Lip Synchronization Modeling for Sinhala Speech

W.A.C.J.Weerathunga

2015/CS/145

This dissertation is submitted to the University of Colombo School of Computing
In partial fulfillment of the requirements for the
Degree of Bachelor of Science Honours in Computer Science

University of Colombo School of Computing
35, Reid Avenue, Colombo 07,
Sri Lanka

July, 2020

Declaration

I, W.A.C.J. Weerathunga (2015/CS/145) hereby certify that this dissertation entitled "Lip Synchronization Modeling for Sinhala Speech" is entirely my own work and it has never been submitted nor is currently been submitted for any other degree.

....

Date Signature of the Student

I, DR A.R. Weerasinghe, certify that I supervised this dissertation entitled "Lip Synchronization Modeling for Sinhala Speech" conducted by W.A.C.J. Weerathunga in partial fulfillment of the requirements for the degree of Bachelor of Science Honours in Computer Science.

....

Date Signature of the Supervisor

I, DR K.D. Sandaruwan, certify that I co-supervised this dissertation entitled "Lip Synchronization Modeling for Sinhala Speech" conducted by W.A.C.J. Weerathunga in partial fulfillment of the requirements for the degree of Bachelor of Science Honours in Computer Science.

....

Date Signature of the Co-Supervisor

Abstract

Lip synchronization also known as visual speech animation, is the process of matching the speech with the lip movements. Visual speech animation has a great impact in the gaming and animation film industry, due to the reason that it provides the realistic experience to the users. Furthermore, this technology also supports better communication for deaf people.

For most of the European languages, lip synchronizing models have been developed and used vastly in the entertainment industries. However, there are still no research experiments conducted towards the speech animation of the Sinhala language. This is due to the Less contribution towards research development and unavailability of resources.

This research is focused on the problem of achieving a lip synchronization model for the Sinhala language. This project presents a study on how to map from acoustic speech to visual speech with the goal of generating perceptually natural speech animation.

Initially, this study follows a deep learning approach and terminates due to not having enough video data to achieve a good performance. Next, the experiments on developing a visemes alphabet is carried out using a static visemes approach on a video data set created by the author.

The implemented lip synchronizing model was evaluated using a subjective evaluation based on six different categories. The model achieved a 69% accuracy for the subjective evaluation using the static visemes approach. As an initiative research on speech animation for the Sinhala language, this model accurately animates individual words rather than long sentences.

Preface

This document has been written for the partial fulfillment of the requirements of the B.Sc. in Computer Science (Hons) Final Year Project in Computer Science(SCS4124). I was engaged in researching and writing this dissertation from January 2019 to February 2020.

The basis for this research idea came from my supervisor Dr. Ruwan Weerasinghe. Implementing a lip synchronization model is a somewhat different task when compared with other Natural Language Processing researches. However, it will be a good research contribution to the researching community, since no one has attempted to implement a lip syncing model for the Sinhala language before. The lack of resources of the Sinhala language increases the difficulty level of the research.

In this research work, I was able to derive a viseme alphabet for the Sinhala language which is a much-needed resource to develop speech animation using the static visemems approach.

Acknowledgement

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. K.D. Sandaruwan, lecturer of University of Colombo School of Computing for providing me the continuous guidance and supervision throughout the research.

I would also like to extend my sincere gratitude to Dr. G.D.S.P.Wimalaratne, senior lecturer of University of Colombo School of Computing and Mr. W.V Welgama, senior lecturer of University of Colombo School of Computing for providing feedback on my research proposal, interim evaluation and pre-final defence to improve my study. I also take the opportunity to acknowledge the assistance provided by Dr.H.E.M.H.B. Ekanayake the final year computer science project coordinator.

I appreciate the feedback and motivation provided by my friends to achieve my research goals. This thesis is also dedicated to my loving family who has been an immense support to me throughout this journey of life. It is a great pleasure for me to acknowledge the assistance and contribution of all the people who helped me to successfully complete my research.

Contents

List of Figures

List of Tables

Acronyms

CNN	Convolutional Neural Network
DAVS	Disentangled Audio-Visual System
LSTM	Long short-term memory
LTRL	Language Technology Research Laboratory
PCA	Principal Component Analysis
RNN	Recurrent Neural Network
NLP	Natural Language Processing
UCSC	University of Colombo School of Computing

Chapter 1

Introduction

1.1 Background to the Research

Communication is one of the most important aspects of human evaluation. In human communication, various things are involved such as face, body, voice and social states. Among them face is a complex surface and very important communication channel. Face can express emotions such as happiness, anger, sadness and secondly in verbal communication it can express information. Besides, the third group of facial cues accompany acoustic speech as visible speech and carry information about phonetic content. Moreover there are different types of facial cues carrying out different communication tasks. Visible speech is one of them and it contains the visual information about speech. It improves the intelligibility significantly, especially in noisy environments as the movements of the lips can compensate for a possible loss in the speech signal. Furthermore, the visual component of speech plays a key role for hearing impaired people.

The use of multiple sources such as acoustic and visual, generally enhances speech perception and understanding. The process of producing an animation of mouth that is synchronized with speech input which is also known as 'Lip Synchronization' is one of the main research topics in this area. There are lots of talking heads that have been invented in different languages such as English, French, Arabic, Indonesian and Chinese (Rong et al., 2012, Setyati et al., 2017). This project is centered on speech directed lip-synching and building a talking head for the Sinhala language.

With the improvement of technologies and computing power, visualizing speech with lips synchronization plays a significant role nowadays and it has directed into new fields such as education, transportation, cognitive education, entertainment industry, film industry, etc.For instance,hearing-impaired people can benefit from synthetically generated talking faces by means of visual speech. Lip reading tutors for hearing-impaired, or language tutors for the children who have difficulties in speaking can be prepared (Rodríguez-Ortiz, 2008*a*). In addition to that, videophones can be produced to make possible the distant communication of the deaf people (Power et al., 2007).

1.2 Motivation

Due to the reason that lip synchronization makes characters more realistic, many industries such as the gaming industry, animation film industry tend to use lip synchronization for the animated characters to provide a realistic experience for the users. Motion capturing (M, 2018) is also another technique that maps the motion of speaking to animation models. However, they suffer from limitations such as high cost. Using a simple lip synchronization model for the same task would reduce the cost.

It has been observed in a research study (Rodríguez-Ortiz, 2008*b*), that people who are not deaf by birth can read lips properly. Concerning this fact, lip synchronization models can be used as a solution for converting the audio waves into the lip movements, thus helping deaf people to understand the audio waves. Furthermore talking heads leads to have a great potential for interactive applications,where user Interface agents can be developed to be employed in e-learning, web navigation or as virtual secretary (Bonamico and Lavagetto, 2001). Interactive computer games with talking faces of virtual characters can be produced. Furthermore, synthetic talking heads can be used as avatars;animated representations of users in virtual meeting rooms, or for low bandwidth video tele-conferencing (Eisert, 2003, Ostermann, 1998).

1.3 Research Problem and Research Questions

To achieve the above-mentioned goal of implementing the lip synchronization model, several research questions are being addressed in the research and they are listed below.

1. What is the visemes alphabet for Sinhala language to generate speech animation ?

   A viseme is a generic facial image that can be used to describe a particular sound and it is the visual equivalent of a phoneme or unit of sound in spoken language. Using visemes, the hearing-impaired can view sounds visually - effectively, "lip-reading" the entire human face. Visemes alphabet is derived from the phoneme alphabet. Some researches have succeeded in deriving their own visemes alphabet to languages such as English, French, Arabic, Indonesian and Chinese (Dehshibi and Shanbezadeh, 2014, Rong et al., 2012).

   However, for the Sinhala language there are no researches have been done so far. Thus,deriving our own viseme alphabet for Sinhala language will be one of the main research questions addressing in this research.Deriving a vismes alphabet needs a deep analysis of the phoneme set for that particular language so as to find a suitable approach.

2. What is the proper way to find visemes sequence from the Sinhala text?

   After deriving the viseme alphabet, the research work should have find a way to encode the Sinhala text into the visems sequence, to generate the visual speech animation.

3. What is the proper way to generate the speech animation?

   For a better visualization, the visemes sequence needs to be animated properly using a 3D human model after generating the visemes sequence.

4. What is the accuracy of the Sinhala speech animation that can be achieved through the model?

   As the final phase, an evaluation will be carried out to find out the accuracy gain by the lip synchronization model.

1.4 Justification for the research

The Sinhala language is a low resourced language since the Sinhala speaking population is very low when compared with the global population. Due to the reason of the less contribution towards the improvement of research areas, there are no research attempts have been taken to implement a Sinhala talking head while the other languages like English, Spanish, Chinese (Rong et al., 2012) has implemented talking heads using various approaches refer .

Deep learning is the state-of-the-art approach to generate the speech animation for the English language while there are other approaches which are static visemes approach (Mattheyses et al., 2013) and dynamic visemes approach (Taylor et al., 2012). These deep learning approaches require large speech datasets and tools for labeling. For English and Chinese languages, there are datasets such as voxlab,LRS3-TED (Afouras et al., 2018) that are readily available in the field. However, there does not exist any such labeled speech video data set for Sinhala language. Therefore, a sinhala speech video dataset along with a derived visemes alphabet for Sinhala language will be contributed to the research community by this research.The research also intends to experiment on both deep learning and static visemes approach at the beginning for finding a better approach for the data set.

1.5 Methodology

There are several approaches for developing a lip synchronization animation model such as static visemes approach (Mattheyses et al., 2013), dynamic approach (Taylor et al., 2012) and deep learning approach(Taylor et al., 2017). Among them, a deep learning approach is the state of the art approach and it shows a better accuracy rather than other approaches for the English language. For the deep learning approach, it requires a large data set which is approximately 10 hours. (Taylor et al., 2017, 2012). Although there are existing standard datasets available for the English language. For the Sinhala language, there are no datasets that have been created regarding visual speech. Since the unavailability of the data and other resources, this research work is conducted using the static visemes approach.

In static visemes approach, it requires a visemes alphabet for the language to derive the speech animation sequence. To derive the visemes alphabet for the Sinhala language this research work has used the combination of two approaches which are K means clustering approach and the subjective analysis.

1.6 Outline of the Dissertation

In the Chapter 2, a comprehensive study in describing technologies and performances observed by authors in previous researches will be discovered.The research design together with the high level architectures for addressing the research questions is presented in the Chapter 3. A comprehensive explanation on the implementation of research designs is carried out in Chapter 4. Then the experiments and results observed throughout the research period is evaluated in Chapter 5 descriptively. Finally, Conclusion and future work are discussed in Chapter 6.

1.7 Delimitations of Scope

The below listed are not covered in this research.

1. Modeling the facial emotions when speaking
2. The changes occur in the mouth shapes due to stress and rhythms of the speech sound

1.8 Conclusion

This chapter present the are of the research study and the questions that are addressed by this research. The research was justified by analyzing the significance of the research and outlined the dissertation.

Chapter 2

Literature Review

Industrial standard audiovisual speech animation often created by manually a skilled animator or actor with motion capturing technologies (Beeler et al., 2011, Cao et al., 2015). Although this approach has some advantages such as an artist can precisely style and time the animation in the manual animator of motion capturing, the problem is, it is extremely costly and time-consuming to produce. Therefore, this approach can be applied to automatically produce a production-quality animated speech for any character when given an audio or text script as input.

Several studies have been conducted on the filed of lip synchronization for the languages;English, Arabian, Malaysian, etc. The base point of earlier approaches of visualizing speech information is visemes which is the visual representation of phonemes. Visemes are based on the phonemes while phonemes are depended on the language since each language has different phoneme alphabets.

Traditionally the relationship between the phonemes and visemes are static and many to one where one viseme can have several phonemes. Two main approaches that are massively used for grouping the visems are subjective assessment with human viewers (Montgomery and Jackson, 1983) and the data-driven approach (Hazen et al., 2004). In subjective assessment, phonemes are mapped to vismes according to the user responses. However, limitations of subjective assessment necessitate that the stimuli be simple, so phonemes are presented in the context of isolated mono- or bi-syllabic words.In this piece of research, authors have used clustering approaches. For the Persian visemes (Dehshibi and Shanbezadeh, 2014),

they have used the hierarchical clustering method (Aghaahmadi et al., 2013) to derive the visemes while this work (Rong et al., 2012) has used fuzzy clustering and gray relation analysis to derive the visemes groups.When compared with the subjective assessment data-driven approach provides much better results with handling the effects to some level. The following figure shows some results of visemes grouping.

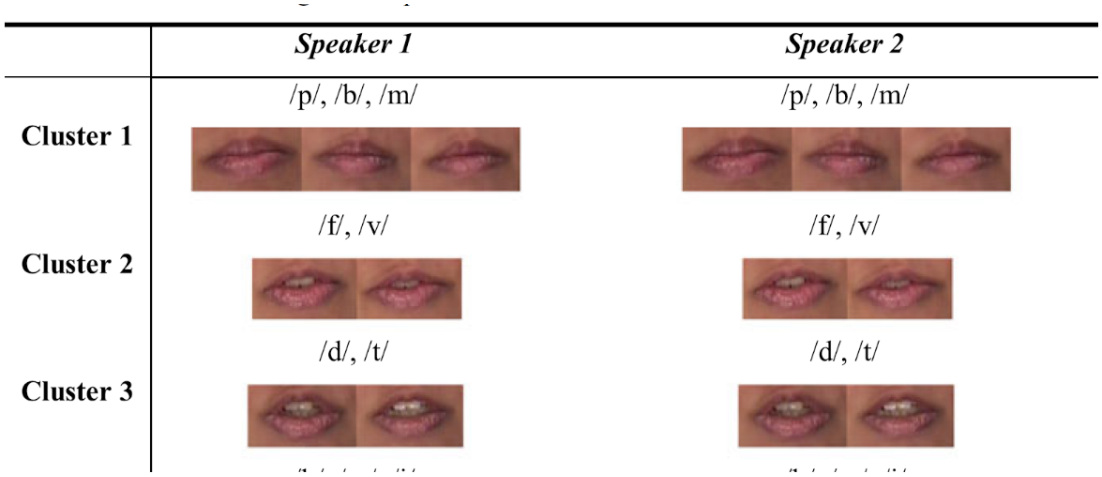

Figure 2.1: Example visemes groups corresponding to phonemes

The goal of speech animation is to present the correct articulately dynamics on a face model. Since a decade of different approaches have been used by researchers with different technologies. According to the Taylor (Taylor et al., 2017) generally, the prior work of automatic speech animation can be categorized into three broad classes: interpolating single-frame visual units, concatenating segments of existing visual data, and sampling generative statistical models.

Single-frame visual unit interpolation involves key-framing static target poses in a sequence and interpolating between them to generate intermediate animation frames (Ezzat et al., 2004). One benefit of this approach is that only a small number of shapes (e.g. one per phoneme) need to be defined. However, the realism of the animation is highly dependent on how well the interpolation captures both visual co articulation and dynamics. When interpolating one can either hand-craft such interpolation functions, but it is time-consuming to refine and ad-hoc or employ a data-driven approach based on statistics of visual speech parameters (Ezzat et al.,

2004). These approaches make strong assumptions regarding the static nature of the interpolant and do not address context-dependent coarticulation. This issue is partially considered in (Ezzat et al., 2004), which uses covariance matrices to define how much a particular lip shape is allowed to deform, but the covariance matrices themselves are fixed which can lead to unnatural deformations.

Research work done by grerri (Mattheyses et al., 2013) has used a short sequence of static length speech data to animate the speech by mapping phonemes to static visemes groups. The issue of this approach is it needs another process to handle co articulations by considering language rules which depend on the selected language. There are no pre-defined rules for languages and since different researches use different rules to handle the coarticulation and the output depends on the language rules.

Instead of using static phoneme and visemes mapping Taylor l (Taylor et al., 2012) has introduced a unit called dynamic visemes which are visems units. The coarticulation issue was handled by this method and provides much better results when comparing the static phoneme viseme approach. For feature extraction, they have used an Active appearance model with multi-segmented appearance parameters and 32 landmarks annotated. But one limitation is that the context typically considers only the phoneme identity, and so a large amount of data is required to ensure sufficient coverage overall contexts. Sample-based animation is also limited in that it can only output units seen in the training data. The below figure shows the dynamic visemes and the way they are combined to derive the words.

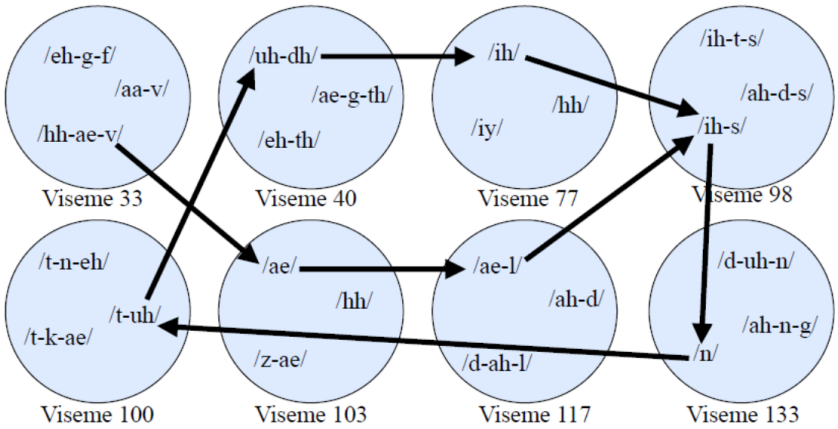

Figure 2.2: Dynamic visemes and combining them to derive the words

A more flexible and latest approach is to use a generative statistical model, such as GMMs (Luo et al., 2014), switching linear dynamical systems switching shared Gaussian process dynamical models (Deena et al., 2010), recurrent neural networks (Fan et al., 2015) or hidden Markov models (HMMs) and their variants (Anderson et al., 2013) or deep learning approaches.

Research work done by Taylor (Taylor et al., 2017) has introduced an effective and well-performed deep learning approach to automatically generate natural-looking speech animation that synchronizes to input speech. Moreover, they have used a sliding window predictor that accurately captures the coarticulation habits and natural motion of the language and arbitrary nonlinear mapping between phonemes label inputs and visemes. In the animation of visemes sequence, they have mainly pose 8 shapes on character rig and create the speech animation linearly combining the rig shapes. Moreover, this approach has used a fully connected network to generate the model and has provided more accurate and efficient results when compared with the previous approaches. Figure 2.3 shows the pipeline of this approach.

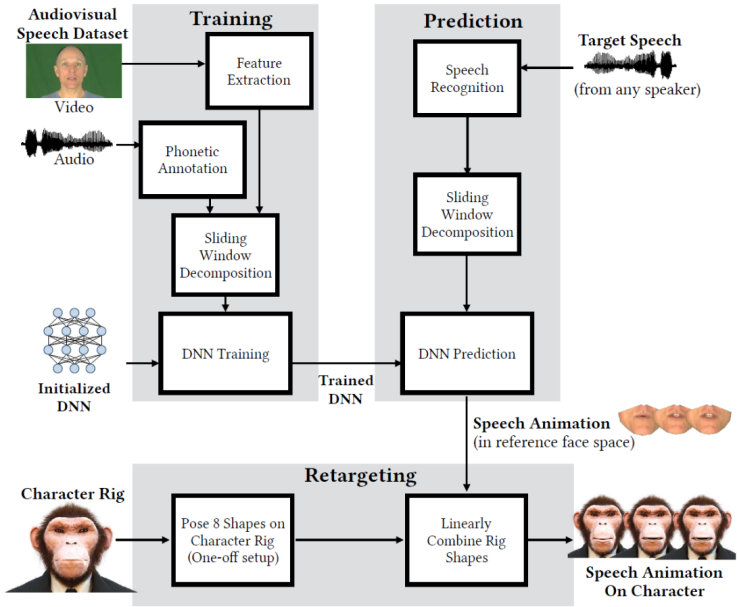

Figure 2.3: An overview of Taylors system

A research study made by yang chou (Zhou, Xu, Landreth, Kalogerakis, Maji and Singh, 2018) has presented a novel deep learning-based approach to produce

the animations directly from the input audio, deriving JALI or FACS based production face rig. In this approach, they have used three-stage Long Short-Term Memory (LSTM) network architecture since the motivation of psycho-linguistic insights (segmenting speech audio into a stream of phonetic groups)are sufficient for viseme construction. Taylors (Taylor et al., 2017) approach which is mentioned earlier needs input text transcripts and there is no compact mapping from audio to animator-centric viseme curves or facial action units. Additionally, this approach has the ability to overlay speech output with facial expressions. Editing or Refining animation of the spoken context or it's style is harder in this approach when compared to the taylors (Taylor et al., 2017) approach. This network is trained to map audio to visemes animation which is both sparse and is low dimensional. Moreover, viseme curves are more directly related to the speech rather than Taylor (Taylor et al., 2017) which has vismes agnostic mouth shape parameterization.

Research work of "you said that" by amir (Jamaludin et al., 2019) has proposed a method to generate videos of a talking face using only an audio speech segment and face images of the target identity. This approach is different from others since their output is not an animation, but a face image sequence of a target person. The speech segment need not be spoken originally by the target person. The authors have developed an encoder-decoder convolutional network (CNN) model which uses a joint embedding of the face and audio to generate video frames of the talking face. They have trained their model using cross-modal self-supervision. Moreover, the authors have successfully proposed an approach to re-dub videos.

The research work done by Hang Zhou (Zhou, Liu, Liu, Luo and Wang, 2018) has proposed Disentangled Audio-Visual System (DAVS), an end-to-end trainable network for talking face generation by learning disentangled audio-visual representations. They have specifically trained their network for different people using a three encoder network to train the Disentangled Audio-Visual System. The outputs generated are as same as amirs (Jamaludin et al., 2019) work. Figure 2.4 shows their output.

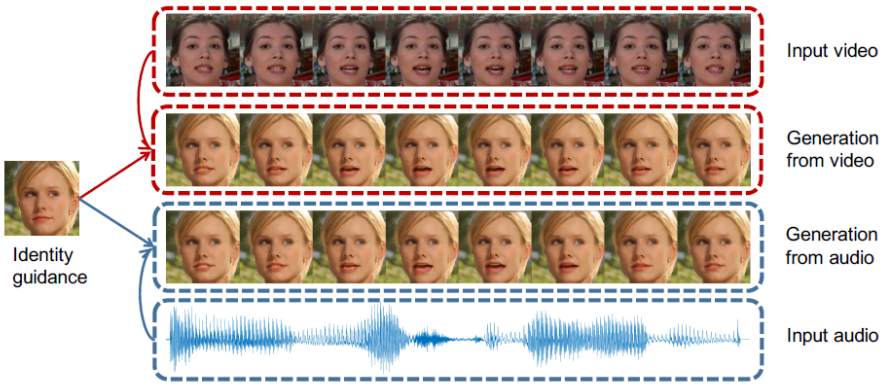

Figure 2.4: Overview of hangs work

2.1 Summary

As a summary of this chapter, the following points can be highlighted.

1. Research has been conducted on speech animation for languages such as English, French, chinese and Persian languages. However, there is no any available research for Sinhala speech animation

2. Mainly researchers have used three different approaches which are static visemes , dynamic visemes and deep learning approaches to generate the speech animation.

3. The dynamic visemes approach provides more accurate result than the static visemes approach and deep learning approach provides the best result among all the three approaches.

4. The state of the art approach is to generate a speech animation model from a deep learning approach.

5. For the static visemes approach, it needs to find a visemes alphabet for the language as the initial step. The two main approaches to derive visemes classes are subjective analysis and clustering approach.

Chapter 3

Research Design

In this chapter, two research designs that are implemented to solve the above mentioned research problems will be addressed in detail. The first approach is based on deep learning techniques while the second approach is based on the static visemes alphabet.

- Deep learning approach
- Static visemes approach

For the deep learning approach, it requires a large amount of data of visual speech to get the better performances. The specialty of static visemes approach is that it can derive the visual mouth shapes to corresponding sounds without training data and by only using the consonant vowel combinations.

3.1 Deep learning approach

The high-level architecture relevant to the deep learning approach is depicted in Figure 3.1 .

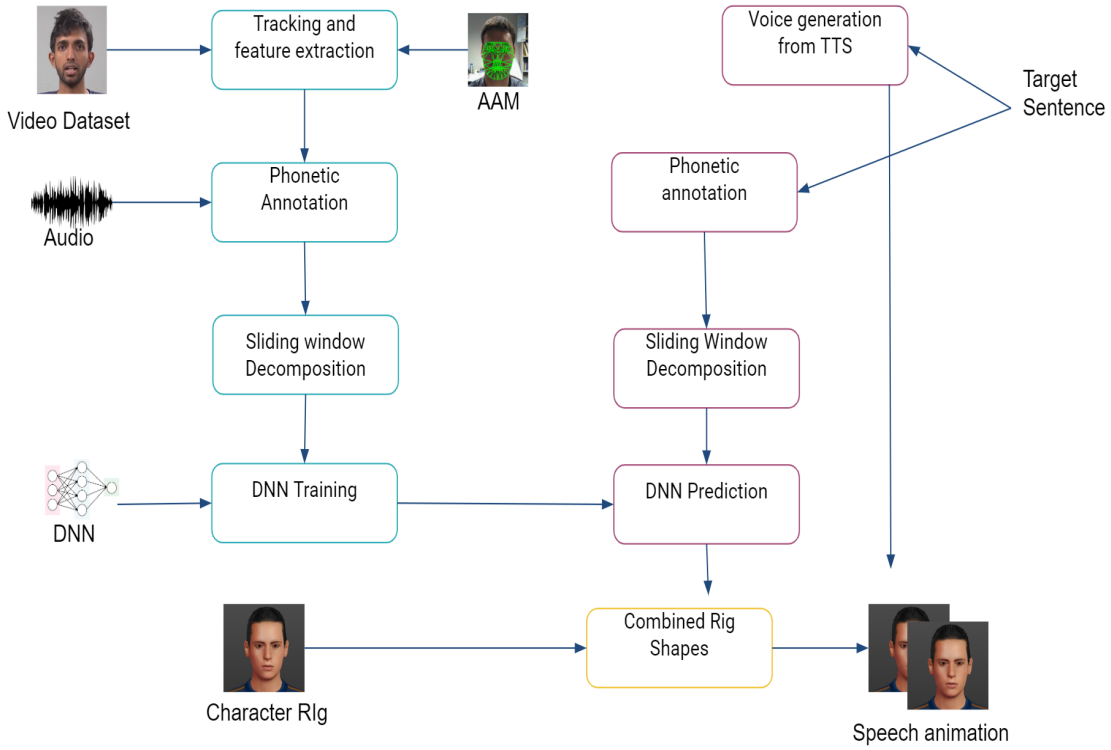

Figure 3.1: High-level architecture of deep learning approach

Above research design can be categorized into three main phases; data set creation, training and retargeting and these will be discussed in the following sections.

3.1.1 Generation of speech corpus

The data set is drawn from an audiovisual corpus containing an actor reciting 2000 distinct Sinhala sentences from the UCSC LTRL sentences collection in a neutral speaking style under the control environment. Videos are recorded to capture the frontal view of the face at 25 frames per second at a resolution of 1280 by 720 progressive scan and the total data set runs to approximately 10 hours.

3.1.2 Visual speech prediction model Generation

A deep learning approach is used to automatically generate natural-looking speech animation that synchronizes to input speech. This approach uses a sliding window predictor that learns arbitrary nonlinear mappings from phoneme label input sequences to mouth movements in a way that accurately captures natural motion and visual coarticulation effects.

3.1.3 Tracking and feature extraction

For this task, an Active Appearance Model is used for feature extraction of the face images. Active Appearance Models (AAMs) (Cootes et al., 2001) provide a means for tracking the speech articulators in a video. The shape of an AAM is defined by the two-dimensional vertex locations of a mesh that delineates the inner and outer lip contours and the jaw. Then the model is built by hand-labeling a small number of training images with the vertices that define the mesh. These training meshes are then normalized for similarity.

The appearance of an AAM is defined over the pixels within the shape mesh, $x = (x; y)T2s0$ and is constructed by warping each training image to the mean shape and then applying PCA to give a compact linear model of appearance variation.

Here, the inverse compositional project-out AAM algorithm Matthews and Baker (2004) is used to track the facial features. For analysis, rather than building an AAM with a single appearance component.

3.1.4 Phonetic Annotation

The video data and the sentence list is aligned correctly and the phonetics are annotated to feed the deep learning neural network. Here, Sinhala letters are annotated with the alphabet.

3.1.5 Deep learning sliding window Approach

The sliding window approach which is used in this research is inspired by [Kim et al. 2015]. Unlike other approaches, this model can handle coarticulation internally. Coarticulation effects can exhibit a wide range of context-dependent curvature along with the temporal domain. For an example, the curvature of the first AAM parameter, vary smoothly or sharply depending on the local phonetic context and also the coarticulation effects are localized and do not exhibit very long-range dependencies. These assumptions motivate the main inductive bias in this learning approach, which is to train a sliding window regressor that learns to predict arbitrary fixed-length subsequences of animation. The prediction pipeline is summarized below,

- Decompose the input phonetic sequence x into a sequence of overlapping fixed-length inputs $({}^{x}1, {}^{x}2, ..., {}^{x}T)$ of window size Kx
- For each ${}^{x}j$, predict using h, resulting in a sequence of overlapping fixed-length outputs $({}^{y}1, {}^{y}2, ..., {}^{y}T)$, each of window size Ky
- Construct the final animation sequence y by blending together $({}^{y}1, {}^{y}2, ..., {}^{y}T)$ using the frame-wise mean

h is the initiate using a deep neural network which is a fully connected network with a(sliding window) input layer connected to three fully connected hidden layers and a final output layer.

3.1.6 Rig-Space Retargeting

To generalize the output face model, predicted animation must be retargeted. AAM model captures both appearance and shape parameters and to retarget these parameters any character model can be used. This work intends to use piece-wise linear retargeting where a small set of poses is manually mapped from the reference face model to the target face model.

3.2 Static viseme approach

The concept of this approach is to derive the visemes alphabet from phonemes. The approach to the first research question is addressed here. The high level design of deriving visemes alphabet is depicted in Figure 3.2.

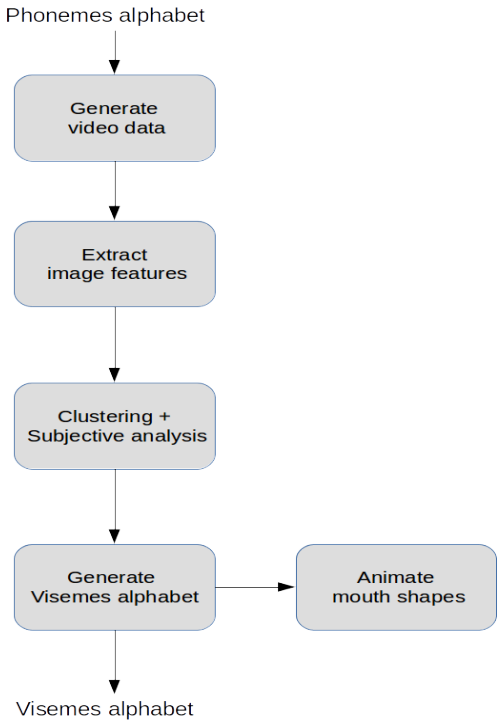

Figure 3.2: Process of deriving visemes alphabet

3.2.1 Sinhala Phonemes alphabet

According to the research (Wasala and Gamage, 2020),Sinhala language is a member of the Indo-Aryan subfamily, which is a member of a still larger family of languages known as Indo-European. Sinhala is the official language of Sri Lanka and the mother tongue of the majority of the people constituting about 74% of its population. Sinhala language is presented in two major varieties: the Spoken and the Literary.

Spoken Sinhala contains 40 segmental phonemes; 14 vowels and 26 consonants,including a set of 4 pre-nasalized voiced stops peculiar to Sinhala, as classified below in figure 3.3 and figure 3.4. (Wasala and Gamage, 2020)

		Labial	Dental	Alveolar	Retroflex	Palatal	Velar	glottal
Stops	Voiceless	p	t		ʈ		k	
	Voiced	b	d		ɖ		g	
Affricates	Voiceless					c		
	Voiced					ɟ		
Pre-nasalized voiced stops		ᵬ	ᵭ		ᶑ		ᶢ	
Nasals		m		n		ɲ	ŋ	
Trill				r				
Lateral				l				
Spirants		f	s			ʂ		h
Semivowels		v				y		

Figure 3.3: Spoken Sinhala consonant classification(Wasala and Gamage, 2020)

	Front		Central		Back	
	Short	long	Short	long	short	long
High	i	i:			u	u:
Mid	e	e:	ə	ə:	o	o:
Low	æ	æ:	a	a:		

Figure 3.4: Spoken Sinhala vowel classification(Wasala and Gamage, 2020)

3.2.2 Generate video data

According to the research (Wasala and Gamage, 2020) spoken Sinhala contains 40 different sounds which include constants and vowel sounds. Research has created a speaking video data set which includes 60 sounds of Sinhala language. The most suitable video frame will be used as the video image of a particular sound. For the vowels, it usually gets an upper part frame of the video and for the constants, it selects the lower part frame of the video. The following figures 3.5 show several frames of the video data set.

Figure 3.5: A sample of frames from the video data set

3.2.3 Extract image features

As the first phase of the feature extraction, selection of frames from the videos will be done. Then using the an Active Appearance Model, the features from the frame images are extracted. Active Appearance Models (AAMs) (Cootes et al., 2001) provide a means for tracking the speech articulators in a video. The shape of an AAM is defined by the two-dimensional vertex locations of a mesh that delineates the inner and outer lip contours and the jaw. Then the model is built by hand-labeling a small number of training images with the vertices that define the mesh.

The AAM is derived in such a way that it can indicate 64 landmarks in the face which have different contours in eyes, nose and mouth. From these contours we only use the mouth area to extract the features. Figure 3.6 shows the 64 landmarks in the face image indicated by the AAM.

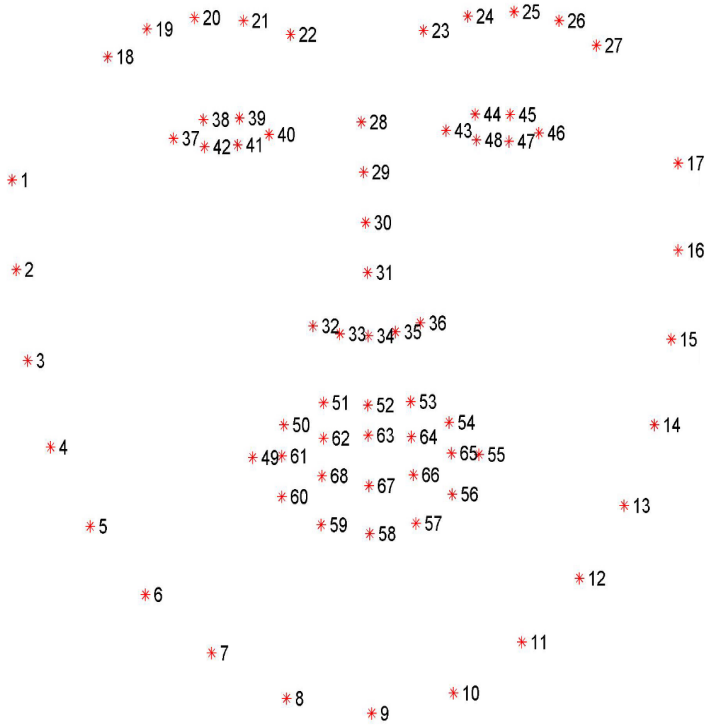

Figure 3.6: The 64 landmarks in face as indicated by AAM

In this research, since we are concerned with the lip and mouth movements, the features will be extracted from the mouth area where the landmarks are denoted by number 48 to number 68.

The contractions of the mouth and lips as shown in figure 3.7 are calculated by measuring the distance between the pre-defined feature points.The main parameters that are calculated are w, h_0,h_1 and h_2.

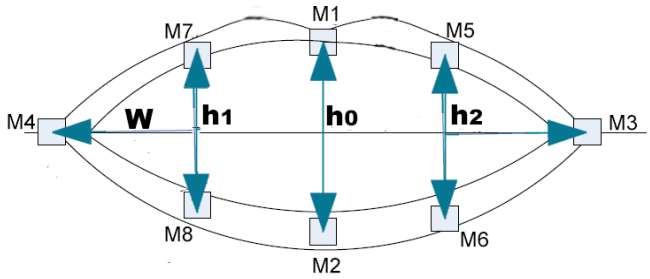

Figure 3.7: The 4 main parameters of the lip shape

w is the horizontal distance between the left corner of the lip to the right corner of the lip. The value for w is taken by getting the euclidean distance between the

points $w_x = |M4_x - M3_x|$ and $w_y = |M4_y - M3_y|$.

h_0 is the vertical distance between middle point of the upper lip and the middle point of the lower lip. It is measured by getting the euclidean distance between the points, $h_x = |M1_x - M2_x|$ and $w_y = |M1_y - M2_y|$.

h1 is the vertical distance of the middle of the left part of the lip and and h2 is the vertical distance of the right part of the lip. These are also measured same as h_0 by getting the euclidean distances of the respective feature points.

Therefore, the features of the mouth can be defined as a 4 parameter vector of $f = [w, h0, h1, h2]$.

Figure 3.8 shows the magnitudes of the 4 feature points respect to vowel sounds.

phoneme	w	h0	h1	h2
e	95.03288904374106	46.51075144523038	46.52418725781247	47.02393007820593
u	89.63955600068532	32.01562118716424	32.14031735997639	32.515380975778214
a	82.25721852822402	53.5	52.53808142671371	54.02082931610732
ae_	96.10541087784809	71.0	69.51618516575834	70.51595564125896
a_	86.3264154242489	58.00215513237418	56.5795899596312	58.502136713115014
u_	85.24816713572204	31.56342820417326	32.190837205639745	32.0624390837628
i	96.5828659752857	43.502873468312416	43.02615483633182	44.52527372178637
e_	100.0199980003999	48.010415536631214	47.510525149697095	49.52272205765753
ae	95.56411460375699	71.0	69.04527500126275	70.01606958406049
i_	99.1526600752597	45.50274716981382	44.52527372178637	46.51075144523038
o	87.64274071479052	37.50333318519835	38.08214804865923	38.5
o_	79.00158226263572	37.0	36.50342449688796	37.013511046643494

Figure 3.8: Magnitudes of the 4 feature points respect to vowel sounds

Figure 3.9 shows the magnitudes of the 4 feature points respect to consonants.

phoneme	w	h0	h1	h2	Constant Letter
ಳ, ಐ	85.248167135722	31.5634282041733	32.1908372056397	32.0624390837628	ಳ, ಐ
ಖ (ಿ)	96.5828659752857	43.5028734683124	43.0261548363318	44.5252737217864	ಖ (ಿ)
ಜ	100.0199980004	48.0104155366312	47.5105251496971	49.5227220576575	ಜ
ಲ, ಪ	89.784535722	31.5634282041733	32.1908372056397	32.0624390837628	ಲ, ಪ
ಪ, ಋ	91.566734857	43.5028734683124	43.0261548363318	44.5252737217864	ಪ, ಋ
ಱ	88.4767135722	31.5634282041733	32.1908372056397	32.0624390837628	ಱ
ೞ	90.73423857	43.5028734683124	43.0261548363318	44.5252737217864	ೞ
ಪ	100.5	48.0104155366312	47.5105251496971	49.5227220576575	ಪ
ಎ, ಏ	95.564114603757	73.64674545	69.0452750012628	70.0160695840605	ಎ, ಏ
ಓ, ಔ	99.1526600752597	45.45334	44.5252737217864	46.5107514452304	ಓ, ಔ
ಓ	87.6427407147905	37.5033331851983	38.0821480486592	37.452323	ಓ
ಜ, ಎ	79.0015822626357	37	36.503424496888	37.0135110466435	ಜ, ಎ
ಧ, ಐ	101.63470004	48.0104155366312	47.5105251496971	49.5227220576575	ಧ, ಐ
ಜ, ಳ	95.564114603757	69.34423243	69.0452750012628	70.0160695840605	ಜ, ಳ
ಧ	99.1526600752597	45.5027471698138	44.5252737217864	46.5107514452304	ಧ
ಪ, ಏ	95.564114603757	71	69.0452750012628	70.0160695840605	ಪ, ಏ

Figure 3.9: Magnitudes of the 4 feature points respect to consonants

3.2.4 Subjective analysis

A subjective comparison will be carried out between each and every mouth shape which is relevant to a particular phoneme, analysing the differences and similarities. Overall there are 60 shapes to compare with each other. The comparison is conducted for two categories which are vowels and constants differently.

After the comparison of all mouth shapes of vowels,we could find 9 different shapes of the mouth for all 14 vowels. Following table 3.10 shows the different visemes groups for the vowels.

Sinhala Letter	Phoneme	Viseme group
අ	a	A
ආ	a:	A_
ඇ	ae	AE
ඈ	ae:	AE
ඉ	i	I
ඊ	i:	I
උ	u	U
ඌ	u:	U
ඍ	/iru/	
ඎ	/iru:/	
ඏ	/ilu/	
ඐ	/ilu:/	
එ	e	E
ඒ	e:	E_
ඓ	ai	
ඔ	o	O
ඕ	o:	O_
ඖ	au	

Figure 3.10: Different viseme groups observed for vowels

After the comparison of all mouth shapes related to constants,we could find 9 different shapes of the mouth for all 26 constants.Following figure 3.11 shows

the different visemes groups for the constants. When analysing both vowels and consonants, it was figured out that when speaking Sinhala, most of the mouth shapes are created when pronouncing vowels.

Constant Letter	Phoneme	Viseme group
ක, ඛ	g	K
ඬ (ංග)	~n	
ඟ	~ g	K
ච, ඡ	c	C
ජ, ඣ	j	K
ඤ,	µ	K
ඤ,	/jµ/	K
ං	/nj/	
ට, ඨ	t	K
ඩ, ඪ	d	K
ණ	`d	K
ත, ථ	t^	K
ද, ධ	d^	K
න, ඦ	n	K
ඳ	~d	K
ප, ඵ	p	P
බ, භ	b	P
ම	m	P
ඹ	mb	P
ය	y	K
ර	r	K
ල, ළ	l	K
ව	w	V
ශ, ෂ	s^	Sha
ස	s	K
හ, (ඃ)	h	K
ෆ	f	F

Figure 3.11: Different viseme groups observed for consonants

The next step after analysing vowels and constants separately, was finding similar mouth shapes from that 18 shapes ,because there can be same mouth movements when speaking vowels and consonants.It was observed that 3 mouth shapes are common to both categories. Thus, after removing the duplicate shapes,15 is left out.Figure 3.12 represents all 15 mouth shapes,from which Sinhala speech can be addressed. This will be used to derive the viseme alphabet for the Sinhala language which then includes 16 viseme groups.

Viseme group	Phonemes
A	a
A_	a:
AE	ae | ae:
I	i | i:
U	u | u:
E	e
E_	e:
O	o
O_	o:
K	g |~ g | j | µ | /jµ/ | t | d | 'd | t^ | d^ | n | ~d | y | r | l | s | h
C	c
P	p | b | m | mb |
V	w
Sha	s^
F	f

Figure 3.12: Fifteen unique viseme groups observed

3.2.5 Clustering

In order to find the visemes classes we get the vowels and constants separately and apply a clustering approach. We use the K-means clustering algorithm for the task of deriving the visemes classes. Since we have already conducted the subjective analysis to derive visemes, the K value will be heuristically determined.

The K means clustering was applied to the vowels and consonants separately as in the subjective analysis. Using the elbow method (Wikipedia contributors, 2019), the optimal number of clusters were figured out by fitting the model with a range of values for K.

Using the elbow curves, the number of clusters for the vowels was recognized as six and for the consonants it was recognized as seven. The total visemes classes observed from the K means clustering approach were 13.

After analysing the results obtained from clustering and subjective analysis, the total number of visemes classes was determined as 15.

3.2.6 Creating mouth shapes

Creating a human graphic model to derive the mouth shapes corresponding to the phonemes is done after getting the visemes groups. The human model generation is done by using a combination of Blender (Blender Online Community, 2017) and Makehuman (Makehuman Community, 2017) softwares. For creating the mouth shapes,only the head section of the model will be take into account. The following figure 3.13 show the human model that was created using Makehuman.

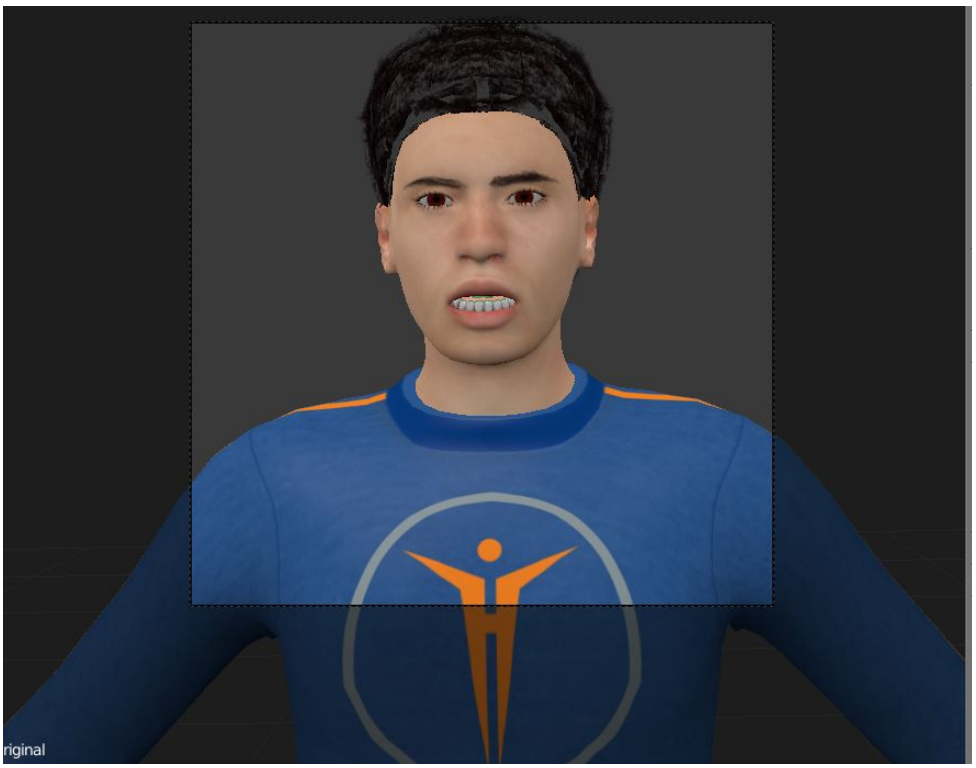

Figure 3.13: Human model created using Makehuman

The mouth shapes modeled with the above human model according to the obtained 15 visemes are presented in the figure 3.14

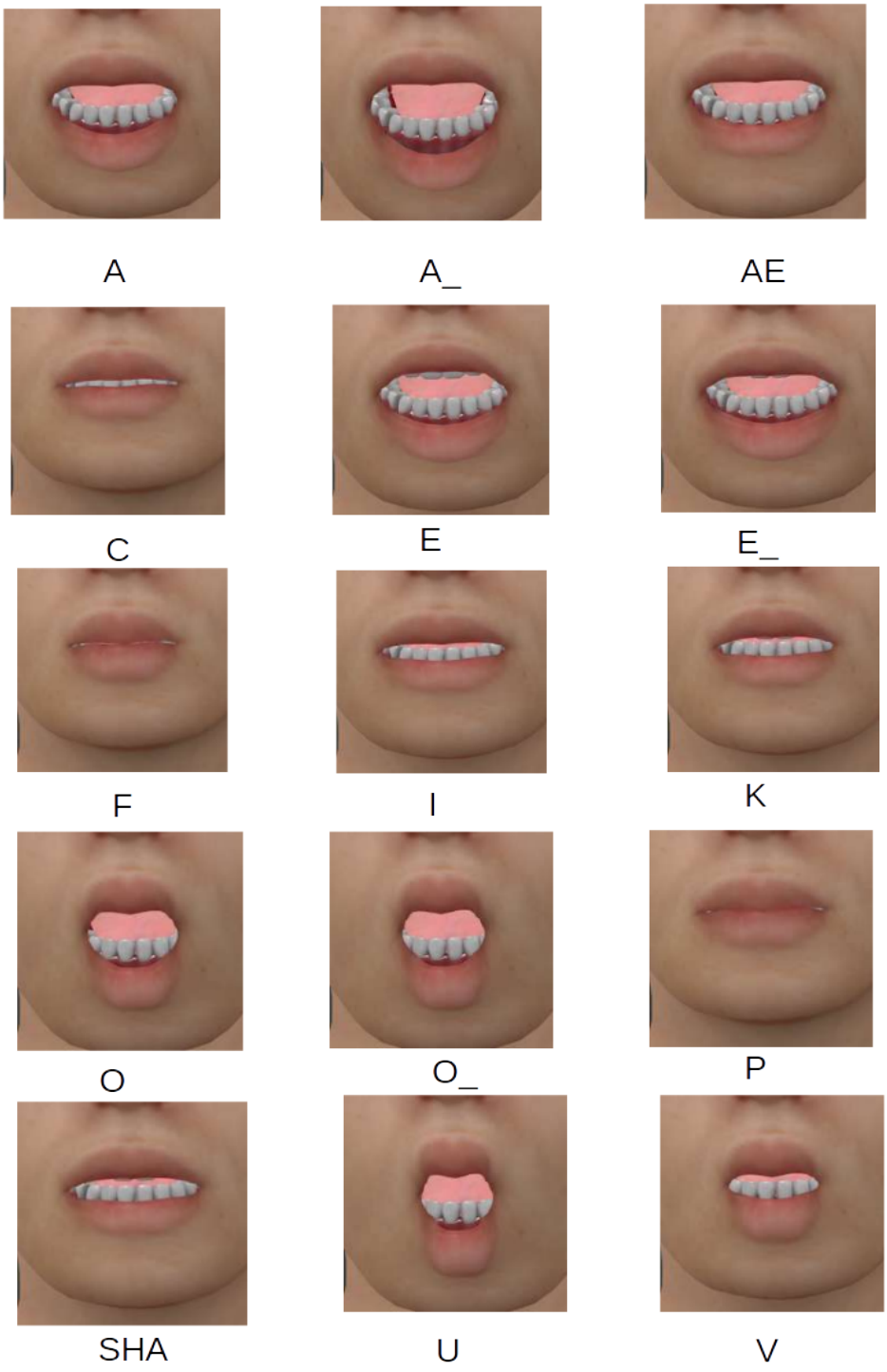

Figure 3.14: Fifteen modeled mouth shapes

After creating the human model, it is saved as an Mhx file format to support Blender. Then the human model is imported to the Blender and changing the shape of the head is done implementing 33 parameters.

Chapter 4

Implementation

Experiments were conducted based on two research designs as described in the chapter:Research design, to address the research questions. This chapter will focus on the implementation details of those 2 research designs. However,the first research design implementation was terminated due to practical problems and second research design is the successful design in this research.

4.1 Implementation for the deep learning approach

The state-of-the-art researches have used deep learning techniques to create better speech animations. Hence as the first phase of implementation, the deep learning approach was tried out for animating the Sinhala speech.

4.1.1 Data set generation

To achieve a better accuracy of the speech animation, generating a good data set that comprises the visemes richness is very important.

Prepare sentences list

The phonetically balanced sentence list used for creating the recordings has been first created to build a natural voice Sinhala TTS system. Since there is a linear mapping between phonemes and visemes, visemes richness can also be achieved through this phonetically rich sentences list. This total sentence list contains up to 5000 sentences including duplicates. After removing the duplicates and numbers,

the pre-processed final sentence list is 3507 sentences. Figure 4.1 depicts a part of the sentence list that is being used for the research.

3486 3486. විසී තමා කැරැ කලක් කලක් දසරජ දම්නෙන්
3487 3487. මහත මිටි හාදයෙකුත් එක්ක ගියා ගමේ ගෙදරකට
3488 3488. මිනිස්කමෙන් පිරිහුණු පුද්ගලයකුට හෙවත් මිනිස් රූපයෙන් යුත් මෘගයකුට නොකළ හැකි
3489 3489. ජේආර් ඇමරිකාවේදී පෙන්වූ විශිෂ්ට රාජ්‍ය තාන්ත්‍රිකත්වය
3490 3490. මෙරට වාර්ගික අර්බුදයට අයර්ලන්තයෙන් පාඩමක්
3491 3491. ඊටද හේතු කාරණා නැතිවා නොවේ
3492 3492. නමුත් පිටිදූවේ සිරිදම්ම යනු සමාජයේ රැල්ල හොඳින් හඳුනන ඉල්ලුමට අනුව සැපයුම ලබ
3493 3493. යනාදී කාරණාවන් පිළිබඳ කොයියම් සහෘදයකු වුවත් තීක්ෂණ අවධානයක් යොමුකරන්න
3494 3494. පගාව ගත්තෙ නෑ
3495 3495. මුල්කාලීනව පැවතියේ ගෝත්‍රික කණ්ඩායම් සමූහයකි
3496 3496. වැඩසිටිකන්දේ කොත නිරාවරණය දෙවැනිදා පැවැත්වේ
3497 3497. ඒවායේ මුළු අවශ්‍යතාවයෙන් සියයට පනහක් සපුරාලීම මේ වසරේ ඉලක්කයයි
3498 3498. ගොවියත් රැකිණි
3499 3499. කාන්තාව හොඳමයි සර්
3500 3500. සංවාදය තුෂාරී කළුබෝවිල
3501 3501. සත්‍ය ගරුකය
3502 3502. ඇන්ගෝලාවේ ආණ්ඩු විරෝධී කැරැලිකරුවෝ වෙඩිබෙහෙත් සහ චීනච්චට්ටි ගැමියන්ට සප
3503 3503. ඔවුන් හදන්නේ අපිව පොල් කට්ටෙන් වහලා තියන්න
3504 3504. මං ආසයි අස් කරන්න නිරන්ජන් ගුරුගේ
3505 3505. කාව්‍ය කියවත්ම මියගිය උවැසියගේ රූපය හදවතේ ළැගුම් ගන්වන පරිද්දෙනි
3506 3506. එන්නත් ලබා ගැනීමේදී පවතිනුයේද මෙවැනිම තත්ත්වයකි
3507 3507. ප්‍රභාකරන්ද විශ්වාස ඝාතකයෙකි

Figure 4.1: Phonetically balanced corpora

Recording data

This project expects to record at least 2000 sentences to ensure the validity of the lip synchronization model. So far, 250 recordings with a neutral speaking style of Sinhala sentences have been completed. The complete set of recordings were done by myself using a remote camera. Moreover, the videos are recorded at 25 frames per second at a resolution of 1280 by 720 progressive scan and run to approximately 2 hours long and only the frontal view of the face is captured. All the recordings were conducted at the UCSC studio under a controlled environment. Below Figure 4.2 depicts a sample frame of the recorded data set.

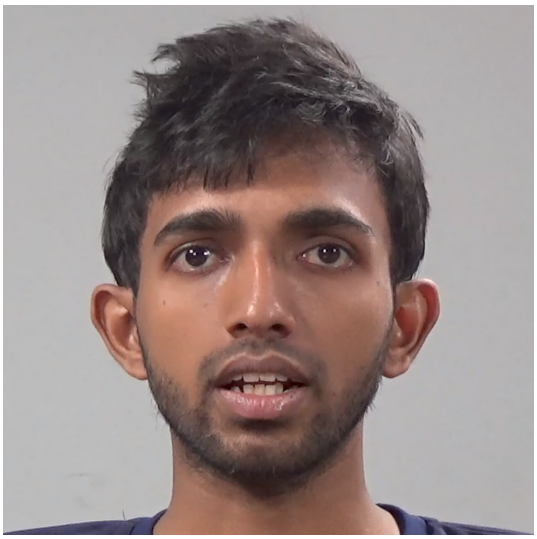

Figure 4.2: Sample frame of recorded data

4.1.2 Feature extraction

Feature extraction was done using Active Appearance model which is a statistical deformable model of the shape and appearance of a deformable object class. Training the AMM was carried out using an existing publicly available dataset named LFPW database ("in-the-wild", 811 and 224 training and testing images, respectively) (Belhumeur et al., 2011). The trained dataset has 68 landmarks of the face. The parameters obtained after the feature extraction process in mouth shapes is depicted in the figure 4.3.

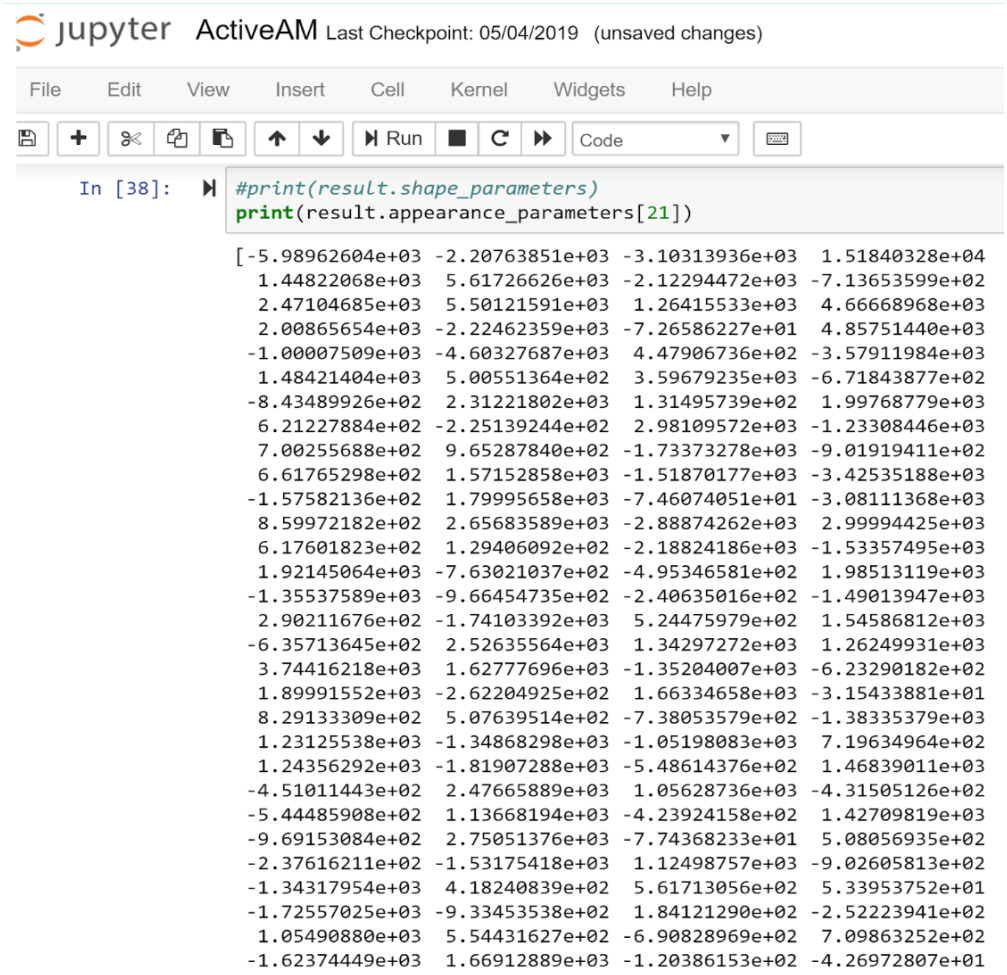

Figure 4.3: Example feature vector of a image

4.1.3 Termination of Deep learning approach

Although the deep learning approach has been considered as a good model for speech animation, the problem arises when it is used for languages with low resources. According to taylor's work (Taylor et al., 2017, 2012), to achieve a successful model, it requires to have at least 10 hours of data.

In this research, we were able to record around 250 videos after spending 2-3 months of time. However, the hardest part is when it comes to labeling the data set. Unlike other languages, for the Sinhala language , a tool for labeling the image frames is not available. Thus, if it were to label all the data set, it would require months of time and manual effort as well. After considering these issues, the research approach was changed from deep learning to Static visemes.

4.2 Static Visemes Approach

As described in Chapter 3, to generate the speech animation, the visemes alphabet is derived from the phonemes. The process involved in deriving a visemes sequence for a given input data will be addressed in this section.

A detailed implementation design of the system is shown by the figure 4.4.

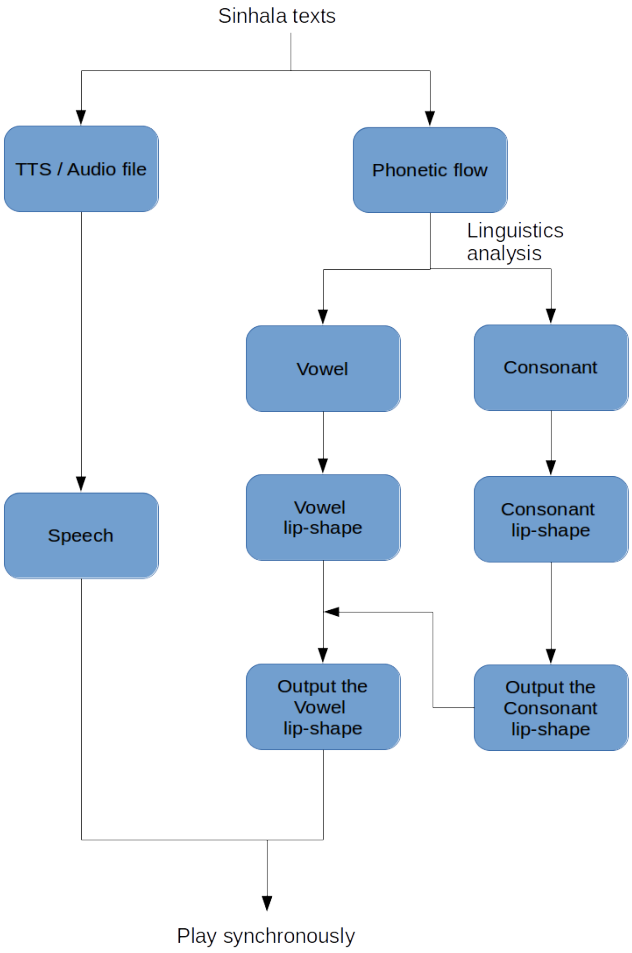

Figure 4.4: The high-level architecture for the static visemes approach

As in the figure 4.4, the input for the system is a Sinhala text. Then the relevant phoneme sequence for the input Sinhala text is converted to its visemes sequence. Afterwards, the speech animation is obtained by processing the visemes sequence synchronously. The following sections will describe the steps above in detail.

4.2.1 Deriving the phonetic flow

The input Sinhala text is broken down into its phonemes sequence according to the UCSC Subasa Sinhala Transliteration project written in Java. It comprises two major steps which are “encoding” and “schwa analysis”. However, integration of this project directly to the Papagayo python software gives compatibility issues. Therefore, the necessary steps were re-implemented using python to meet the requirements of this project. In the encoding step, the vowels and constants are labelled separately using a predefined schema as in the figure 4.5.

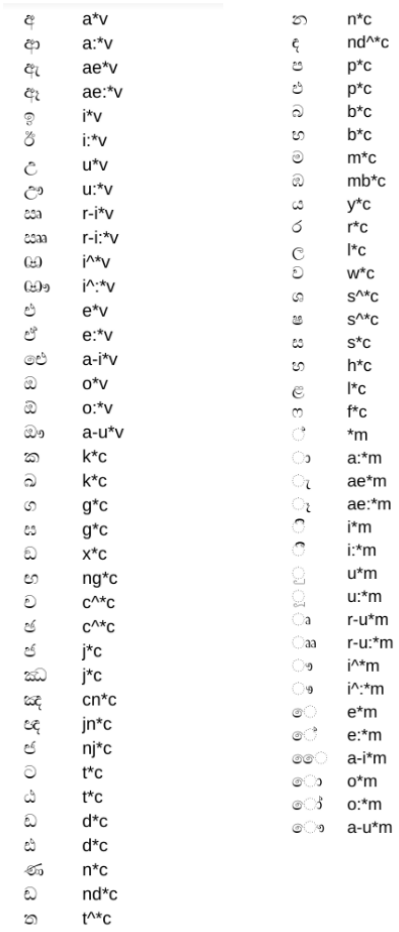

Figure 4.5: The encode schema which is used to label vowels and consonants

In figure 4.5, the left side denotes the Sinhala letter and the right side denotes its relevant phoneme along with its type: vowel (v) or consonant (c).

After the encoding step, a Sinhala word will take the structure as in figure 4.6.

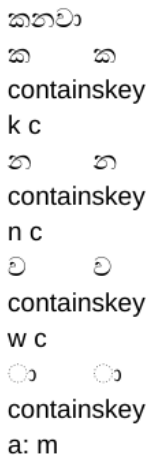

Figure 4.6: Example of a word output after the encoding step

All the words are converted according to the above structure in figure 4.6 to make an encoded phoneme sequence.

According to the research work (phoneme report), the Sinhala character set has 20 vowels and 40 consonants. There are 18 diacritics, 17 of them denote the vowels and are known as vowel modifiers, and the other one (◌්) represents the pure consonant form. (i.e. without ◌්, the consonant is pronounced either associating it with schwa or the vowel "a" . i.e. k and = kə or ka). ". In the encoding section shewa analysis is not handled.

The next stage of the phoneme breakdown is schwa analysis and here it replaces vowels from shewa (@ symbol represents the shewa) according to the pronunciation. This can be easily shown by an example word in Sinhala. In word "මකනවා ", the pronunciation of letter "ක " is different from the pronunciation of letter "ක" in word "කනවා".

In the word "කනවා", the phoneme sequence is as follows.

['k', 'a', 'n', '@', 'w', 'a:']

In the word "මකනවා", the phoneme sequence is as follows.

['m', 'a', 'k', '@', 'n', '@', 'w', 'a:']

In the word "මකනවා", the vowel phoneme "a" is replaced by shewa (@) since the pronunciation is different.

Eight rules were re-implemented in python to handle the shewa analysis. These rules are presented in the appendix A.1. After applying them, the final correct phonemes sequence for a given Sinhala word or sentence is returned.

4.2.2 Mapping phonemes to visemes

The next step involves mapping the phonemes into its corresponding viseme character and deriving the visemes sequence of the Sinhala text. This is a one to one mapping which uses a visemes dictionary.

```
phoneme_conversion = {
    a': 'A',        d':'K',     k': 'K',
    a:': 'A_',      t^':'K',    g': 'K',
    ae': 'AE',      d^':'K',    j': 'k',
    ae:': 'AE',     n':'K',     µ': 'k',
    i': 'I',        p':'P',     c':'C',
    i:': 'I',       b':'P',     t':'K',
    u': 'U',        m':'P',     d':'K',
    u:': 'U',       mb':'P',    t^':'K',
    ri': 'i',       y':'K',     ii':'I_',
    ru': 'i',       r':'K',     ai':'AE',
    ilu': 'i',      l':'K',     o:': 'O_',
    ilu:': 'i',     w':'V',     ou': 'U',
    e': 'E',        s^':'SHA',  x': 'rest',
    e:': 'E_',      s':'K',
    ai': 'I',       h':'K',
    o': 'O',        f':'F',
```

Figure 4.7: One to one mapping between phonemes and visemes

4.2.3 Speech animation

To model the speech animation, the derived visemes sequence is mapped to mouth shapes. I use the Papagayo (Papagayo Community, 2011) open source tool which has a GPL license and modified it to meet my requirements. The Papagayo interface was integrated to my code base to demonstrate the animated speech.

The Papagayo requires both text form and audio to generate the speech animation. To obtain the audio form from the text, a Sinhala TTS system was used at first. However, for some input Sinhala sentences, the TTS system is not flexible enough to control the speed of audio. Therefore, they had to be recorded manually

and imported to Papagayo. The Papagayo interface also supports making changes to the frame rate of the animation which is helpful for obtaining realistic behaviour. Figure 4.8 shows the integrated Papagayo interface.

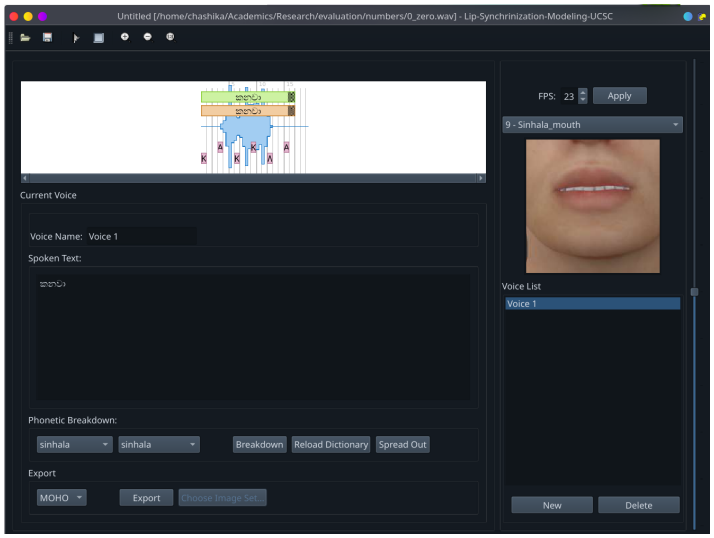

Figure 4.8: The Papagayo interface

When the Sinhala sentence is passed, it is breakdown into its visemes sequence step by step as described in previous sections. This sequence of processes can be visualized using the Papagayo as in figure 4.9.

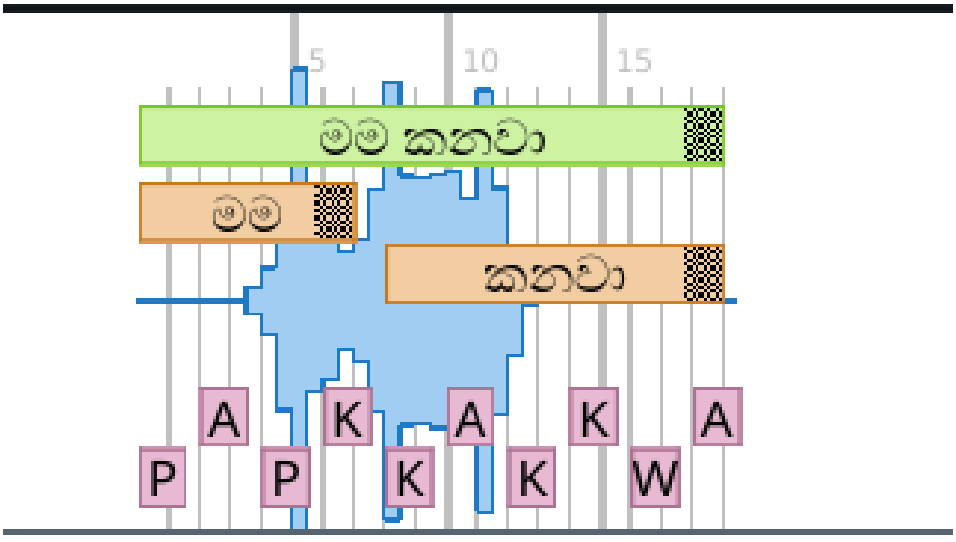

Figure 4.9: Breakdown of Sinhala text into visemes. First line shows the input Sinhala text, second line shows the separated words chunks and last line shows the mapped visemes sequence

In this chapter, the conversion steps of sinhala text to visual speech animation

using the static visemes approach were presented . The main steps that were followed can be summarized as below,

1. convert the Sinhala text to phonemes using the re-implemented 8 rules
2. Map phonemes to the viseme classes using a dictionary
3. Play the visemes sequence sequentially to generate the speech animation
4. Papagayo open source tool is used to display the output of the model

Chapter 5

Results and Evaluation

It is difficult to find any metrics other than human evaluation to evaluate the naturalness of a speech animation. Moreover, there are no available research activities in this area to be compared with. Hence, the evaluation of this research will be based on the subjective analysis.

The subjective evaluation process is used to evaluate three main aspects of the Sinhala animation model. They are,

- The efficacy of Sinhala visemes
- The quality
- Internal performance

We chose 50 different Sinhala sentences from different categories as below to test the reaction of the model.

- Sinhala digits
- Sinhala words
- Named entities
- Short sentences (less than 5 words)
- Long sentences (more than 5 words)
- Mix of English sentences

The short and long sentences were collected from UCSC LTRL speech corpus as they are rich in phoneme diversity. The rest of the sentences or words from other categories were selected to test the visemes diversity of the Sinhala language.

Mainly, the subjective evaluation process was carried out using the rating method.

5.1 Rating evaluation

Six questionnaires for the six categories where each carrying 10 questions and a rating system of 1 to 5 scale (1 is an unsatisfied opinion and 5 meaning that you are very satisfied) were distributed among undergraduates. The participants should rate the performance of the model according to how satisfied they are with the speech animation video. Figures 5.1, 5.2 shows a part of the evaluation questionnaire.

The accuracy of any item related to a category was calculated using the below equation,

$$accuracy = \frac{score}{responses * 5} * 100\% \tag{5.1}$$

$$accuracyQuestion = \frac{NoOfCorrectAnswers}{responses} * 100\% \tag{5.2}$$

$$accuracyCategory = \frac{\sum_{i=1}^{n} accuracyQuestion_i}{n} \tag{5.3}$$

$$accuracyOverall = \frac{\sum_{i=1}^{c} accuracyCategory}{c} \tag{5.4}$$

$$accuracyOverall = \frac{\sum_{i=1}^{n} accuracyQuestion_i}{n} \tag{5.5}$$

In equation (5.1), the score is calculated by getting the sum of ratings of all the responses for that item. Rating 5 has the maximum value of 5 and rating 1 has the lowest value of 1.

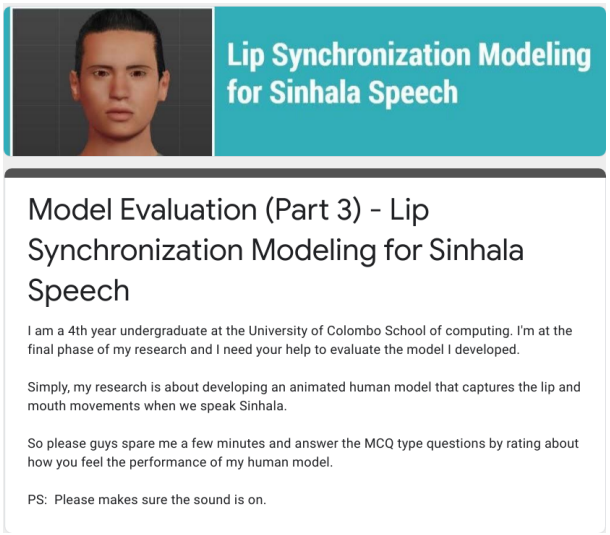

Figure 5.1: Part of the questionnaire for rating evaluation -1

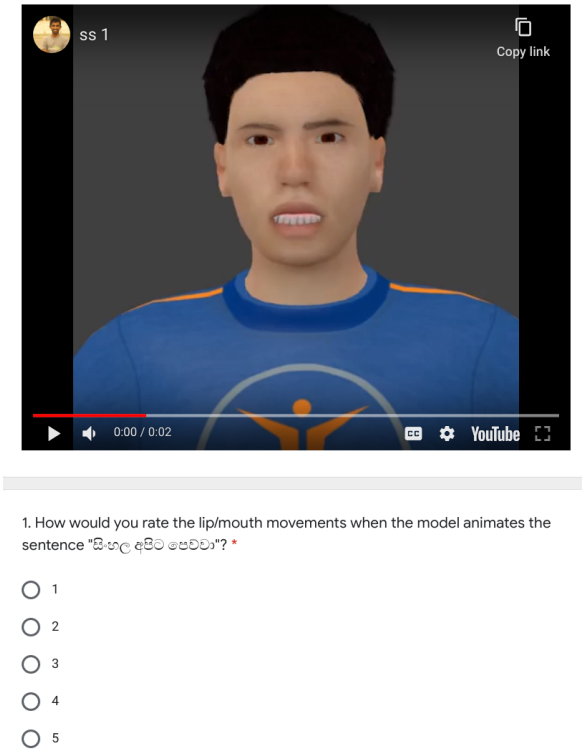

Figure 5.2: Part of the questionnaire for rating evaluation -2

5.1.1 Subjective analysis of digits,words and named entities

Table 5.1 represents the results of the digits category reported from 30 responses.

Table 5.1: Results of the digits category reported from 30 responses.Rate 1 means unsatisfied and rate 5 means very satisfied

Digit	% of rate 1	% of rate 2	% of rate 3	% of rate 4	% of rate 5
එක	0	6.7	16.7	43.3	33.3
දෙක	0	6.7	16.7	43.3	33.3
තුන	0	3.3	23.3	43.3	30
හතර	0	3.3	20	50	26.7
පහ	0	10	20	36.7	33.3
හය	0	3.3	10	43.3	43.3
හත	0	0	20	36.7	43.3
අට	0	0	10	43.3	46.7
නවය	0	0	10	43.3	46.7
බින්දුව	3.3	6.7	33.3	40	16.7

The overall accuracy observed for the digits category is displayed by the bar chart in figure 5.3.

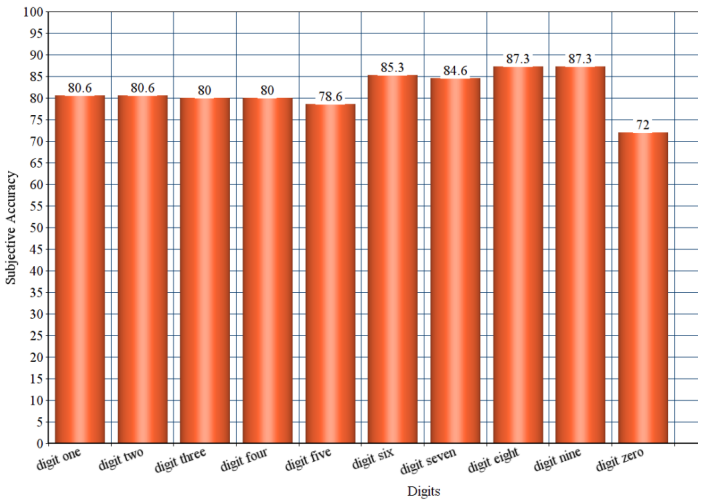

Figure 5.3: Overall accuracy for digits category. x axis denotes digits while y axis denotes the subjective accuracy

The overall accuracy for the digit category was observed to be 75% and the lowest accuracy has been gained for the word “බින්දුව”. When analyzing this word, it was found that this word includes a different viseme class that represents “න්” which is rarely used. This may lead to lower accuracy.

Table 5.2 represents the results of the words category reported from 27 responses.

Table 5.2: Results of the word category reported from 27 responses.Rate 1 means unsatisfied and rate 5 means very satisfied

Word	% of rate 1	% of rate 2	% of rate 3	% of rate 4	% of rate 5
අම්මා	3.7	25.9	18.5	33.3	18.5
ඇනය	3.7	11.1	25.9	33.3	25.9
ඉඳල	3.7	3.7	18.5	48.1	25.9
උයන්වත්ත	0	7.4	25.9	51.9	14.8
ඒකල	0	7.4	7.4	44.4	40.7
ඔරුව	7.4	18.5	33.3	29.6	11.1
කනවා	3.7	7.4	29.6	48.1	11.1
පනාව	3.7	18.5	22.2	40.7	14.8
ටොෆි	0	0	11.1	48.1	40.7
වේදනාව	0	0	11.1	51.9	37

The overall accuracy observed for the word category is displayed by the bar chart in figure 5.4.

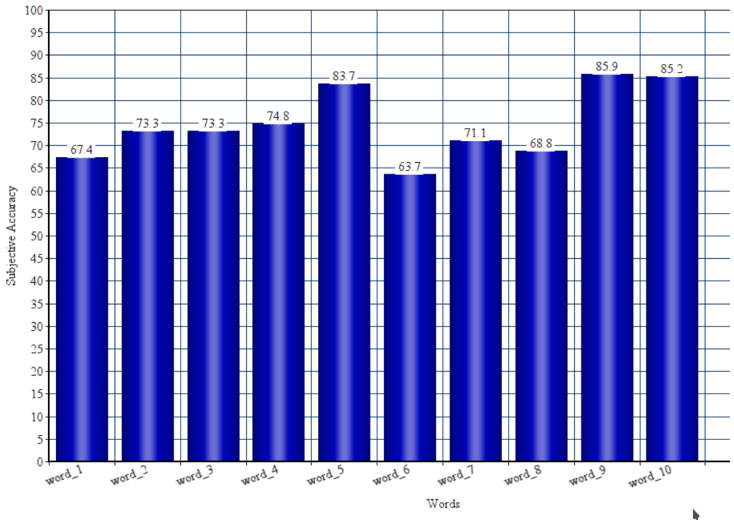

Figure 5.4: Overall accuracy for word category. x axis denotes digits while y axis denotes the subjective accuracy

The overall accuracy for the word category was 74.7%. It can be observed that most of the words in this category, have obtained a lower accuracy when compared with the digits category. For words such as “ඔරැවි”, the accuracy has fallen down to 63.7%. Here, the visemes that reflect “ඔ” and “රැ” have not synchronized properly and due to that the naturalness of the lip movements are not displayed.

Table 5.3 represents the results of the named entity category reported from 12 responses.

Table 5.3: Results of the named entity category reported from 12 responses. Rate 1 means unsatisfied and rate 5 means very satisfied

Word	% of rate 1	% of rate 2	% of rate 3	% of rate 4	% of rate 5
කමල්	0	8.3	66.7	25	18.5
අමර	0	0	41.7	33.3	25
නුගෙගොඩ	0	15	25	50	10
සාගර	0	0	41.7	41.7	16.7
කමලාවතී	0	16.7	50	33.3	0
වශික	0	0	33.3	50	16.7
රාම	0	0	25	66.7	8.3
වල් බල්ලා	0	16.70	41.73	33.30	8.7
දෙල්කද	0	0	33.35	41.74	251
කොලඹ		0	11.1	51.9	37

The overall accuracy observed for the named entity category is displayed by the bar chart in figure 5.5.

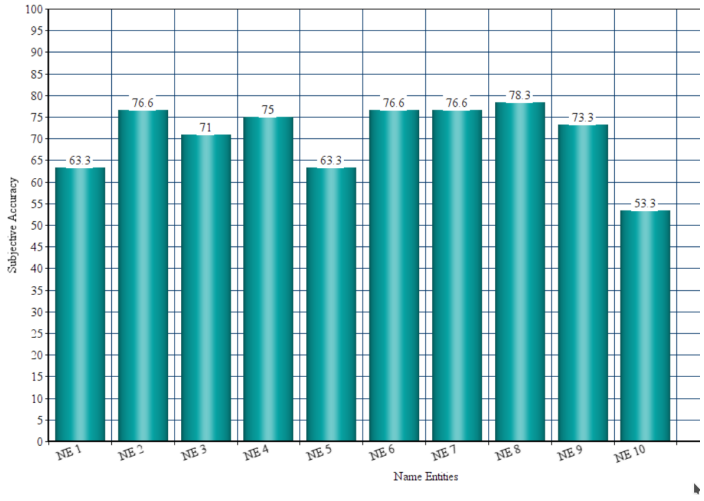

Figure 5.5: Overall accuracy for named entity category. x axis denotes digits while y axis denotes the subjective accuracy

5.1.2 Subjective analysis of short sentences

The short sentences list used for the evaluation is shown in the table 5.4.

Table 5.4: The short sentences list use for evaluation

Sentence	Sinhala text
sentence 1	සිංහල අපිට පෙව්වා
sentence 2	එහි ප්‍රතිඵල අද අපි යහමින් භුක්ති විඳිමු
sentence 3	එංගලන්ත ක්‍රීඩකයෝ
sentence 4	ආනන්ද නාලන්දා සුලෝහිත සංග්‍රාමය එස්එස්සී පිටියේදී
sentence 5	ප්‍රාණඝාතය ද සොරකම ද එ සේ ම ය
sentence 6	ලංකාව ඒ අතර විශේෂිත ය
sentence 7	ජනාධිපතිවරයා දූෂිත නම් ඡන්දවලට යන්නේ නැත
sentence 8	උද්ධච්ච බව මුලිනුපුටා දැමීමය
sentence 9	හොඳ සහ නරක
sentence 10	ඒක කාටවත් ප්‍රතික්ෂේප කරන්න බැහැ

Table 5.5 represents the results of the short sentence category recorded from 34 responses.

Table 5.5: Results of the short sentences category reported from 34 responses.Rate 1 means unsatisfied and rate 5 means very satisfied

Sentence	% of rate 1	% of rate 2	% of rate 3	% of rate 4	% of rate 5
sentence 1	0	2.9	20.6	52.9	23.5
sentence 2	2.9	5.9	50	29.4	11.8
sentence 3	2.9	23.5	35.3	32.4	5.9
sentence 4	2.9	8.8	41.2	35.3	11.8
sentence 5	2.9	11.8	38.2	41.2	5.9
sentence 6	0	23.5	23.5	41.2	11.8
sentence 7	2.9	14.7	29.4	44.1	8.8
sentence 8	2.9	8.8	14.7	58.8	14.7
sentence 9	0	8.8	26.5	29.4	35.3
sentence 10	0	8.8	26.5	47.1	17.6

The overall accuracy observed for the short sentence category is displayed by the bar chart in figure 5.6.

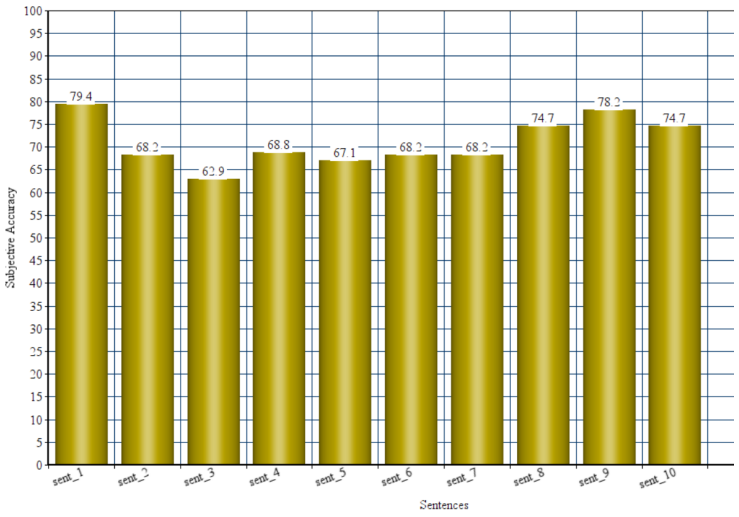

Figure 5.6: Overall accuracy for short sentence category. x axis denotes sentence while y axis denotes the subjective accuracy

An overall accuracy of 71% was observed from the short sentences category which is lower than both digit and word category. The reason for evaluating the

short sentences is to measure the model accuracy when animating more than one word. However, the results show that the model has not performed very well in animating transitions between words.

5.1.3 Subjective analysis of long sentences

The long sentences list used for the evaluation is shown in the table **??**.

1. Sentence 1 : ස්ටාලින් අර්බුදයේ තැති ගැන්ම හා නිලධාරිවාදයේ ශාපය සමාජවාදී ලෝකයෙහි යම් දෙදරවීමක් ඇති කෙළේ ය
2. Sentence 2 : ලෝකයේ පිළිගත් ආර්ථික ශ්‍රේණිගත කිරීමේ සමාගමක් බැංකු පද්ධතියක් ගැන සැක පහළ කිරීමක්ම මේ සඳහා අවශ්‍ය නොවේ
3. Sentence 3 : එසේ නැතුව ලෝක වෙළඳපොලේ කෙසෙල් සඳහා ශ්‍රී ලංකාවේ නිපදවන කෙසෙල් වලට තරහ කිරීමට කොහෙත්ම බැරිය
4. Sentence 4 : සිලා මේඝ වර්ණ රජතුමා යුද්ධයේදී අල්වා ගත් සිරකරුවන් දාසයන් ලෙස ආරාමවල සේවයට යෙදූ තොරතුරක් මහාවංශයේ එයි
5. Sentence 5 : බිංගිරිය පොලිසියේ පොප කමල් රත්නායක සමාජ ප්‍රවර්ධන නිලධාරී සේපාල හේරත් කුලියාපිටිය සේවා වනිතා ඒකකයේ ලේකම් කාපොසැ
6. Sentence 6 : නමුත් විජේතුංග ගත් හැරවුම් ක්‍රියාමාර්ගය මහජනයාට ඒත්තු ගැන්වීමට නොහැකි පරිදි එජාපය විසින් පාර වරද්දාගෙන තිබිණි
7. Sentence 7 : හොවැංහෝ යුප්‍රටීස් ටයිග්‍රීස් ඉන්දුනිම්න යන ගංගා ආශ්‍රිත ශිෂ්ටාචාරයන් ලෝක ඉතිහාසයේ නොමැකෙන ලෙස සටහන් ව තිබේ
8. Sentence 8 : ඉන් එක් ස්ථානයක් අනුරාධපුරයට නුදුරින් පිහිටා ඇති අතර අනෙක් ස්ථානය දඹදෙණිය විජයසුන්දරාරාම පුරාණ රජමහා විහාරය වෙයිය
9. Sentence 9 : එක්සත් ජාතීන්ගේ ආරක්ෂක මණ්ඩලය දැනටමත් එක්සත් ජනපදයේ හා බ්‍රිතාන්‍යයේ ඉල්ලීම අනුව අවස්ථා හතරකදී සම්බාධක පනවා ඇත
10. Sentence 10 : මෙම නඩුව වසර පහළොවක් තිස්සේ හම්බන්තොට මහාධිකරණයේ විභාගවෙමින් තිබුණු අතර පසුව තංගල්ල මහාධිකරණයට මාරුකර තිබුණි

Table 5.6 represents the results of the long sentence category recorded from 12 responses.

Table 5.6: Results of the long sentences category reported from 12 responses.Rate 1 means unsatisfied and rate 5 means very satisfied

Sentence	% of rate 1	% of rate 2	% of rate 3	% of rate 4	% of rate 5
sentence 1	8.3	33.3	41.7	16.7	0
sentence 2	0	41.7	58.3	0	0
sentence 3	0	16.7	75	8.3	0
sentence 4	0	16.7	41.7	41.7	0
sentence 5	8.3	33.3	58.3	0	0
sentence 6	16.7	41.7	41.7	0	0
sentence 7	8.3	50	33.3	8.3	0
sentence 8	0	41.7	50	8.3	0
sentence 9	0	8.3	75	16.7	0
sentence 10	0	41.7	50	8.3	0

The overall accuracy observed for the long sentence category is displayed by the bar chart in figure 5.7.

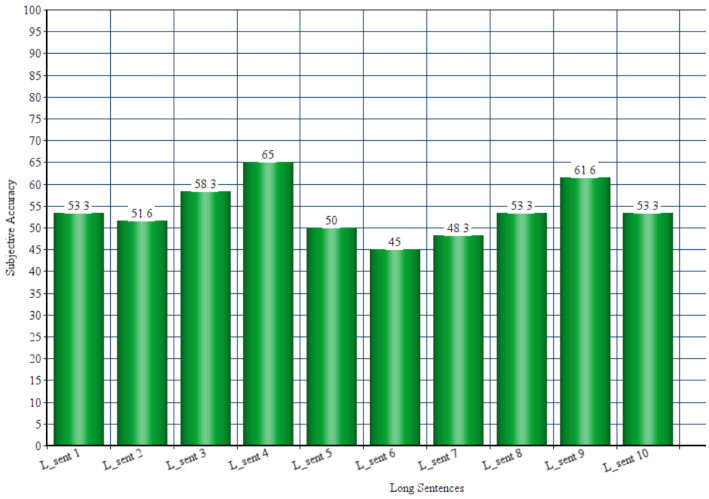

Figure 5.7: Overall accuracy for long sentence category. x axis denotes sentence while y axis denotes the subjective accuracy

The long sentence category could obtain only a 54% accuracy for the subjective analysis. Nor participants have rated 5 for the long sentences. This is the lowest accuracy observed from all the categories. Based on the results, it is evident that when the number of transitions between the words increases, the performance of the model tends to decrease.

5.1.4 Subjective analysis of mix of English and Sinhala sentences

The mix of Sinhala and English sentences list used for the evaluation is shown in the table **??**.

1. Sentence 1 : අපි වැම්පියන්ස්ලාය
2. Sentence 2 : අද තරගයේ විනර්ස්ලා වෙන්නායි සුපර් කින්ග්ස්ේ
3. Sentence 3 : එක එක්ස්ප්‍රස් ට්‍රේන් එකක්
4. Sentence 4 : අපේ ස්කූල් බස් එක කලින් ගිහින්
5. Sentence 5 : ගියවර ටර්ම් ටෙස්ට් එකේ ක්ලාස් පර්ස්ට් මමමස
6. Sentence 6 : එයාට තමයි වෙස්ටන් මියුසික් වල වැඩිම ලකුනු
7. Sentence 7 : ෆුට්බෝල් තමයි ලෝකයේ ෙම්මස්ම ක්‍රීඩාව
8. Sentence 8 : උසේන් බෝල්ට් තමා වර්ල්ඩ ෆාස්ටස්ම ප්ලේයර් ය
9. Sentence 9 : එයා හරිම කියුටියු
10. Sentence 10 : මගේ බෝයිෆ්රෙන්ඩ හරිම ඉන්ටලිජන්ටණි

Table 5.7 represents the results of the mix of English and Sinhala sentence category recorded from 11 responses.

Table 5.7: Results of the Sinhala, English mixed sentences category reported from 11 responses.Rate 1 means unsatisfied and rate 5 means very satisfied

Sentence	% of rate 1	% of rate 2	% of rate 3	% of rate 4	% of rate 5
sentence 1	0	9.1	45.5	36.4	9.1
sentence 2	0	9.1	63.6	27.3	0
sentence 3	0	9.1	45.5	36.4	9.1
sentence 4	0	0	18.2	54.5	27.3
sentence 5	0	27.3	54.5	18.2	0
sentence 6	0	18.2	54.5	27.3	0
sentence 7	0	9.1	72.7	18.2	0
sentence 8	0	9.1	45.5	36.4	9.1
sentence 9	0	27.3	36.4	36.4	0
sentence 10	0	9.1	54.5	36.4	0

The overall accuracy observed for the mix of English and Sinhala sentence category is displayed by the bar chart in figure 5.7.

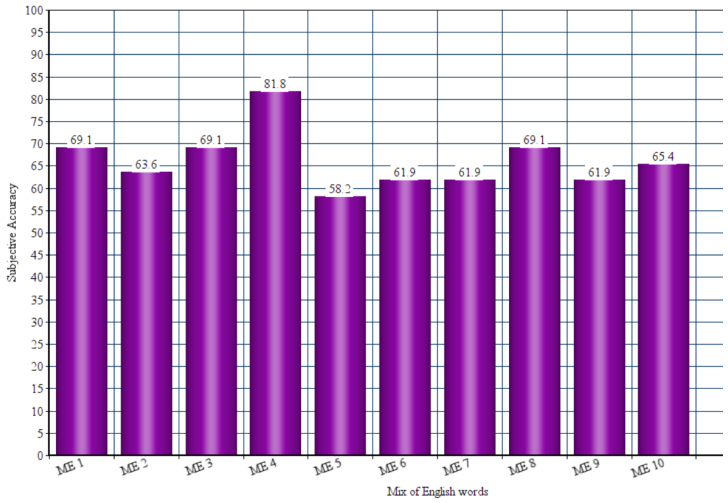

Figure 5.8: Overall accuracy for mix of English and Sinhala sentence category. x axis denotes sentence while y axis denotes the subjective accuracy

The overall accuracy observed from the mix sentences was 66.2

5.2 Overall Observation and Discussion

The intention of the ranking approach is to evaluate the model performance based on the subjective analysis. The accuracy gained for each evaluated category is summarized by figure 5.9.

The average accuracy for digits, words, named entities, short sentences, long sentences and Sinhala,English mixed sentences are 75, 74.7, 72, 71,54 and 66.2 percent respectively.

Based on the results, it can be stated that the model performs better for individual words rather than sentences. For long sentences, the model failed to animate accurately due to the increase in the number of transitions between the words.

When analysing the results, it was observed that the model is only able to correctly animate the speeches which have a constant speed. Although the sentence is uttered with different speeds, the model will only consider its phoneme sequence and then animate the relevant viseme frames at a constant speed. Therefore, the synchronization between the audio and lip movements fail vastly. In the comments section of the questionnaire for long sentences, two participants had commented out this synchronization issue. Thus, this can be considered as a drawback of the model when animating long sentences. Adopting the model to handle speed variations of the input speech is one of the notified improvements to be done.

The overall accuracy based on all the categories in the subjective evaluation is 68.8%.

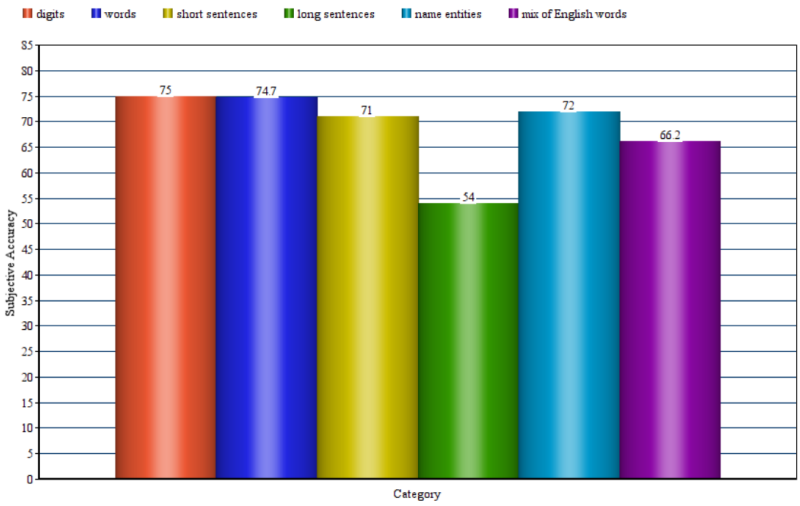

Figure 5.9: Overall accuracy the six categories. x axis denotes sentence while y axis denotes the subjective accuracy

Chapter 6

Conclusions

6.1 Introduction

This research is on implementing a lip synchronization animation model for the Sinhala language by using a static visemes approach. Initially the research started with the state of the art approach which is the deep learning approach. With the limited time frame, since it is difficult to tackle the problem of the unavailability of the resources and tools, the deep learning approach was terminated.

In this chapter, the overall conclusion of the research work conducted will be presented.

6.2 Conclusions about research questions and objectives

Implementing a lip synchronization model for the Sinhala language is the main research problem in this research.. To address the main objective, we answer the three research questions throughout this research.

The first research question is "What is the visemes alphabet for the Sinhala language to generate speech animation". Since there were no available viseme alphabets for Sinhala, the first and foremost step of the research was deriving the Sinhala visemes alphabet. Two methods which are clustering and subjective analysis were experimented for this co-task. As the initial step, the subjective analysis was conducted using the video dataset which was recorded by myself.

The Spoken Sinhala contains 40 segmental phonemes; 14 vowels and 26 consonants according to the (Wasala and Gamage, 2020). The subjective analysis was conducted on three levels separately for the vowels and consonants and the third level is to combine both vowels and consonants and conduct the analysis again. After the experiments, 15 visemes classes were derived.

As the second step, we conducted the clustering approach to derive the viseme classes. The K means clustering algorithm was used since the k value was approximately known from the subjective analysis. Using the clustering method, 13 viseme classes were found. After combining it with subjective analysis, a total of 15 visemes classes were derived.

The second research question is "What is the proper way to generate the speech animation?".To generate the speech animation it is required to have graphic designing tools. In this research work, the human model was designed using Makehuman and blender open source tools by controlling the mouth variations using 33 mouth parameters.

To answer the third research question which is "What is the proper way to find visemes sequence from the Sinhala text in order to generate the visual speech animation?", we converted the Sinhala text into phonemes sequence using the UCSC Subasa project. Using the obtained phoneme sequence, the visemes sequence was generated from which was then used to sequentially generate the speech animation.

A subjective evaluation process was conducted to evaluate the speech animation model . Since there are no developed speech animation models for the Sinhala language, an objective evaluation process can not be conducted.

6.3 Conclusions about research problem

According to the results and findings concluded in section 6.2, this research work has been able to implement a lip synchronization model that yields an overall accuracy of 69% for the subjective evaluation using the static visemes approach. As an initiative research on speech animation for Sinhala, this model accurately animates individual words rather than long sentences.

This paper is also expected to serve as a starting point for those interested in

initiating projects for low resourced languages that follows from non-Latin linguistic traditions.

6.3.1 Conclusions about the deep learning approach

Initially, this research followed the deep learning approach which is the state of the art for generating speech animations. However, the deep learning approach was terminated in the middle due to the practical issues occured and then moved to the static visemes approach. This section will summarize the challenges and the path we applied for implementing the speech animation model using the deep learning approach.

The main issue was the limited data set . Although other languages have standard visual speech data sets (Afouras et al., 2018) for Sinhala, there are no existing data sets related to speech animation. The recorded data set for this research was only 250 sentences that would roughly estimate to around 30 minutes. According to (Taylor et al., 2017, 2012), the authors have used a 10 hour video data set to implement a lip synchronization model for English language. Thus, with 250 sentences, achieving a good performance from deep neural networks would be difficult, since deep neural networks inherently performs better with more data. The next issue comes when labeling the data set because of not having a labeling tool for the Sinhala language.

After creating and labeling the data set, the next step is to select the deep learning technique to train a model. From the literature review, few deep learning methods that can be applied are gathered below,

1. Prepare the input using the sliding window approach and use deep neural network to train

2. Train data using LSTM

According to (Taylor et al., 2017) , both methods have provided good results. For animating the predicted output, most of the researchers have used the Maya software (Maya Community, 2019) which is a commercial product to animate the mouth shapes.

6.4 Limitations

The visemes sequence that are generated by the implemented model requires synchronizing with the input audio. In a speech utterance, speed can vary in different time frames. But this model was not able to handle those speech variations. It correctly animates the speech when the uttered at a constant speed. Therefore, if the speech is subjected to different speed variations, the model performs poorly when synchronized with the audio.

According to the subjective evaluation, the model is not capable of creating accurate speech animation for long sentences. Thus, when the number of transitions between words increases, the model performs poorly.

6.5 Implications for further research

Speech models can be implemented from different approaches such as static visemes , dynamic visemes and deep learning approaches. This research work followed the static visemes approach and obtained a model accuracy of 69% for subjective evaluation. Further research enhancements can be conducted to address the limitations of this research work which are handling different speeds of speech and long sentences with higher word transitions. In addition to that, further research can be conducted using deep learning or dynamic visemes approaches to increase the accuracy of the models.

References

Afouras, T., Chung, J. S. and Zisserman, A. (2018), 'LRS3-TED: a large-scale dataset for visual speech recognition', *CoRR* **abs/1809.00496**.
URL: *http://arxiv.org/abs/1809.00496*

Aghaahmadi, M., Dehshibi, M. M., Bastanfard, A. and Fazlali, M. (2013), 'Clustering persian viseme using phoneme subspace for developing visual speech application', *Multimedia Tools and Applications* **65**, 521–541.

Anderson, R., Stenger, B., Wan, V. and Cipolla, R. (2013), Expressive visual text-to-speech using active appearance models.

Beeler, T., Hahn, F., Bradley, D., Bickel, B., Beardsley, P., Gotsman, C., Sumner, R. and Gross, M. (2011), 'High-quality passive facial performance capture using anchor frames', *ACM Trans. Graph.* **30**, 75.

Belhumeur, P., Jacobs, D., Kriegman, D. and Kumar, N. (2011), Localizing parts of faces using a consensus of exemplars, pp. 545–552.

Blender Online Community (2017), *Blender - a 3D modelling and rendering package*, Blender Foundation, Blender Institute, Amsterdam.
URL: *http://www.blender.org*

Bonamico, C. and Lavagetto, F. (2001), Virtual talking heads for tele-education applications.

Cao, C., Bradley, D., Zhou, K. and Beeler, T. (2015), 'Real-time high-fidelity facial performance capture', *ACM Transactions on Graphics* **34**, 46:1–46:9.

Cootes, T., Edwards, G. and Taylor, C. (2001), 'Active appearance models', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **23**, 681 – 685.

Deena, S., Hou, S. and Galata, A. (2010), Visual speech synthesis by modelling coarticulation dynamics using a non-parametric switching state-space model, p. 29.

Dehshibi, M. M. and Shanbezadeh, J. (2014), 'Persian viseme classification using interlaced derivative patterns and support vector machine', *Journal of Information Assurance and Security* **9**, 148–156.

Eisert, P. (2003), Immersive 3-d video conferencing: Challenges, concepts, and implementations, Vol. 5150, pp. 69–79.

Ezzat, T., Geiger, G. and Poggio, T. (2004), Trainable videorealistic speech animation, Vol. 3, pp. 57– 64.

Fan, B., Wang, L., Soong, F. and Xie, L. (2015), Photo-real talking head with deep bidirectional lstm.

Hazen, T., Saenko, K., La, C.-H. and Glass, J. (2004), A segment-based audio-visual speech recognizer: data collection, development, and initial experiments., pp. 235–242.

Jamaludin, A., Chung, J. S. and Zisserman, A. (2019), 'You said that?: Synthesising talking faces from audio', *International Journal of Computer Vision* .

Luo, C., Yu, J., Li, X. and Wang, Z. (2014), Realtime speech-driven facial animation using gaussian mixture models, pp. 1–6.

M, R. (2018), 'Review on motion capture technology', *Global Journal of Computer Science and Technology* .
URL: *https://computerresearch.org/index.php/computer/article/view/1851*

Makehuman Community (2017), *Makehuman - open source tool to make 3D human characters*, Makehuman foundation, Makehuman Institute.
URL: *http://www.makehumancommunity.org/*

Matthews, I. and Baker, S. (2004), 'Active appearance models revisited', *International Journal of Computer Vision* **60**.

Mattheyses, W., Latacz, L. and Verhelst, W. (2013), 'Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis', *Speech Communication* **55**, 857–876.

Maya Community (2019), *Autodesk Maya - 3D computer animation, modeling, simulation, and rendering software*, Maya Foundation, Maya Institute.
URL: *https://www.autodesk.com/products/maya/overview*

Montgomery, A. A. and Jackson, P. L. (1983), 'Physical characteristics of the lips underlying vowel lipreading performance', *The Journal of the Acoustical Society of America* **73**(6), 2134–2144.

Ostermann, J. (1998), Animation of synthetic faces in mpeg-4, pp. 49 – 55.

Papagayo Community (2011), *Papagayo - a 3Lip Sync tool*, Papagayo Foundation, Papagayo Institute.
URL: *https://my.smithmicro.com/papagayo.html*

Power, M., Power, D. and Horstmanshof, L. (2007), 'Deaf people communicating via sms, tty, relay service, fax, and computers in australia', *Journal of deaf studies and deaf education* **12**, 80–92.

Rodríguez-Ortiz, I. (2008*a*), 'Lipreading in the prelingually deaf: What makes a skilled speechreader?', *The Spanish journal of psychology* **11**, 488–502.

Rodríguez-Ortiz, I. (2008*b*), 'Lipreading in the prelingually deaf: What makes a skilled speechreader?', *The Spanish journal of psychology* **11**, 488–502.

Rong, C., Zhenjun, Y., Yuan, W. and Yu, Y. (2012), 'Research on chinese viseme based on fuzzy clustering and grey relation analysis'.

Setyati, E., Susandono, O., Zaman, L., Pranoto, Y., Sumpeno, S. and Hery Purnomo, M. (2017), Establishment of indonesian viseme sequences using hidden markov model based on affection, pp. 275–280.

Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A., Hodgins, J. and Matthews, I. (2017), 'A deep learning approach for generalized speech animation', *ACM Transactions on Graphics* **36**, 1–11.

Taylor, S., Mahler, M., Theobald, B.-J. and Matthews, I. (2012), Dynamic units of visual speech, pp. 275–284.

Wasala, A. and Gamage, K. (2020), 'Research report on phonetics and phonology of sinhala'.

Wikipedia contributors (2019), 'Elbow method (clustering) — Wikipedia, the free encyclopedia', `https://en.wikipedia.org/w/index.php?title=Elbow_method_(clustering)&oldid=930355592`. [Online; accessed 18-February-2020].

Zhou, H., Liu, Y., Liu, Z., Luo, P. and Wang, X. (2018), 'Talking face generation by adversarially disentangled audio-visual representation'.

Zhou, Y., Xu, Z., Landreth, C., Kalogerakis, E., Maji, S. and Singh, K. (2018), 'Visemenet: Audio-driven animator-centric speech animation', *ACM Transactions on Graphics* **37**, 1–10.

Appendices

Appendix A

Code snippets

A.1 Rules Appendix

- Replace first syllable schwa with /a/

  If the first phone is a consonant and If word length > 2 (omit one or two-letter words) and If the First phone is /k/ and second phone is schwa and the third phone is NOT /r/, then Replace schwa with /a/ or If second phone is Schwa, replace it with /a/.

- Occurrences with /r/ and /h/

  If the phone is /r/ and if word length > 2 and if the preceding phone is a consonant and /r/ is followed by /@/ and /@/ is followed by any consonant (/h/ or !/h/) ,replace it with /a/ and if the phone is /r/ and If word length > 2 and If the preceding phone is a consonant and /r/ is followed by /a/ and /a/ is followed by any consonant other than /h/ , replace it with /@/

- If /a/, /e/, /ae/, /o/, /@/ is followed by /h/ and /h/ is preceded by /@/

  If the phone is one of the given phones (”a”, ”e”, ”ae”, ”o”, ”@”) and If that phone is followed by /h/ and If /h/ is preceded by /@/ , replace it with /a/

- If /@/ is followed by a consonant cluster

  If the phone is /@/ and If the next two phones are consonants ,replace it with /a/

- If /@/ is followed by words final consonant

If word length > 3 and If the phone before final phone is /@/ and If final phone is not in selected list(”r”, ”b”, ”t”, ”d”) and If the final phone is a constant ,replace it with /a/

- If word’s final syllable is ’yi’ or ’wu’ and it is preceded by /@/

  If word length > 4 and If the last syllable of the word is ’yi’ or ’wu’ and If it is preceded by /@/ , replace it with /a

- If /k/ is followed by /@/, and /@/ is followed by /r/ ot /l/ and then by /u/

  If the considering phone is /k/ and If /k/ is preceded by /@/ and If next phone is /r/ or /l/ and If /r/ or /l/ is followed by /u/ , replace it with /a/

- If word start’s with /kal/ in several words as follows, /kal(a:|e:|o:)y/->/k@l(a:|e:|o:)y/ /kale(m|h)(u|i)/->/k@le(m|h)(u|i)/ /kal@h(u|i)/->/k@l@h(u|i)/ /kal@/->/k@l@/

  If word length >= 5 and If word starts with /kal/ , /kal(a:|e:|o:)y/->/k@l(a:|e:|o:)y/ , replace it with /@/

  If word length >= 6 and If word starts with /kale/, /kal@h(u|i)/->/k@l@(u|i)/, replace it with /@/ If word length < 5 and > 3 and If word starts with /kal@/, /kal@/->/k@l@/ , replace it with /@/

Appendix B

Model evaluation questionnaire

B.1 Questionnaire Appendix

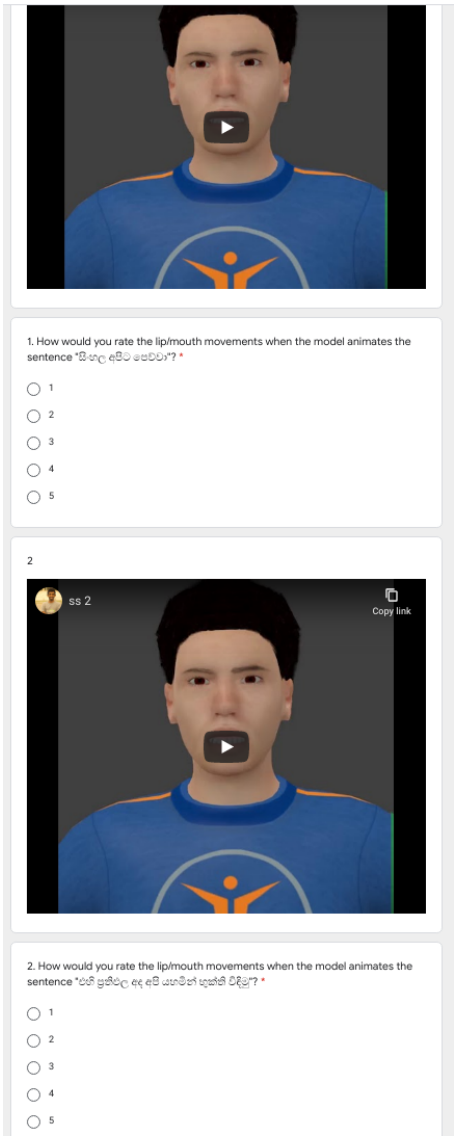

Figure B.1: sample question of the short sentences evaluation google form

B.2 Responses Appendix

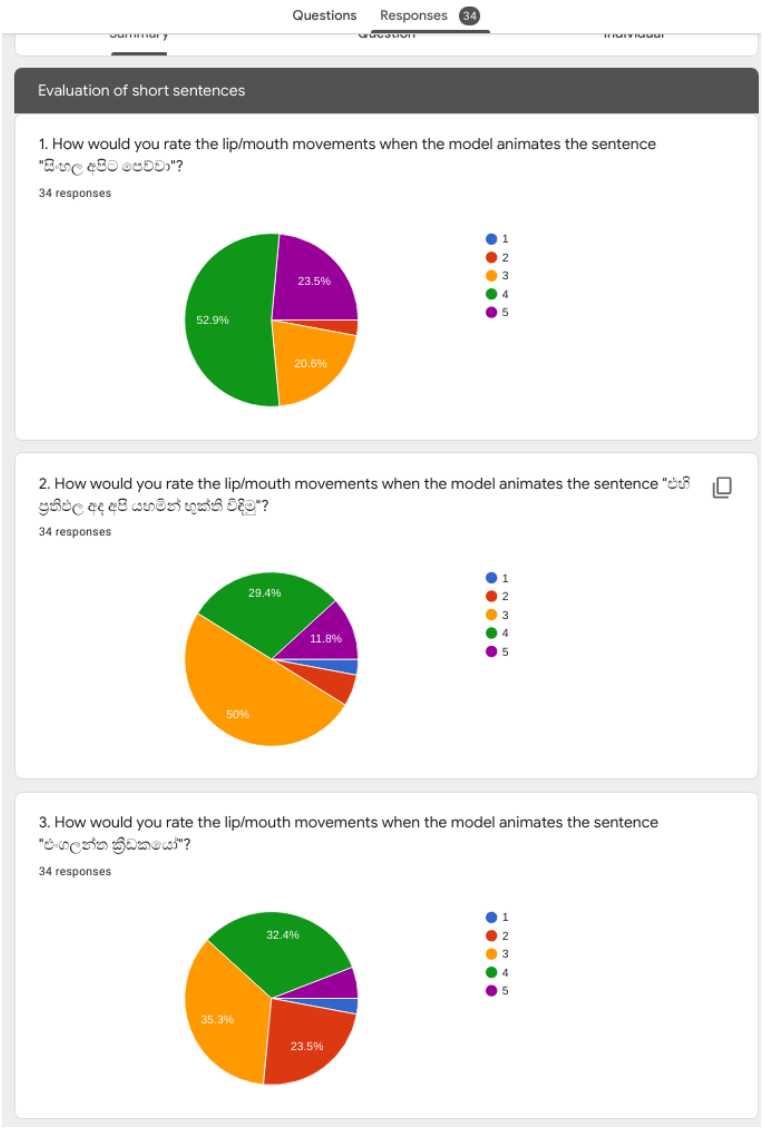

Figure B.2: sample responses of the short sentences evaluation google form