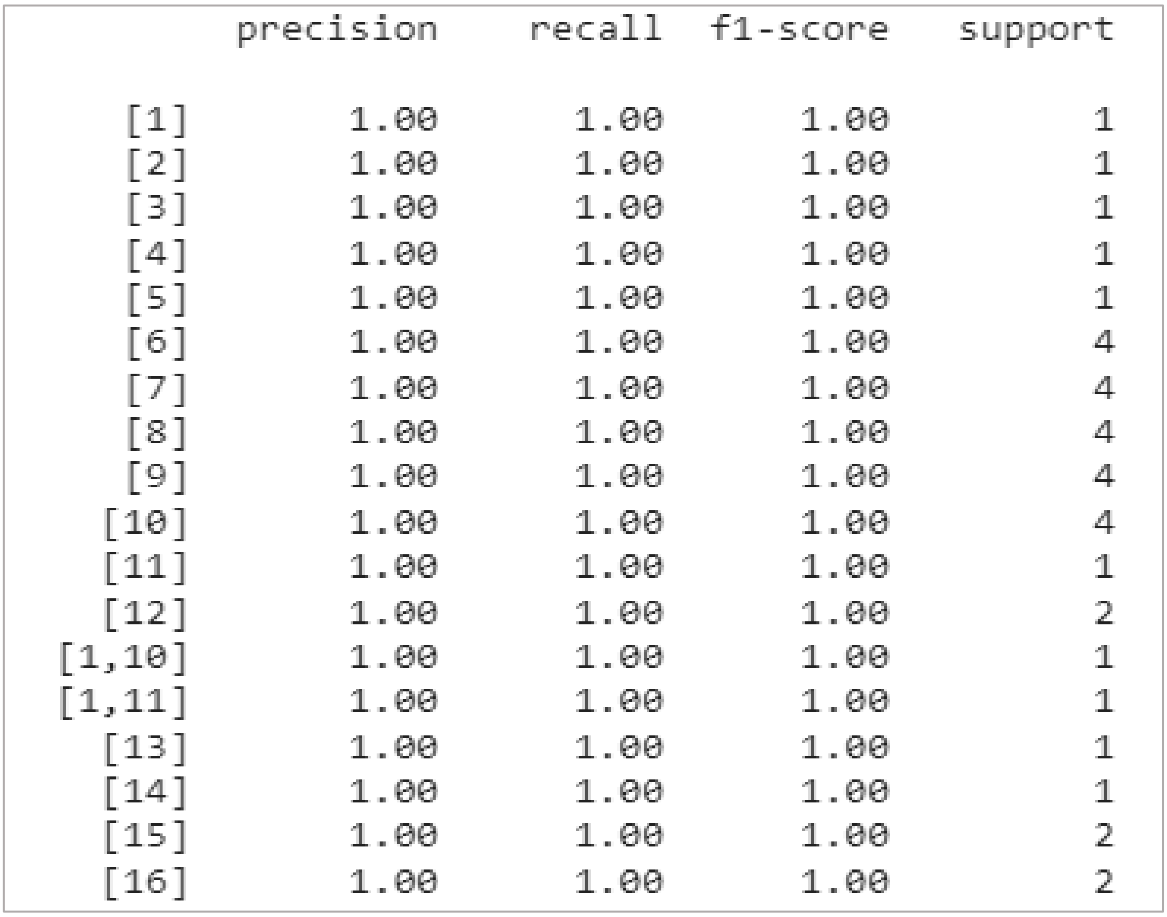
$$where\ n\ is\ number\ of\ classes$$

In Table: 13 macro average recall is also 0.99 and it shows the percentage of correctly identified classes. Recall and precision values are same therefore F1 score which is harmonic mean of precision and recall. That is also 0.99 and it can be used when the precision and recall values are showing larger difference. But in our case precision and recall is showing same value, therefore the do not need to go for F1 score to measure the performance of the metric.

Table 13: Classification Metrics for DNN model

	Precision	**Recall**	**F1 score**	**Support**
Accuracy			0.98	601
Macro Average	0.99	0.99	0.99	601
Weighted Average	0.99	0.98	0.98	601

A part of classification metrics generated for each class has been shown in the following figure. Most of the probabilities for each classes are one because the data set contain only one feature vector that is a unique Sinhala letter.



```
          precision    recall  f1-score   support

     [1]       1.00      1.00      1.00         1
     [2]       1.00      1.00      1.00         1
     [3]       1.00      1.00      1.00         1
     [4]       1.00      1.00      1.00         1
     [5]       1.00      1.00      1.00         1
     [6]       1.00      1.00      1.00         4
     [7]       1.00      1.00      1.00         4
     [8]       1.00      1.00      1.00         4
     [9]       1.00      1.00      1.00         4
    [10]       1.00      1.00      1.00         4
    [11]       1.00      1.00      1.00         1
    [12]       1.00      1.00      1.00         2
  [1,10]       1.00      1.00      1.00         1
  [1,11]       1.00      1.00      1.00         1
    [13]       1.00      1.00      1.00         1
    [14]       1.00      1.00      1.00         1
    [15]       1.00      1.00      1.00         2
    [16]       1.00      1.00      1.00         2
```

Figure 20: Part of classification matrices for each class

The deep learning model has been compared with different activation functions and also with different optimizers. Following table will shows the comparison among those.

According to the Table: 23 it is showing that when using the different configuration how the model perform with average loss and accuracy. Lowest average loss (0.4022) and highest accuracy(0.975) could able to find using “ReLU” as activation function , Adam as optimizer with two hidden layers with 30 and 10 nodes while the number of batch size is 2800 and 32 respectively.

Table 14: Comparison between Activation functions and Optimizers

Activation Function	Optimizer	Hidden Layers	Number of steps	Batch Size	Average loss	Accuracy
Relu	Adam	30x10	1000	512	0.49025488	0.975
Relu	Adam	30x10	2200	64	0.4107133	0.975
Relu	Adam	30x10	2800	32	0.4022	0.975
Relu	Adam	20x5	4400	32	0.98696935	0.975
Relu	Adam	30x5x2	2800	32	4.453558	0.157
Relu	Adam	30x5	3600	32	1.2887037	0.975
Softmax	AdaGrad	30x10	600000	512	5.843	0.4748
Softmax	Adam	30x10	4800	8	0.5222415	0.975
Relu6	Adam	30x10	5000	512	0.48508015	0.975
Relu6	AdaGrad	30x10	5000	512	6.2037153	0.016
Tanh	AdaGrad	30x10	8000	512	4.785494	0.143

Each letter has shown better probability when it is evaluated and if the input given from another language it is showing lower probability. Model has been evaluated using Sinhala and non-Sinhala letters and according to the Table 15 model is showing higher probability for Sinhala letters and but very low probability for the English letters.

Table 15: Probabilities generated for Sinhala and non-Sinhala letters

Letter	produced viseme label	Probability
ඉ	[5]	95.76%
ඊ	[6]	90.47%
A	[15,2]	18.30%
b	[15,2]	18.30%
r	[15,2]	18.30%

Threshold value is added in the system implementation by considering those probabilities to avoid non Sinhala letters in the text input. If any letter has the lower probability than the threshold value the system will showing the notice message to the user about input.

4.2 System Evaluation

Entire system has been evaluated using three main methods to identify user satisfaction about the system using rating method, to get a feedback on correct viseme production of the model and to identify whether the visemes gives a help to understand the words we are listening rather than only hearing the sound.

All three methods are done using Google forms by sharing with users. Lip synchronization system outputs are recorded at normal speed using a screen record tool "BANDICAM" and it recorded videos for all three evaluation methods are attached to the Google form. Also all three evaluations are highly guided to obtain the accurate results.

4.2.1. Method 01 : Rating method

Rating method is used to check whether the satisfaction of synchronization of the animation of the model and the audio. A questionnaire is provided with a model animation to the given digits, words and sentences. The given questions are rated as linear scaled between 1(satisfaction of the synchronization) to 5(satisfaction of the synchronization). Questionnaire is shared as a Google form with forty (40) participants.

Results for Sinhala Digits: The results are tabulated as given in following table for the Sinhala digits which have been evaluated using questionnaire using ten (10) Sinhala digits. Those ten digits includes vowels, base consonants and dependent letters which are built using vowels or vowels and consonants together.

Table 16: Results for Sinhala digits

Sinhala Digits	Ratings (%)				
	1	2	3	4	5
ඈ	0	0	12.8	33.3	53.8
ඖ	2.5	7.5	12.5	37.5	40
උ	5	5	10	27.5	52.5
ශ්‍රී	0	2.5	12.5	25	60
ඵ	0	0	10.8	24.3	64.9
ඹ	0	0	12.8	25.6	61.5
ඥ	0	12.5	12.5	40	35
ණෝ	0	0	12.5	25	62.5
ට	10	5	17.5	25	42.5

ඏ	0	2.5	5	20	72.5

In the results of the above table letter ඏ has a highest satisfaction for the model 72.5% among 40 participants and the 40% lowest at the scale 4 for the letter ඥ**.** On the other hand letter ඛ**,** උ and ට has dissatisfaction among participants as 2.5%, 5% and 10% respectively.

According to the green colored highlighted cells most of the digits are showing highest satisfaction percentage for the synchronization by showing highest percentage at the scale 5. Nine words out of ten number of words are showing highest satisfaction rate at scale 5 among 40 participants.

Results for Sinhala words: The same questionnaire the next section is provided with ten Sinhala words to evaluate satisfaction of the lip synchronization for the words. Following table shows the results obtained from forty participants for the ten words. Those words are included Sinhala naming words and Singlish words.

Table 17: Results for Sinhala words

Sinhala Words	**Ratings (%)**				
	1	**2**	**3**	**4**	**5**
වෛර	0	2.5	15	62.5	20
ධජය	0	7.5	37.5	37.5	17.5
ඖෂධ	2.5	0	17.5	35	45
මඤ්ඤොක්කා	0	0	12.5	52.5	35
ප්‍රතිඵල	0	5	15	45	35
ඥානපාල	0	0	7.5	35	57.5
ඕල්වේස්	0	10	20	25	45
දධය	0	10.3	30.8	28.2	30.8
කෲර	2.5	5	20	35	37.5
වයෝවෘද්ධ	2.5	12.5	27.5	35	22.5

According to the Table: 15 highest satisfaction 62.5% at scale 4 is for the word **"වෛර"** and 30.8% lowest value at scale 3 and scale 5 showing for the "දධය". Five out of ten words are showing highest percentage at scale 5, those are "ඖෂධ", "ඥානපාල", "ඕල්වේස්", "දධය" and "කෲර". Other words are spread over the scale 3 and scale 4 as average satisfaction.

Dissatisfaction of the lip synchronized system for three words "ඖෂධ", "කෲර" and "වයෝවෘද්ධ" are showing lowest rate at scale 1 as 2.5 % for all three words. When considering green colored cells those are spread over the scale 5, 4 and 3 of ten words.

Then in the next section of the questionnaire has been included “Short sentences” to check whether the system is satisfied while the synchronization happens while playing sequence of words.

Results for Sinhala short sentences: Short sentences are included more than two (2) words but less than eight (8) words. There are ten words including written and spoken Sinhala word structure is rated by 40 participants. Not only that but also included some partial set of words which is not showing complete sentence and Singlish sentences are included in the sentences. Following table shows the sentences has been used in the questionnaire.

Table 18: Short sentences

Short Sentence (SS)	Sinhala Short Sentences
SS_1	ඒ වෙනුවෙන් පෙළඹීමේ අඩුවක්
SS_2	නිමල් එළුවාට කන්න දුන්නා
SS_3	දරුවෝ අඳෝනා තබමින් අඬති
SS_4	ධනාත්මක චින්තනය අපට දැන් අවශ්‍ය වී ඇත
SS_5	අ.පො.ස උසස් පෙල භෞතික විද්‍යාව
SS_6	ඕල්වේස් බ්‍රේක් ඩවුන්
SS_7	එලදායී ක්‍රම දහයක්
SS_8	සේවය සැපයිය හැකි ප්‍රජාව
SS_9	බෞද්ධයෝ චෛත්‍ය වන්දනා කරති
SS_10	පාරිභෝගිකයෙකු යම් භාණ්ඩයක් ඇණවුම් කල විට

For the easiness of the representation each sentence has been labeled as SS_1, SS_2, etc... Satisfaction of the lip synchronization model is obtained as following table.

Table 19: Results for Short sentences

Short Sentences	Ratings (%)				
	1	2	3	4	5
SS_1	0	7.7	30.8	41	20.5
SS_2	0	15	27.5	35	22.5
SS_3	0	7.5	22.5	40	30
SS_4	7.5	25	27.5	32.5	7.5
SS_5	0	5	27.5	35	32.5
SS_6	0	7.7	28.2	33.3	30.8
SS_7	0	0	35	27.5	37.5
SS_8	0	7.5	25	25	32
SS_9	2.5	0	40	42.5	15
SS_10	0	20	32.5	27.5	20

Highest percentage 42.5% at scale 4 is showing for SS_9 sentences that is “බෞද්ධයෝ චෛත්‍ය වන්දනා කරති”. For the lowest 32.5% is showing SS_10 that is at scale 3 as average satisfaction. There are three sentences SS_6, SS_7 and SS_8 are showing scale 5 as their highest satisfaction as 30.8%, 37.5% and 32% respectively. Sentence SS_4 and SS_9 is showing lowest satisfaction at scale 1 as 7.5% and 2.5% respectively.

The results in Table: 17 has spread over the scale 4 in the case of satisfaction of short sentences as highlighted in green colored cells. Six sentences are now in the scale of 4 and three sentences are at scale 5.

Then the questionnaire consists with “Long sentences” for the rating evaluation.

Results for Sinhala long sentences: Long sentences are contain more than eight (8) words but less than fifteen words including spoken and written sentence structure. There are ten long sentences are found from Sinhala news articles including dialogs, news reading sentence structure and written Sinhala language structure. Those sentences are also included Singlish words.

Table 20: Long Sinhala sentences

Long Sentence (SS)	Sinhala Long Sentences
SL_1	කොරෝනා වෛරස සනයට අයත් වෙනත් වෛරස මොනවාද කියලා හොයලා බලන එක ඉතාම වැදගත් වෙන්න පුලුවන්
SL_2	බොහෝ දෙනා මුදල් බැංකුවේ ස්ථාවර ගිණුමක තැන්පත් කිරීමට පෙළඹුනත්, ආයෝජනයන් වෙත පෙළඹීමේ අඩුවක් තිබෙනවා
SL_3	පැරණි සාහිත්‍යයේ විවිධ රචනා ශෛලීන් සඳහා නිදසුන් කිහිපයක් මතු දැක්වේ
SL_4	තින් කන්ටෙන්ට් තියන ඒවා නන්ඉන්ඩෙක්ස් කලාම අනිත් පේජස් වලට හොද බූස්ට් එකක් හම්බ වෙනවා
SL_5	යට ගිය දවස බරණැස් නුවර බ්‍රහ්මදත්ත නම් රජ කෙනෙකුන් රාජ්‍ය කරන කල්හි
SL_6	මහ රැජිණගේ අභාවයත් සමග වාල්ස් කුමරු බ්‍රිතාන්‍යයේ රජු ලෙස පත්ව සිටි
SL_7	“ආයුබෝවන් ශ්‍රාවක හිතවතුනි. අද ලෝක ජල දිනය මාර්තු විසි දෙක”
SL_8	භාෂාව කියන්නෙ පොත් පත් වගෙ අතට අහු නොවන දෙයක් නිසා ප්‍රමාණයක් කියන්න අමාරුයි
SL_9	මැතිවරණයක් කැඳවිය යුතු බව ආදිවාසී නායක ඌරුවරිගේ වන්නිලඇත්තන් ඊයේ අවධාරණය කළේය
SL_10	විශ්ව සාහිත්‍යයේ පුරෝගාමී ලේඛකයන් අතර විලියම් ශේක්ස්පියර් අමරණීය චරිතයකි

Above table shows Long sentences which have been selected for the evaluation. Ten sentences are labeled as SL_1, SL_2, etc... for the easiness for representing in tabular format. Following are the results obtained from the 40 participants about the satisfaction of the lip synchronization of the long sentences.

Table 21: Results for Long sentences

Long Sentences	Ratings (%)				
	1	2	3	4	5
SL_1	0	10	37.5	40	12.5
SL_2	2.5	20	42.5	25	10
SL_3	0	7.5	27.5	37.5	27.5
SL_4	10	10	40	32.5	7.5
SL_5	0	7.5	40	40	12.5
SL_6	0	0	22.5	30	47.5
SL_7	0	7.5	42.5	32.5	17.5
SL_8	0	2.5	22.5	35	40
SL_9	0	10	52.5	20	17.5
SL_10	0	5	35	30	30

According to the Table: 19 highest rating percentage among 40 participant is 52.5% at scale 3 for sentence SL_9, which is "මැතිවරණයක් කැඳවිය යුතු බව ආදිවාසී නායක ඌරුවරිගේ වන්නිලඇත්තන් ඊයේ අවධාරණය කළේය". There only two sentences, SL_6 and SL_8 is showing best satisfaction at scale 5. Dissatisfaction is showing for the two sentences SL_2 and SL_4 with rating percentages 2.5% and 10% respectively.

When considering spread of the green colored cells about the highest satisfaction of each long sentence is lies on scale 3 and scale 4.

4.2.2. Method 02

This evaluation method is used to check whether the system produce the correct visemes for given text. Providing a lengthy sentences are not practical because none of the participants are not an experts in lip reading. Therefore ten words are provided with a time gap between each word. Following table contains the words including named entities, verbs, Singlish words, months and numbers given in two videos with different sequences.

Table 22: Set of words for Evaluation 2

Word	Sinhala text
W1	මාලඹේ
W2	මාර්තු
W3	මෝටර් රථ

W4	පන්සිය හැට අට
W5	දුවනවා
W6	පොලීසිය
W7	ෆැන්ටා බීම
W8	එළවා
W9	සීයා
W10	ගුහාවක්

A Google form is shared with 40 participants and that include a sequence of words with audio and mouth movement in the first section. Participants has to watch the video by carefully listening the audio at once. Then in the next section another video given without an audio and different word order of the previous word set. By looking at the viseme playing in the video user has to select the identified words in the words given in the next section by playing the second video several attempts. There are ten words has been given in the different orders in two videos. In the questionnaire user has to input number of attempts taken to find the word in the list.

Figure 15 is a bar chart drawn to represents the percentages of identified words by only seen on visemes. “මාලඹේ”, “මාර්තු” and “එළවා” three words are having highest percentages as identified words using visemes. Lowest percentage is for the word “මෝටර් රථ” with 50% percentage. Overall all the ten words are identified with more than 50% higher rate in the bar chart.

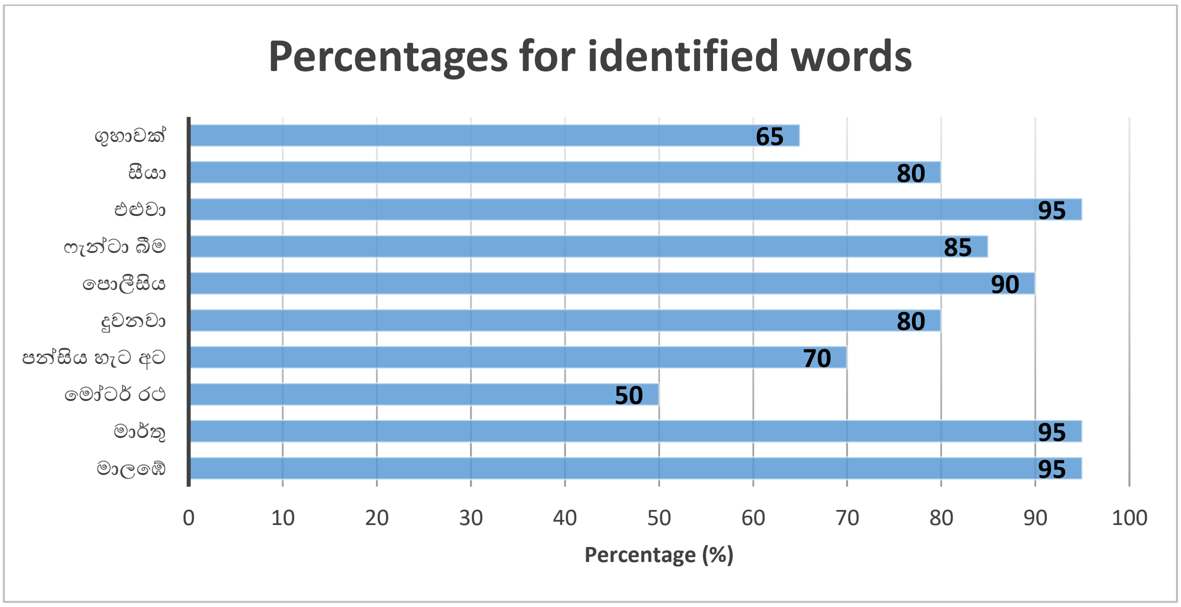


Figure 21: Bar chart for percentages of identified words

Following Figure 16 shows a bar chart drown to indicate the percentages of number of attempts taken to identify word by looking only at visemes. Two attempts are taken by 32.5% in the overall data set as the highest percentage. Then the 30% has been taken one attempt to identify the words and 20% has been taken three attempts for identification using visemes. Lowest percentages 5% is for the five and six attempts for the identification.

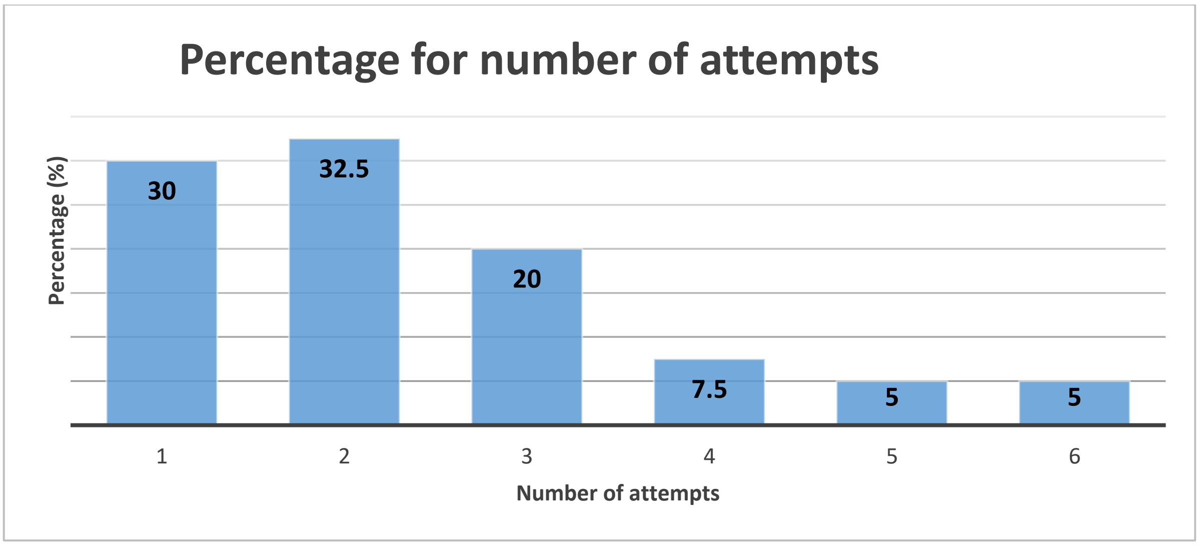


Figure 22: Bar chart for percentages of number of attempts

Overall higher percentages are taken by one, two and three attempts to identify the word without listening to the audio and most of the words are identified with more than 50% among 40 participants. None of the words are could not identified in the dataset.

4.2.3. Method 03

Evaluation method 03 is used to check whether the visual representation of the mouth movements gives the help to memorize the words rather than only hearing the audio. Therefore a questionnaire has been prepared with two sections and it is prepared using Google form.

In the first section of the questionnaire audio is given for the ten words and it should be listen at once. Then the participants has to select the words in the given lists which are memorized while listening the audio. That audio cannot be listen again. After that in the next section video is provided with the audio and different word order of previous word set. That video also can be watched at once and the last section is given the word set in different order to select the most memorable words of the user.

Following are the words given in evaluation method 3 in Table 21. It contains named entities, verbs and some words comes from English language.

Table 23: Words for Evaluation 03

Word	Sinhala text
W1	රෝදය
W2	මකුළුවා
W3	පූජාව
W4	රෙජිමේන්තු
W5	පොහොසත්
W6	බෞද්ධයෝ
W7	විදේශිකයන්
W8	ආශිර්වාද
W9	දේශපාලන
W10	අත්පොත

Prepared questionnaire has been shared with 40 participants to collect the feedback about the system. Following graph representing how the participants react based only for the audio and also audio with viseme representation for the word list.

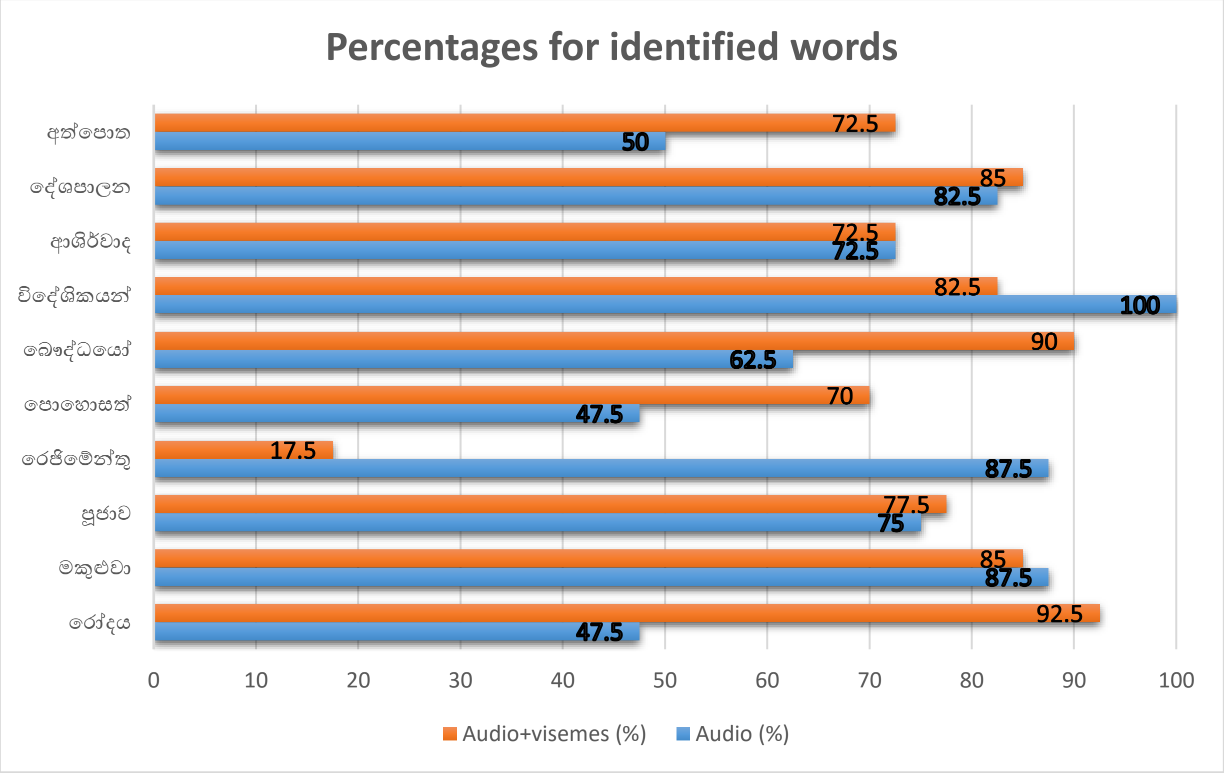


Figure 23: Bar chart for percentages of identified words

The percentages for both the audio alone and the audio with visemes are shown in the figure above for word identification. When the data is analyzed, seven words except words රෙජිමේන්තු, විදේශිකයන් and මකුළුවා stand out as having higher percentages for the audio with viseme video for word identification. When identifying words solely from audio, those seven words display lower values. Visemes being played along with audio has a greater impact on word recognition as listen.

Table 24: Words identified with higher percentage than audio

Words	Audio (%)	Audio and visemes (%)
රෝදය	47.5	92.5
පූජාව	75	77.5
පොහොසත්	47.5	70
බෞද්ධයෝ	62.5	90
ආශිර්වාද	72.5	72.5
දේශපාලන	82.5	85
අත්පොත	50	72.5

Another two words are showing lower percentages for the audio with visual representation, but it is closer to the percentage calculated in audio. The word “මකුළුවා” is showing 2.5 % difference between audio and audio with visemes and for the word “රෙජිමේන්තු” has the lowest percentage 17.5% for the audio with visemes that can be neglected because it is showing unusual identification with audio than the visemes with audio. That word should needs to be identified correctly in the second time with audio and video because the first phase it has been already correctly identified only with the audio.

4.3 Overall observation and Discussion

DNN Model Evaluation: Evaluation results under the section 4.1 higher accuracy and precision has been obtained for the model. Recall values calculated for macro average and weighted average is also having higher values for each. By considering the precisions and recall model has higher rate of identifying viseme labels and also the higher rate of correctly identified classes in the data set.

Evaluation method 01: Rating method has shown how the overall system satisfaction digits, words, short sentences and the long sentences. Based on Table 14 and Table 15, nine digits and five digits are having highest percentage for scale 5. Therefore, digits and words lip synchronization is more satisfied by the users because most of the highest percentages are available in the scale 5 that is satisfaction end. In other words green colored cell area which is showing highest rating is on the scale 5.

Then for the short sentences six sentences are having highest rating at scale 4, that means lip synchronization for short sentences are not at a highest satisfaction but the system gives output at good level. Then for the long sentences six sentences are having majority of the highest satisfaction at scale 3.

Therefore Sinhala letters and words are well synchronized with the system we have developed and short sentences also satisfied the lip synchronization. Long sentences are satisfied with average performance without going to the dissatisfaction of the synchronization. Overall the system is work well for the Sinhala digits, words, short sentences and long sentences with respect to the user satisfaction. No results are showing dissatisfaction for the lip synchronization in the system we have developed.

Evaluation method 02: Then the DNN model performance for viseme generation also work well based on the results obtained for the evaluation method 2. All the words have been identified with more than average rate of identification without the audio. There are no words are having lower percentages than 50% for the word identification.

When considering number of attempts have been taken to identify the words by looking at the visemes majority of participants has used one, two or three attempts. Therefore we can say within first three attempts user can identify the words without hearing the audio. Hence we can say the DNN model provides correct visemes for the given text.

Evaluation method 03: Lip movements are affected to identify the words while we are having discussion with someone. That feature has been tested with the evaluation 3 for the lip synchronized system and it is successfully showing better results by showing higher percentages for the words identification with visemes and audio playing together than listening the audio alone.

Considering the results of evaluation 03, seven words out of 10 has been shown higher or equal percentages for the memorizing word using visemes with audio. That is proved about how the lip movements along with the sound we are hearing help to identify and memorize the words we are hearing rather than only listening it. Therefore that results indicates that visemes which is the lip movement relates to the sound of a language gives a support to identify the word while the system is playing different words.

Overall three evaluations are proof that the lip synchronization is more satisfied for the Sinhala digits, words, short and long sentences while the DNN model producing correct and accurate visemes for the synchronization with the generated audio. Also the visemes are more helpful to identify and memorize the words in the system.

CHAPTER 05 CONCLUSION AND FUTURE WORK

This chapter concludes about the overall research solutions find for the research problems and objectives and other deliverables relates to the research study. Also the limitations and the future developments regarding the research study has been included to open up the new research areas in the fields of natural language processing, static viseme approaches and DNN for multi class classifications.

5.1 Conclusion

Lip synchronization is a process of moving the lip according to the audio to visualize the talking in a cartoon characters. Lip synchronization are highly affected to the liveness of the cartoon character with smoothness of the lip movements. Several studies have been done to different languages such as English, Portuguese, Korean etc.. By many research teams in the world. Currently the real time processing of cartoon characters also implemented for the broadcasting television programs in some countries.

In Sri Lanka lip synchronization models cannot be found for the Sinhala language. Only one research study done with static viseme rule based approach could be able to find for lip synchronization for Sinhala. But it also having issues with synchronization on long Sinhala sentences which are having more transitions between each word.

There are three research questions has been addressed at the beginning to find solutions through the research study. In next sub section all the research questions and solutions given from the study are stated.

5.1.1 Conclusion about the research problems

The first research question is to **find appropriate way to synchronize lip with text input given in different lengths**. As describe in the methodology section 3.2.2 time calculation has been done to find out the time allocation for one viseme by analyzing the time length of each audio which is produced for the different text input by the system. Also the set of visemes find by the DNN model has been played within the total time calculated for the audio. In the user interface there are three speed levels provided to see the outputs for the given text using different speed levels called fast, normal and slow. Those speed levels can be used to see whether the output is best synchronizes with which speed level.

Sinhala language rules and structures are highly considered while creating the data samples for alphabetical letters. Sinhala is a logical language and it has logics on producing different letters in the language and each letter representing a unique sound or a combination of two or more sounds. All those rules are studied and all the combinations of letters are taken to the data set.

As a solution for the second question that is about “**how to create training dataset by mapping phoneme to viseme?”** data set has been created for the Sinhala alphabetical letters with visemes label as a target. In this study out input is in the text format and the data set also needed to create for the text. Therefore the data set has been created for the alphabetical letters which are formed in language logics and rules with the target as viseme label along with each character. Because Sinhala alphabetical characters are representing sounds in the language. That sound does not change where the character has been placed. Therefore, direct mapping between letters and visemes are done to create the data set.

Then the third research problem is **about the latency that can synchronize the synthetic mouth according to the phoneme generated to the given input**. That latency has been calculated by considering the audio length and total number of set of visemes generated in the model. As mentioned in section 3.2.2 by calculating a time for a one viseme it is applied in the place where the visemes are playing in the system. Visemes will play within the total length of the audio generated.

5.1.2 Conclusion about the objectives and deliverables

This research has been developed a system to synchronize the lip movements with Sinhala text input given using static viseme approach with DNN model to overcome the issues happens in the previous study done with rule based approach. By going through the evaluation results overall system perform well to synchronize Sinhala digits, words, short and long sentences by generating correct viseme set via DNN model.

When considering the main objectives, those are; developing a DNN model to produce correct viseme labels for the given text input, identify the correct frame rate for synchronization between visemes and audio and develop a synthetic frontal view of the mouth for the Sinhala text synchronization has been successfully completed through this study.

Two deliverables of this research are delivering a DNN model for viseme selection and simple system implementation with frontal view of a synthetic mouth including three different speed levels to control. Also this system is not depend only one sentence structure such as written or spoken. Any type of sentence structure can be given to the system because the model has been trained for not for the sentences that is trained for single atomic letter in the language. DNN model development has reduced additional work on the rule based approach that is writing rules for all the language logics and structure.

When considering the evaluation of the DNN model it is showing better accuracy and lower loss for the given input. Therefore the DNN model is working fine for the viseme selection of the Sinhala letters given as an input.

Sinhala is a live language which has been added and removed some letters time to time. For instance letter “ෆ” comes to the Sinhala because we are using English words with Sinhala. Likewise if a new letter added we can easily add it to the data set with its variations ignoring the rules and structures of the language. Also the system uses text input then it avoid the processing of audio which uses most of the lip synchronization models as an input.

5.2 Limitations

This research has been done to develop a lip synchronization model for the Sinhala language using machine learning with static viseme approach. Our system will synchronizes lip movement with the audio generated for the input text. That audio has been generated using Google Text to speech translator and it works with internet. Therefore for the developed system should needs to have better internet connection to generate the audio.

Another limitation is when the input text size is more than 15 words that calculated time for the one viseme should needs to be adjusted to have a proper synchronization. Visemes will ends up earlier while the audio is playing in the system in normal speed rate. Audio is later start playing in the system. Therefore the system is not showing much better synchronization for the sentences with more than 15 words.

5.3 Further developments

In this research overall system has been developed for lip synchronization for Sinhala language using static viseme approach with DNN model for the viseme classification. Further developments can be addressed to the following areas of this research. Those are;

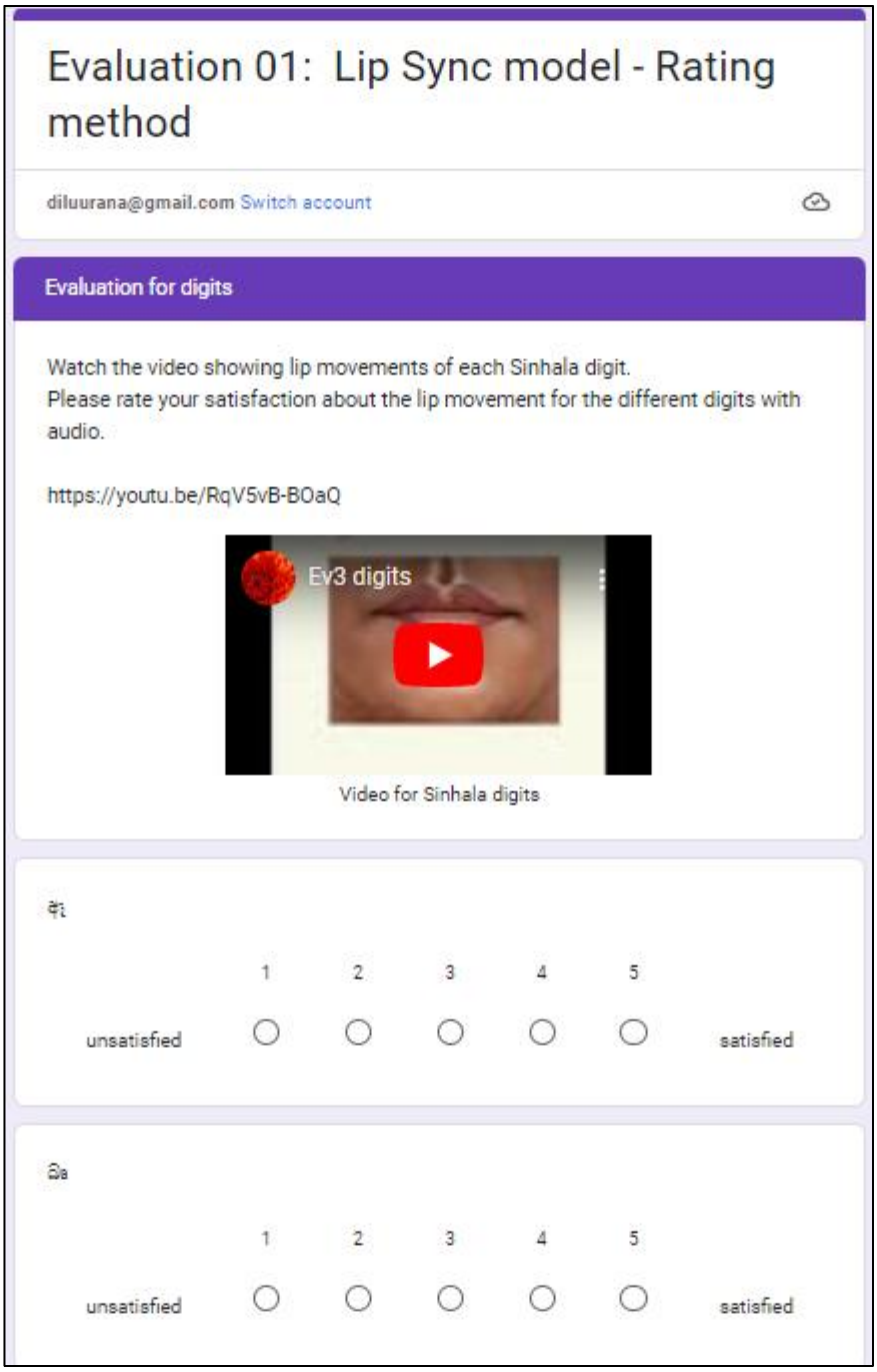
- this system can be further developed as a web application or mobile application for the users. Current development has been done as a desktop application.
- by using static viseme with machine learning technology we can enhance this system to view the side view of the face, tong movement and to showing how the air flow is going through vocal tract
- System implementation for the full face by showing the emotions based using text input.

In the text should be analyzed to find the emotional feature indications of the Sinhala language.

- Full face can be implemented for the text input to represents different types of facial expressions by analyzing the language properties belongs with facial expressions. Dynamic viseme approach is more suitable by processing video data set with different words with facial expressions.
- Full face for the Sinhala language can be developed using dynamic viseme approach by creating large video data set and using deep learning technique to extract the features in the video dataset.
- Real time cartoon character can be implemented for the Sinhala language for live broadcasting television shows.

APPENDICES

Appendix A: Questionnaire for Evaluation1



Evaluation 01: Lip Sync model - Rating method

diluurana@gmail.com Switch account

Evaluation for digits

Watch the video showing lip movements of each Sinhala digit.
Please rate your satisfaction about the lip movement for the different digits with audio.

https://youtu.be/RqV5vB-BOaQ

Video for Sinhala digits

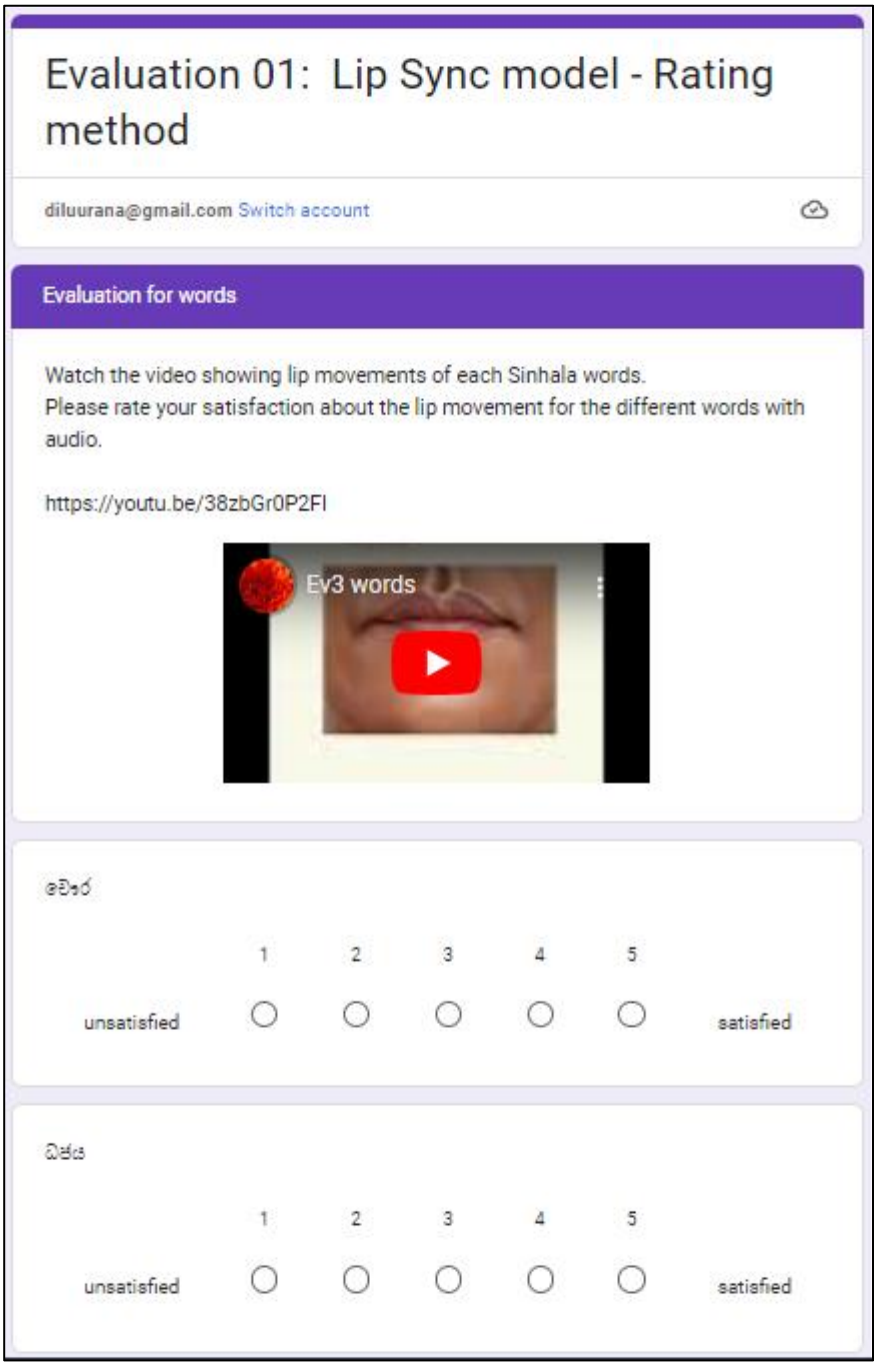
එක

		1	2	3	4	5	
unsatisfied		○	○	○	○	○	satisfied

දෙක

		1	2	3	4	5	
unsatisfied		○	○	○	○	○	satisfied

Figure A 1: Sample questions for Sinhala digits



Evaluation 01: Lip Sync model - Rating method

diluurana@gmail.com Switch account

Evaluation for words

Watch the video showing lip movements of each Sinhala words.
Please rate your satisfaction about the lip movement for the different words with audio.

https://youtu.be/38zbGr0P2FI

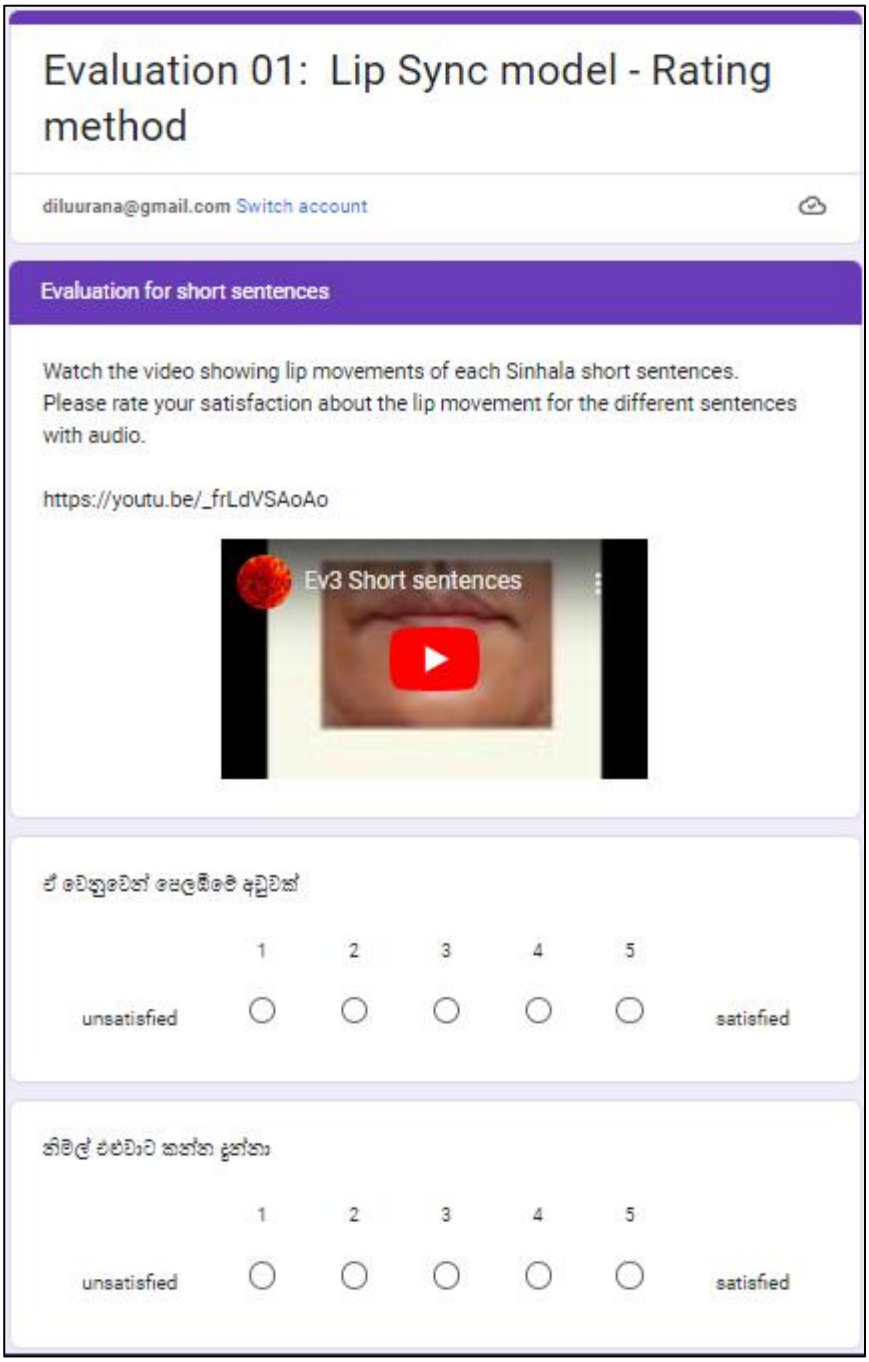
අවිෂර්

	1	2	3	4	5	
unsatisfied	○	○	○	○	○	satisfied

ධජය

	1	2	3	4	5	
unsatisfied	○	○	○	○	○	satisfied

Figure A 2: Sample questions for Sinhala words



Evaluation 01: Lip Sync model - Rating method

diluurana@gmail.com Switch account

Evaluation for short sentences

Watch the video showing lip movements of each Sinhala short sentences.
Please rate your satisfaction about the lip movement for the different sentences with audio.

https://youtu.be/_frLdVSAoAo

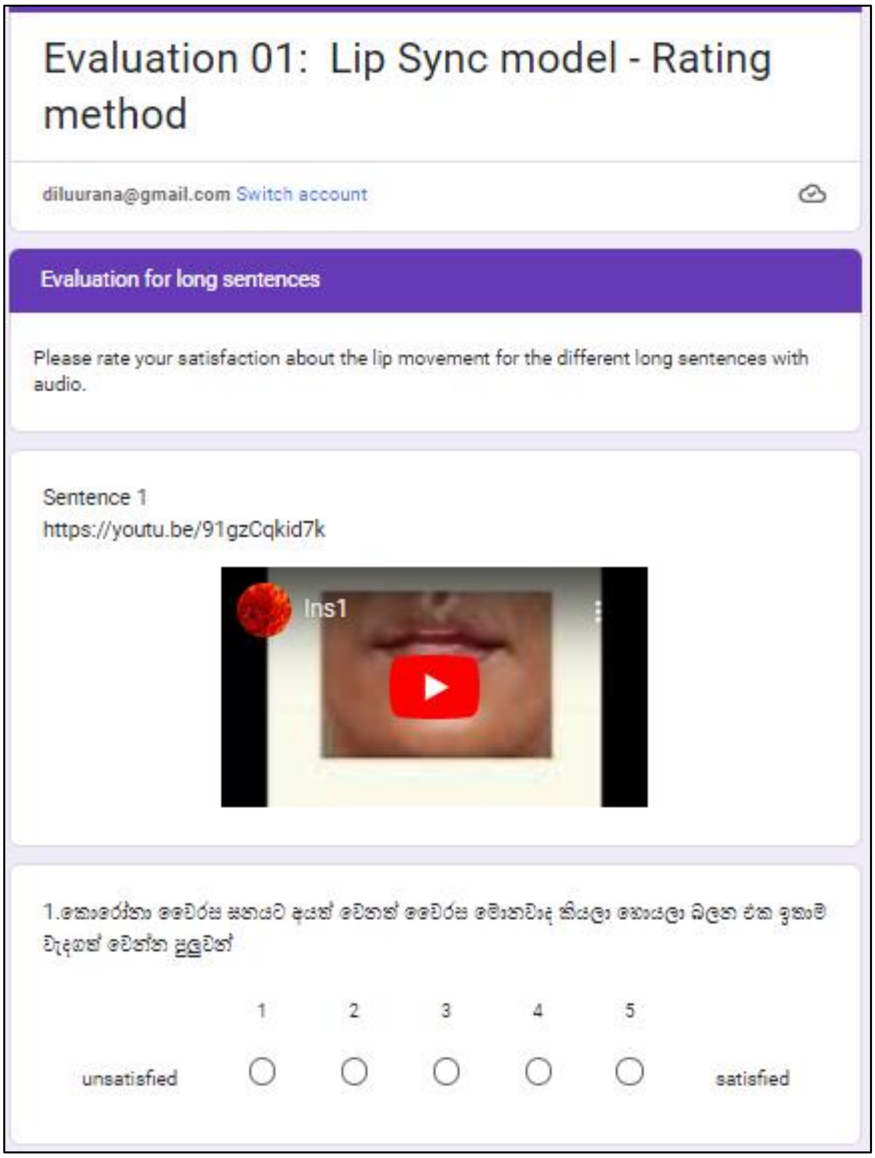
ඒ වෙනුවෙන් පෙලඹීමේ අඩුවක්

	1	2	3	4	5	
unsatisfied	○	○	○	○	○	satisfied

නිමල් එළුවාට කන්න දුන්නා

	1	2	3	4	5	
unsatisfied	○	○	○	○	○	satisfied

Figure A 3: Sample Questions for Sinhala short sentences



Evaluation 01: Lip Sync model - Rating method

diluurana@gmail.com Switch account

Evaluation for long sentences

Please rate your satisfaction about the lip movement for the different long sentences with audio.

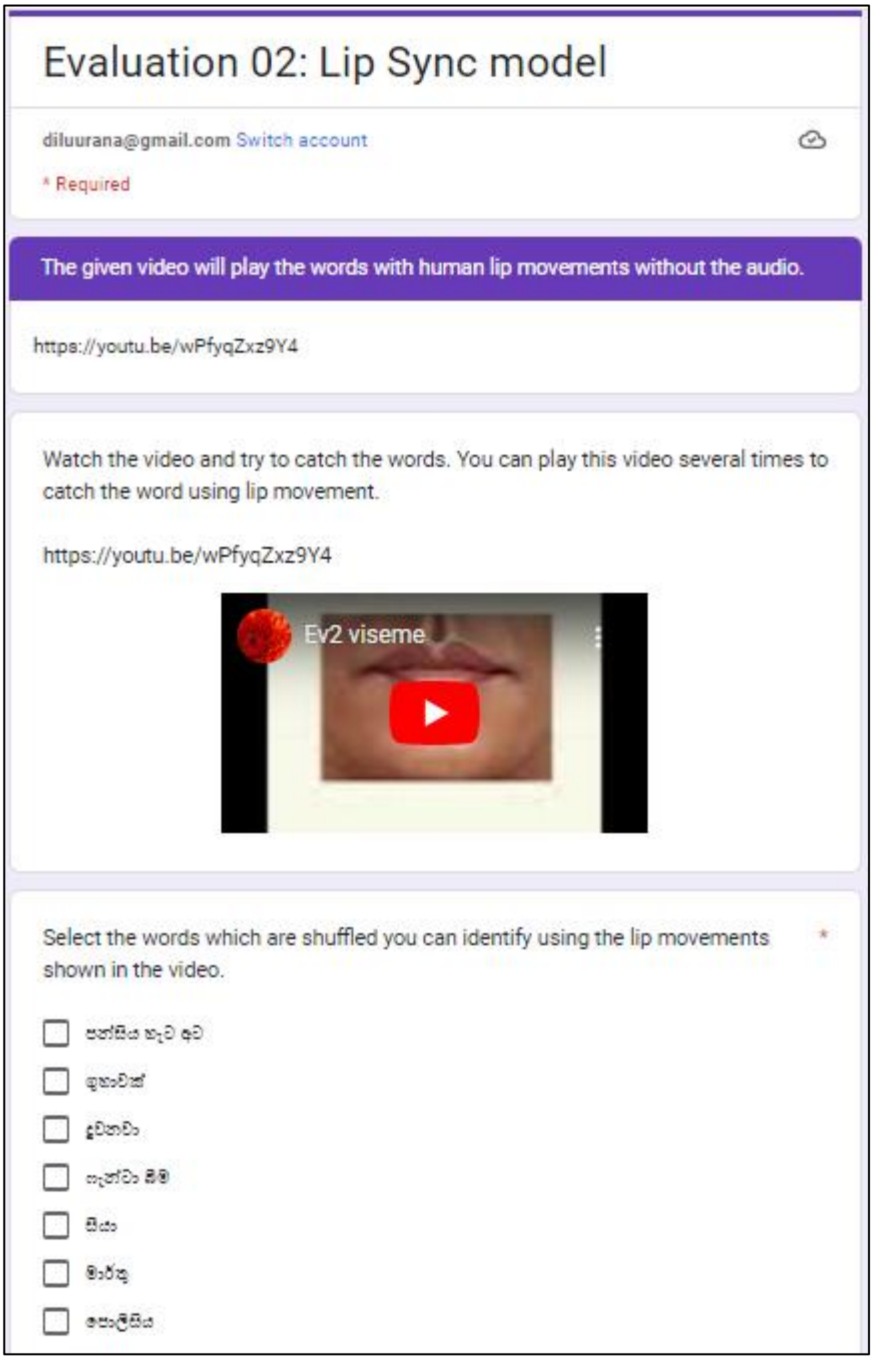
Sentence 1
https://youtu.be/91gzCqkid7k

1.කොරෝනා වෛරස සනයට අයත් වෙනත් වෛරස මොනවාද කියලා හොයලා බලන එක ඉතාම වැදගත් වෙන්න පුලුවන්

	1	2	3	4	5	
unsatisfied	○	○	○	○	○	satisfied

Figure A 4: Sample questions for Sinhala Long sentences

Appendix B: Questionnaire for Evaluation 2



Evaluation 02: Lip Sync model

diluurana@gmail.com Switch account

* Required

The given video will play the words with human lip movements without the audio.

https://youtu.be/wPfyqZxz9Y4

Watch the video and try to catch the words. You can play this video several times to catch the word using lip movement.

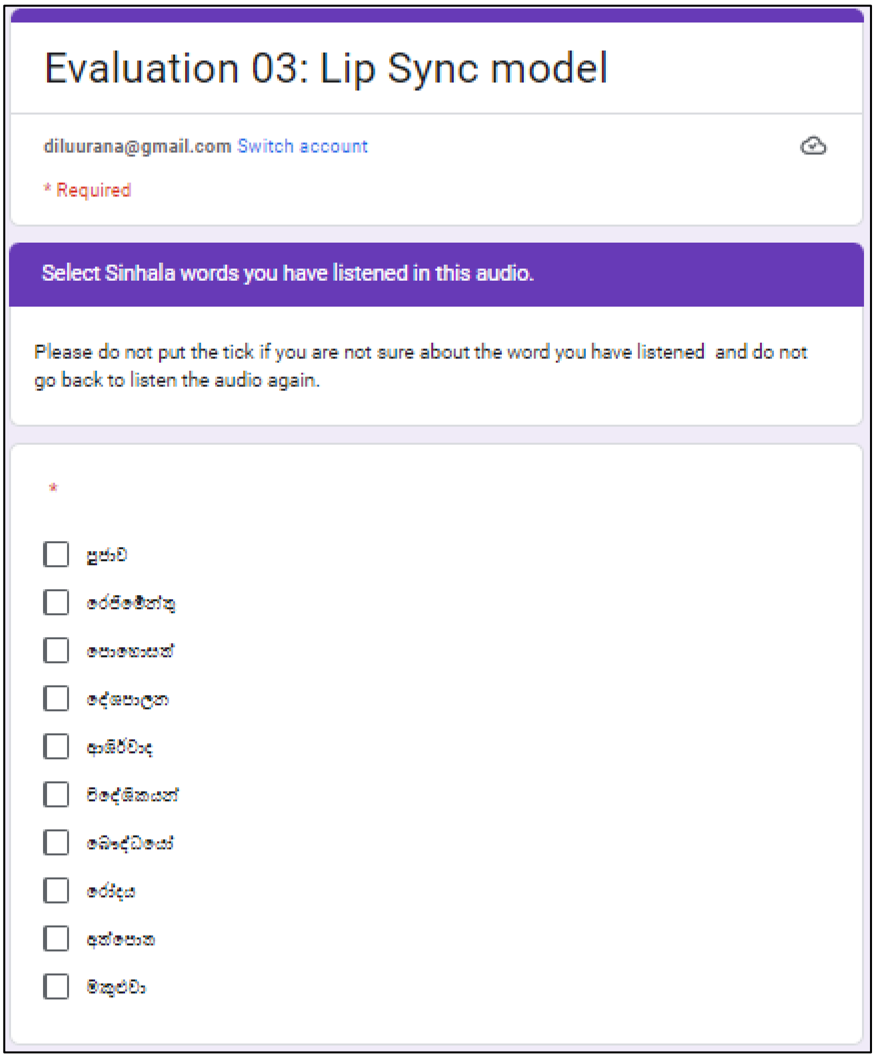
https://youtu.be/wPfyqZxz9Y4

Select the words which are shuffled you can identify using the lip movements shown in the video. *

- ☐ පන්සිය හැට අට
- ☐ ගුහාවක්
- ☐ දුවනවා
- ☐ හැන්වා සිම
- ☐ සියා
- ☐ මාර්තු
- ☐ පොලීසිය

Figure B 1: Part of a questionnaire of Evaluation 02

Appendix C: Questionnaire for Evaluation 3



Evaluation 03: Lip Sync model

diluurana@gmail.com Switch account

* Required

Select Sinhala words you have listened in this audio.

Please do not put the tick if you are not sure about the word you have listened and do not go back to listen the audio again.

*

□ පූජාව

□ රෙජිමේන්තු

□ පොහොසත්

□ දේශපාලන

□ ආශිර්වාද

□ විදේශිකයන්

□ බෞද්ධයෝ

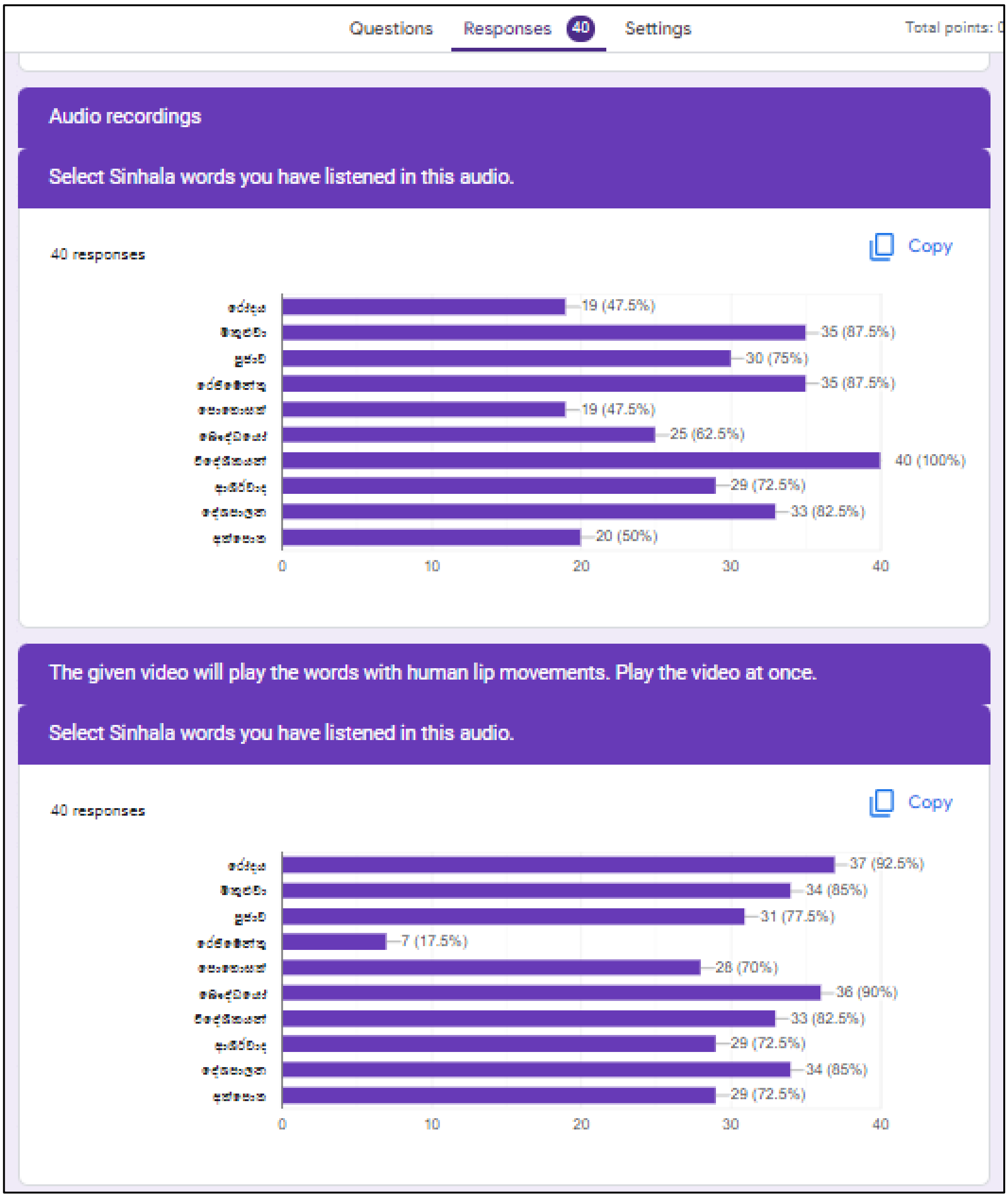
□ රෝදය

□ අත්පොත

□ මිතුරුවා

Figure C 1: Part of a questionnaire of Evaluation 03

Appendix D: Responses for Questionnaires



Audio recordings

Select Sinhala words you have listened in this audio.

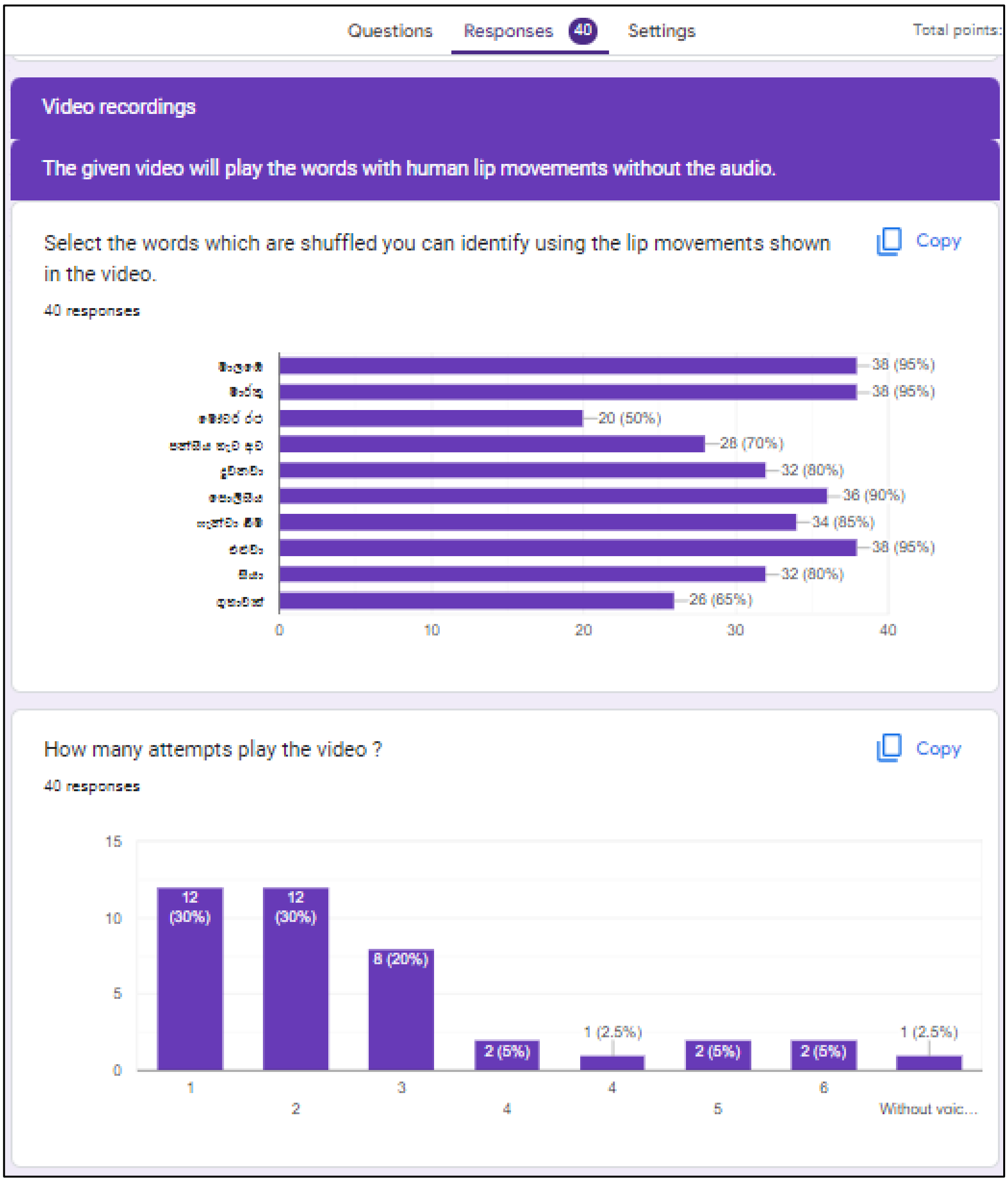
40 responses

The given video will play the words with human lip movements. Play the video at once.

Select Sinhala words you have listened in this audio.

40 responses

Figure D 1: Response for Evaluation 03



Video recordings

The given video will play the words with human lip movements without the audio.

Select the words which are shuffled you can identify using the lip movements shown in the video.

40 responses

How many attempts play the video ?

40 responses

Figure D 2: Sample Response for Evaluation 03

REFERENCES

Aneja, D., Li, W., 2019. Real-Time Lip Sync for Live 2D Animation. ArXiv191008685 Cs.

Bear, H.L., Harvey, R., 2019. Alternative Visual Units for an Optimized Phoneme-Based Lipreading System. Appl. Sci. 9, 3870. https://doi.org/10.3390/app9183870

Bhuiyan, M.I., 2021. A Deep Learning Approach to Learn Lip Sync from Audio 43.

Britto Mattos, A., Borges Oliveira, D.A., da silva Morais, E., 2018. Improving CNN-Based Viseme Recognition Using Synthetic Data, in: 2018 IEEE International Conference on Multimedia and Expo (ICME). Presented at the 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE, San Diego, CA, pp. 1–6. https://doi.org/10.1109/ICME.2018.8486470

Ivanko, D., Ryumin, D., Karpov, A., 2019. AUTOMATIC LIP-READING OF HEARING IMPAIRED PEOPLE. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLII-2/W12, 97–101. https://doi.org/10.5194/isprs-archives-XLII-2-W12-97-2019

K R, P., Mukhopadhyay, R., Philip, J., Jha, A., Namboodiri, V., Jawahar, C.V., 2019. Towards Automatic Face-to-Face Translation, in: Proceedings of the 27th ACM International Conference on Multimedia. Presented at the MM ’19: The 27th ACM International Conference on Multimedia, ACM, Nice France, pp. 1428–1436. https://doi.org/10.1145/3343031.3351066

Ko, H., Han, D.K., 2006. SVM-BASED PHONEME CLASSIFICATION AND LIP SHAPE REFINEMENT IN REAL-TIME LIP-SYNCH SYSTEM. Int. J. Pattern Recognit. Artif. Intell. 20, 1029–1051. https://doi.org/10.1142/S0218001406005113

Loh, N.H., 2014. Development of Real-Time Lip Sync Animation Framework Based On Viseme Human Speech. Arch. Des. Res. 112, 19. https://doi.org/10.15187/adr.2014.11.112.4.19

Nadungodage, T., Liyanage, C., Prerera, A., Pushpananda, R., Weerasinghe, R., 2018. Sinhala G2P Conversion for Speech Processing, in: 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018). Presented at the 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018), ISCA, pp. 112–116. https://doi.org/10.21437/SLTU.2018-24

Serra, J., Ribeiro, M., Freitas, J., Orvalho, V., Dias, M.S., 2012. A Proposal for a Visual Speech Animation System for European Portuguese, in: Torre Toledano, D., Ortega Giménez, A., Teixeira, A., González Rodríguez, J., Hernández Gómez, L., San Segundo Hernández, R., Ramos Castro, D. (Eds.), Advances in Speech and Language Technologies for Iberian Languages, Communications in Computer and Information Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 267–276. https://doi.org/10.1007/978-3-642-35292-8_28

Shrestha, K., 2019. Lip Reading using Neural Network and Deep learning 7.

Son, J.S., Zisserman, A., 2017. Lip Reading in Profile, in: Procedings of the British Machine Vision Conference 2017. Presented at the British Machine Vision Conference 2017, British Machine Vision Association, London, UK, p. 155. https://doi.org/10.5244/C.31.155

Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I., 2017. Synthesizing Obama: learning lip sync from audio. ACM Trans. Graph. 36, 1–13. https://doi.org/10.1145/3072959.3073640

Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A.G., Hodgins, J., Matthews, I., 2017. A deep learning approach for generalized speech animation. ACM Trans. Graph. 36, 1–11. https://doi.org/10.1145/3072959.3073699

Thangthai, A., Milner, B., Taylor, S., 2019. Synthesising visual speech using dynamic visemes and deep learning architectures. Comput. Speech Lang. 55, 101–119. https://doi.org/10.1016/j.csl.2018.11.003

Weerathunga, C., Weerasinghe, R., Sandaruwan, D., 2020. Lip Synchronization Modeling for Sinhala Speech, in: 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer). Presented at the 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), IEEE, Colombo, Sri Lanka, pp. 208–213. https://doi.org/10.1109/ICTer51097.2020.9325489

Xu, Y., Feng, A.W., Marsella, S., Shapiro, A., 2013. A Practical and Configurable Lip Sync Method for Games, in: Proceedings of Motion on Games. Presented at the MIG ’13: Motion in Games, ACM, Dublin 2 Ireland, pp. 131–140. https://doi.org/10.1145/2522628.2522904