Speaking Lip Animation based on Active Appearance Models: A Unified Framework for Malayalam Visual Speech Synthesis

Thesis
submitted to the University of Calicut
in partial fulfilment of the requirement for the award of

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

By

SANDESH E. PA

Under the guidance of

Dr. Lajish. V. L.
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF CALICUT
KERALA, INDIA - 673635

FEBRUARY 2019

UNIVERSITY OF CALICUT DEPARTMENT OF COMPUTER SCIENCE

Dr. Lajish. V. L.
Assistant Professor

Calicut University (P.O.)
Kerala. India-673635

CERTIFICATE

This is to certify that the thesis entitled "**Speaking Lip Animation based on Active Appearance Models: A Unified Framework for Malayalam Visual Speech Synthesis"** is a report of the original work carried out by Mr.Sandesh. E. PA. under my supervision and guidance in the Department of Computer Science, University of Calicut, Kerala and that no part thereof has been presented for the award of any other degree.

Dr. LAJISH. V.L
Head of Department
Department of Computer Science &
Director, Calicut University Computer Centre
University of Calicut, Kerala-673635, INDIA

CU Campus
February 28 , 2019

DECLARATION

I hereby declare that the work presented in this thesis is based on the original work done by me under the supervision of Dr.Lajish V.L., Department of Computer Science, University of Calicut, Kerala and that no part thereof has been presented for the award of any other degree.

Sandesh E.PA

CU Campus
February 28 , 2019

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere thanks to my research guide Dr. Lajish. V.L., Head, Department of Computer Science, University of Calicut, Kerala. I appreciate his patience, ideas, and contributions of time, immense knowledge and encouragement which made my Ph. D. experience enlightening and productive. I wish to express my deep felt gratitude to him for the insightful guidance and invaluable help in steering the course of this study.

I extend my gratitude towards Dr.K.Samudravijaya, Scientific Officer, Tata Institute of Fundamental Research, Mumbai, India. I am deeply obliged to Dr. Nagabhushan, Professor and Director, IIIT Allahabad, India for his inestimable support. I am indebted to Dr. M. Sreenathan, Professor, Department of Linguistics, Thunchath Ezhuthachan Malayalam University, Kerala for valuable suggestions and encouragement.

I express my gratitude to Dr. R. K. Sunil Kumar, Assistant Professor, Department of IT, Kannur University, Kerala, India for his advice, guidance and also for his care and concern over the past years. I am greatly indebted to Dr. G. Harikrishnan, Assistant Professor, Government College Madappally and Syke A.K. HSA GHSS chorode and Executive Commitee Member, Digital wing of International Association of Teachers of English as a Foreign Language for their insightful reviews and support. I thank Mr. K. Suresh Babu, Former Associate Professor & Head, Department of Physics, Govt. College Madappally, Kerala and Dr. P. Ramakrishnan, Former Principal, Government College Madappally Kerala for extending their valuable suggestion and support during the tenure of the work. I whole heartedly thank Mrs. Manjula K. A., Assistant Professor, Department of Computer Science, University of Calicut, Kerala, for the encouragement and support given during the entire course of my work.

My special thanks are due to my research collegues Dr.Vivek P, Dr. Benson C C, Mr. Baiju K. B, Ms. Reshma P.K, Mr. Bibeesh K, Ms. Habeebath K.P, Mr.Anoop K. and Ms. Manjary P.Gangan for their support and collaboration during the course of this work.

Last, but certainly not the least I offer my thanks and appreciations to family and friends whose encouragement and support was unremitting source of inspiration for this work.

Sandesh E.PA

Dedicated to

Friends and Family

ABSTRACT

This work proposes a unified visual speech synthesis framework for Malayalam and lip movement animation synthesis is performed with given grapheme sequence as input. Visual speech synthesisers are used in diverse applications including man machine interfaces, animated movies and assistive technology solution for deaf people. The current research and development in the domain of visual speech synthesis is oriented more towards developing systems which can learn from utterances. The proposed framework uses Active Appearance Models (AAMs) for learning and synthesising facial movements. Comprehensive, carefully compiled and annoted audio visual speech corpora, developed as part of this work is used for training the AAM based Malayalam visual speech synthesiser. Many sub modules in the framework, including durational model, transcriptror and set of visually discernable phonemes are also developed as part of this work. An extensive study is conducted to model the colour peculiarities of mouth region pixels using the facial images in the audio visual speech corpora. ASM (Active Shape Model) and Convolution Neural Network (CNN) based lip tracking systems are also implemented using landmark points detection strategy. Most of the lip annoted images for the AAM based lip movement synthesiser is obtained through the lip tracking systems.

Grapheme to audio visual speech synthesis systems need to understand text, audio and visual domains of language representations. Standardisations, mappings from one domain to another and statistical analysis based on atomic units in each domain need to be carried out

for the target language. Graphemes are the basic unit in the textual representation of a language. Phonemes, defined as the smallest distinctive sound units in a language, are considered to be the basic unit for speech. But the properties of phonemes exhibit wide variations based on its position in the word and context. In Malayalam, phonemes are further categorised in to allophones based on the positional and contextual variability, i.e. the contextual and positional variability is encoded in the allophone characterisation of Malayalam language. This computational linguistic feature of Malayalam is exploited in designing many components of the proposed work. Corpora in text, audio and visual modalities are created for the training and the conduct of various experiments related to lip tracking and visual speech synthesis. The developed bench marked data sets including the word based text corpora, isolated phoneme audio visual speech corpora and isolated word audio visual speech corpora are designed and developed by accommodating the allophonic variability in Malayalam.

In the first phase of this work, phoneme and allophone based durational models are developed. In a text to speech synthesis system, the durational model predicts the duration sequence corresponding to the speech segment sequence. Most of the co-articulation effects in Malayalam are encoded in the allophonic characterisations for each phoneme. So an allophone centric durational model is developed in this work. The silence corresponding to the articulatory formation stage for the anticipatory plosive is identified to be the main factor creating audio visual asynchrony in Malayalam. This intra phoneme silence is analysed from the speech corpora and a silence model is developed based on statistical methods. The silence duration information is incorporated to the durational model and used in the visual speech synthesis framework. A rule based text to phoneme and allophone transcription system is developed as part of this study. The sequence of graphemes is converted first in to a sequence of phonemes and then to the corresponding sequence of allophones.

The next phase of the work addresses the issues pertaining in the visual domain of Malayalam speech. Identifying the viseme set in a language is the fundamental step. Viseme set is the set of visually discernable mouth appearances, which is considered as the basic unit of visual speech. Many to one phoneme to viseme maps for Malayalam are developed using linguistic knowledge, perception testing and data driven approaches. Linguistic viseme set is formed by exploiting the expert knowledge in the linguistics of the language. Experiments are conducted to understand the visual perception of Malayalam phonemes by human subjects. Viseme set is prepared by recording and analysing the responses of human subjects. Geometric features of lips and Discrete Cosine Transform (DCT) based features of lips are used for data driven clustering. Viseme set is formed by clustering in the feature vector space. Hierarchical Agglomerative Clustering (HAC) is used for clustering. In HAC, initially each instance is treated as a separate cluster to iteratively form a single cluster accommodating all instances. Many to many phoneme to viseme maps is also developed using data driven method by taking allophone as the basic unit.

The proposed visual speech synthesis framework is applied for synthesising mouth area movements with lips as the boundary. Mouth area mainly consists of lip, skin neighbourhood of lips, teeth, tongue and dark portions of mouth cavity. The colour of these mouth regions shows significant variation with respect to ethnicity. The study as part of this thesis performed the statistical and probabilistic analysis of colour of mouth region pixels in Indian context. The analysis is performed in different colour spaces. The colour spaces are ranked according to their performances against a multiclass Bayesian classifier. Lip tracking by land mark point localisation is another major problem addressed in this thesis. Active Shape Modelling (ASM) and Convolution Neural Networks (CNN) are used for lip land mark point localisation. The ranking of colour spaces is exploited in designing the ASM based lip tracking system. The ASM annoted images are used for training CNNs and as training images for AAM based visual speech synthesis.

In the last phase of the work all the components developed is integrated in to a text to visual speech synthesis frame work using independent Active Appearance Models (AAM).The lip movements corresponding to a Malayalam text is synthesised using the framework. The frame work uses durational models, transcriptors, viseme set and tracking modules to develop the AAM based synthesiser.The conversion of input text into a sequence of atomic grapheme units is the first step in the synthesis process pipeline. The atomic unit can be a sequence of phonemes, allophones, visemes obtained from phonemes and visemes obtained from allophones. Finding the number of frames for each atomic unit is the second step in the synthesis framework, which uses the duration of actual utterance and average unit duration.

The visual speech synthesiser in Malayalam using independent AAM creates the shape and texture of lip movements' separately. The appearance and shape coefficients of frames corresponding to a phoneme or allophone are stored in a code book. Intermediate frames are generated using morphing in the coefficient space or morphing in the image space. Morphing in the image space is observed to be generating impossible frames. Finally, perception experiments are conducted with different visual basic units using morphing in the coefficient space based framework.From the experimental results, it is concluded that allophone is the perfect Atomic Visual Unit (AVU) for Malayalam visual speech synthesis.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1. Background

Language is not just a tool for expressing ideas, but a social, historical and aesthetic phenomenon. Knowledge, arts, literature and every such form of human achievements is born, evolved and memorised in one of the languages. Language is expressed either as speech or text. Both forms are now used equally, even though every language is evolved through speech. In the initial stage of learning any language, one learns the conversion of speech to text and text to speech after familiarising with the basic symbols in each domain. To automate these conversions, scientists have been exploring various machine learning techniques. In this regard, the last decade witnessed plenty of landmark achievements. The perception of speech is primarily based on the audio cues generated from the speech signal. But the visual cues during speech are equally relevant in the speech perception and in the communication of non-linguistic factors such as emotional state. The emphasis and punctuations added through facial expressions make direct conversation an effective mechanism of communication. This makes the design and development of visual component of speech in the man machine interface and speech synthesisers noteworthy.

The visual component of speech manifests as facial motion. Modelling the facial motion has been a subject of research for more than hundred years. Before the computing era, hand crafted cartoons of visual speech were generated by successively drawing a predefined set of mouth shapes. Computer generated human faces in character animations became a reality in 1970s. Over the years intelligibility and realism of the synthesised face images have improved drastically through incessant research and development in the domain. The input to the facial motion synthesiser can either be a speech signal (speech driven systems) or a grapheme sequence (text driven systems). The techniques for generating synthetic visual speech can be broadly classified as image based and model based approaches. The effective implementations of model based approaches use HMM (Hidden Markov Modelling), Active Appearance Models (AAMs) and Artificial Neural Networks (ANNs). Lip movement synthesis, a sub problem of facial animation creates realistic lip movements synchronised with the audio. Creation of realistic lip movements is one of the critical and costly component in character animation based multimedia productions. The mouth area movement studies are also significant for developing sign language synthesisers for deaf community.

The tools in the domain of language computing have been attaining multilingual capability like never before and audio visual speech synthesis is no exception. The disparity in the lip movements of characters with the spoken words in multilingual productions is identified to be the most distracting factor effecting the appreciation. In this context, a production framework accommodating multilingual lip movement for characters is springing up across the globe. Audio-visual speech synthesis in a language demands many language specific linguistic and computational explorations. Text to audio visual speech synthesis systems have to work in text, audio and visual domains of language representations. Standardisations, mappings from one domain to another domain and statistical analysis need to be carried out for the target language. Many of the Indian languages including Malayalam still continues to be a low resource language in the language computing domain.

This thesis performs analysis and modelling supporting the development of tools for visual speech synthesis in Malayalam. This spans several areas of work such as transcriptions from text to phonetic representations, mappings to unique mouth shapes, modelling co-articulation effects and finally combining them to build a framework for visual speech synthesis. Importance is given to the analysis of the mouth regions of talking faces and the experiments on model based lip movement animation in Malayalam.

1.2. Motivation

Conversing in the mother tongue of a person, in which he/she is evolved into a social being, is natural, fast and comfortable for a human being. So equipping the machine to communicate in a person's mother tongue is one of the main goals of research and development in our time with remarkable achievements. To exploit this favourable situation Malayalam still has to complete many preliminary investigations in the domain of language computing. This thesis is a collection of investigations towards bringing out the inherent computational features of Malayalam language. The limitations of the current practice of localising and adapting English language computing tools to Indian regional languages is evident in many spheres. It happens due to the considerable disparity in the computational linguistic structure of Indian languages with English like languages.

A bench marked data set is the prime constraint in the research and development of language computing tools. An audio visual speech corpus in Malayalam has been developed as part of this work with the help of linguistics, which can be used for audio visual speech synthesis and recognition. The data is procured by considering allophone as the basic speech unit. Malayalam is one of the few languages which have got a well-defined rule set for allophone characterisation. Allophone centric speech processing attempts have started recently in Malayalam. Allophone Transcription systems need to be developed for further advances in this direction. This work has developed a text to allophone transcription system for Malayalam language.

The importance of visual component in speech perception is an established fact. The intimacy and feeling of completion in face to face communication can be mainly attributed to visual information. Facial motion image sequence synthesis corresponding to an utterance is important for man-machine interface and facial animation automation. Analysis of Parametric representations of Visual speech modality for classification performed in this work is crucial both for automatic lip reading systems and lip motion synthesis systems. Even personalised versions of facial motion synthesis systems are emerging in the world.

Synthesis by analysis is the most successful approach. Colour is the most prominent component in the analysis and its variation across ethnicities is studied and established by many researchers. So the thesis performed an analysis to understand the colour probabilistic nature of mouth region components such as lip, teeth and tongue. Point Distribution Models (PDM) with learning capabilities such as Active Shape Model (ASM) for segmenting deformable shapes is applied for lip segmentation. The massive visual corpus of talking mouth region with land mark points for outer and inner lip, developed as part of this work, is used for trying lip segmentation using deep learning networks. A Convolution Neural Network (CNN) implementation is developed for lip segmentation and analysis is performed by changing network parameters. Finally the analysis performed on text, speech and visual forms of Malayalam speech is integrated to develop a visual speech synthesis framework for Malayalam. Active Appearance Model (AAM), trained on the own developed and labelled visual corpus, is the main component of speech animation framework derived out of this study for Malayalam Language.

1.3. Outline of the Thesis Organisation

The thesis is structured as 8 chapters. The review of literature in Chapter 2 places this work in the context of established and ongoing enquiries in the relevant areas. It starts with a general review on phoneme based duration models with special emphasis on models that captures contextual variability. The review in also conducted for the rule based, model based and data driven approaches to text to phoneme transcription. The works in languages like Malayalam with an established rule set for transcription and allophone characterisation is explored in detail. Visual phoneme or viseme is the visual speech equivalent of a phoneme or the set of visually separable phonemes. Viseme set in language is created either by a many to one phoneme mapping or many to many phoneme mapping. The review on viseme set formation encompasses all type of mappings using different techniques. Review on the analysis and segmentation attempts on different mouth regions such as skin, lip, tongue and teeth is also discussed. The detailed review on lip segmentation and tracking uses image based, model based and edge based techniques in lip segmentation and tracking is also performed. The final review section describes the various recent attempts reported for facial animation.

Chapter 3 explains the processes involved in the creation of comprehensive Malayalam audio visual speech corpora used in the work and consolidates the findings of in-depth investigations performed on the duration of allophonic variations of Malayalam phonemes, which will be used as the durational model of a visual speech synthesis system. Duration and its modelling are important cues in the intelligibility and naturality of synthesised audio visual speech. In a text to speech synthesis system durational model predicts the duration sequence corresponding to the speech unit sequence and those values are predicted based on many factors affecting duration. It is possible to capture most of positional and contextual variability from the textual representation of the sentence. A phoneme can appear in the start, middle and end of a word creating positional variability. The change in duration due to the effect of surrounding speech units is called contextual variability. The effect due to neighbouring phonemes and position are generally known as co-articulation effects. Co-articulation effects in Malayalam are modelled by Malayalam linguists as allophonic characterisations for each phoneme. A well-defined allophone formation rule set exists for Malayalam. This is exploited to develop a durational model accommodating co-articulation effects due to positional and contextual variability of phonemes by understanding the durational pattern of allophones.

Chapter 4 explains the implementation constituents of the proposed grapheme to phoneme and allophone transcripter. The given grapheme sequence, after appropriate pre-processing in the grapheme domain is converted first into sequence of phonemes, and this sequence is converted to the corresponding sequence of allophones. Grapheme to Phoneme transcripter is reported for many languages. The underlying implementation strategies can be classified into dictionary based, data driven and rule based approaches. Linguistic rule set can be formulated for phonetically perfect languages such as Sanskrit with written and spoken form correspondence. The transcripter uses a rule based approach both for grapheme to phoneme and phoneme to allophone transcription. The pre-processing performed for the transcription in the grapheme domain is explained in the initial section of the chapter. The next two sections explain the proposed grapheme to phoneme and phoneme to allophone transcription algorithms. The implementation of these transcriptors are verified by computing the frequency of occurrence of phonemes and allophones in a word corpora.

Chapter 5 illustrate different phoneme to viseme mapping strategies developed for Malayalam. The first section of the chapter explains the pre-processing performed on the images in the audio visual data set for executing the work. Visual phoneme or viseme is the visual speech equivalent of a phoneme or the set of visually separable phonemes. The nature of mappings from phoneme set to viseme set is either many to one or many to many. Many to one phoneme to viseme mappings are developed based on the linguistic knowledge, perception testing and data driven approaches. Geometric features extracted from lip region and Discrete Cosine Transform (DCT) of lip area are the features used for data driven clustering. The work exploited the benefit of an allophone set modelling co-articulation in Malayalam to design a many to many phoneme to viseme map. Data driven approaches are then used for developing allophone to viseme maps.

Chapter 6 reports the in-depth analysis and facial feature segmentation experiments performed on the own developed visual speech corpora. Mouth motion accounts for the prominent non-rigid facial motion, especially during talking. Tracking of mouth regions such as lips, teeth and oral cavity has potential applications both in speech recognition and facial motion synthesis. The performance of colour components in different colour space is compared against Bayesian mouth region segmentation algorithms on image sequences of the frontal face while talking. Lip segmentation is attempted as a separate module using Active Shape Model (ASM) and Convolution Neural Networks (CNNs). The performance of different colour components in different colour spaces is compared with Beysian classifier and ASM. The date set for training the CNN is obtained from manual landmarking and from ASM based segmentation in a single speaker mode.

Chapter 7 explains the visual speech synthesis framework developed for Malayalam using independent Active Appearance Models (AAMs). The framework uses the components developed as part of the work such as allophone based duration model, transcripter and viseme set. The system is conceptualised as a text to visual speech synthesis system with an additional component for receiving duration cue from actual utterance. The framework is successfully applied for generating lip movements corresponding to a Malayalam text. The lip landmarked dataset for AAM training is obtained by manual land marking and automatic techniques. Three separate models are implemented for lip movement synthesis, one for synthesising only shape and the other two synthesises both shape and texture. The process flow of the proposed visual speech synthesis developmental framework is shown in figure 1.1. Perception experiments are conducted by taking allophone, phoneme, viseme mapped from allophone and viseme mapped from phoneme as inputs to the synthesiser.

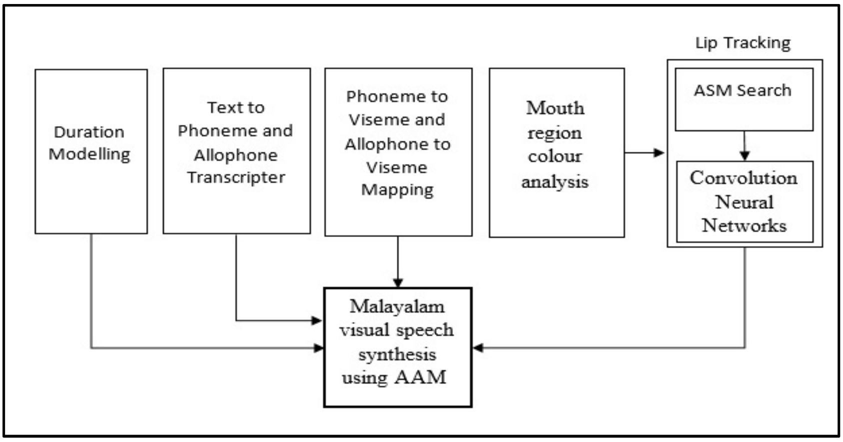

Lip Tracking
Duration Modelling
Text to Phoneme and Allophone Transcripter
Phoneme to Viseme and Allophone to Viseme Mapping
Mouth region colour analysis
ASM Search
Convolution Neural Networks
Malayalam visual speech synthesis using AAM

Figure 1.1: Process flow of the proposed visual speech synthesis developmental framework

Chapter 8 concludes the work with important directions of future research. Audio visual asynchrony modelling, phonetic transcripter which can model non textural cues, analysis-synthesis approach to speech animation in Malayalam and Deep Neural Network (DNN) based speech animation are the important domains identified for related future work.

CHAPTER 2

REVIEW OF RELATED WORK

2.1 Introduction

The bimodal nature of speech perception is an established fact. Talking face image sequence corresponding to an audio can create realistic conversation environment and will surely make machines more human-friendly. The movement of visible articulators creates the varying image sequence corresponding to visual speech. Hand drawing of expressive facial movements is time consuming and demands the involvement of expensive manpower. The lower jaw, tongue, teeth, and lips are the visible articulators of the human speech production system. The attempts to automise the movements of these organs started in 1970s, which is a basic polygonal mesh framework with mouth and eyes closing and openings. The last decades witnessed the emergence of a plethora of new techniques for facial movement automation, especially while talking. This chapter reviews the related work in various domains relevant to visual speech synthesis. This spans several areas of work such as transcriptions from text to phonetic representations, mappings to unique mouth shapes and modelling co-articulation effects. Importance is given to the visual speech element by performing the analysis of the mouth regions of talking faces analysis and the experiments on model based speech animation in Malayalam.

This study focuses on the development of visual speech synthesis framework in Malayalam, using Active Appearance Models (AAM). Active Appearance Model (AAM), a statistical data driven model is used for modeling lip shapes and texture. Independent AAM, which models shape and textures using separate set of coefficients, is employed in this work. The framework is conceptualized as a text to visual speech synthesis system. The detailed review of related works is structured in this chapter as follows. After the brief introduction in section 2.1, section 2.2 reviews the phoneme based and allophone based duration models and audio visual speech corpora reported from various languages. Section 2.3 lists the graphme to phoneme and allophone transcriptors developed in various languages using dictionary based, data driven and rule based approaches. Section 2.4 reviews viseme set generation attempts based on linguistic knowledge, perception experiments and data driven approaches. Section 2.5 reviews facial feature segmentation techniques reported in literature and section 2.6 reviews various visual speech synthesis frameworks reported in the literature. The review is concluded in section 2.7.

2.2 Review on Durational Model for Speech Synthesis and the Development of Audio Visual Speech Corpora

Allophone based speech processing studies are reported in many languages. Piotr Kozierski *et al.* use an allophone centric approach for polish whispery speech recognition [1]. The paper checked the use of extended allophone set to improve the speech recognition. Long Nguyen *et al.* analysed the most frequenly occuring CV (Consonant Vowel) pairs in the Japaneese language and added corresponding allophones to the phoneme set for speech processing[2]. Ten new phonemes are added from the frequently occurring CV unit list. In the work by Ji Xu *et al.* allophones representing pronunciation variations are automatically derived from Korean Speech data [3]. Gaussian Mixture Models (GMM) are used for measuring the candidate allophones and allophone set is derived using a decision tree based approach. The automatically derived allophone set is reported to outperform linguistically derived allophone set in experiments. A number of works uses allophones as the basic unit for concatenative speech synthesis.

Imed douben *et al.* integrated allophone based segments in to the data base for improving Arabic speech synthesis system [4].Report by Pavel A. Skrelin compares diphone and allophone based approach for Russian speech synthesis. In this work, the actual allophones are selected from utterances characterising the allophone phonetic context [5]. Barkhoda *et al.* compares the kurdish syllable, allophone, and diphone based speech synthesis systems [6]. Subjective tests with human evaluators are used for comparing the performance. Allophone based concatinative speech synthesis using neural networks is also attempted [7]. In their work different Linear Predictive Coding (LPC) based speech parameterisation is attempted and evaluated for concatenative speech synthesis. A sliding window input layer based neural network architecture is used in the implementation. Mazin *et al.* uses allophone/ diphone concatenation method for speech synthesis [8]. Alex carpov uses allophones and multi allophones for text to speech synthesis system [9]. An allophone rule set based strategy is employed by Ka-Ho Wong to develop a language learning framework [10]. Gregor A. Kalberer has used an allophone-viseme transcription system for visual speech synthesis [11][12].

Phoneme duration is dynamic and the variation is mainly attributed to effects due to neighbouring phonemes. Rule-based approaches, statistical approaches, model-based approaches and its combinations are used for representing this variability. Ommer *et al.*primarily used a statistical approach for duration modelling in Turkish TTS [13]. Robert et al uses a decision tree based approach for duration modelling of Czech speech segments [14]. A Classification and Regression Tree (CART) approach is used by Alexandros *et al.*to model the duration of Greek phonemes [15]. A Long Short Term Memory Recurrent Neural Network framework, which tries to model the countable duration values are used by Bo Chenn *et al.*[16]. A neural network based Arabic phoneme prediction system is proposed by Yasser Hifny *et al.*[17]. Giedrius performs a decision tree based phoneme duration analysis for Lithuanian language [18]. Janneet *et al.*employes expanded state HMM for modelling Finnish language duration modelling [19]. Yonas *et al.* developed a data driven duration model for Amharic phonemes using classification and regression trees [20]. To improve the intelligibility of synthesised speech an HMM and Multilayer Perceptron hybrid model duration prediction is introduced by kalu *et al.*[21]. Tree based machine learning approaches is also used for Serbian language duration modelling [22]. Table 2.1 summarises the durational modelling attempts in various Indian languages.

Table 2.1: Summary of the durational modelling works in various Indian languages

Sl. No	Title of the Paper	Author	Language	method/Work used
1	Duration characteristics of Hindi phonemes in continuous speech	K SamudraVijaya	Hindi	Technical Report [23]
2	Durational Characteristics of Hindi Stop Consonants	K SamudraVijaya	Hindi	Stop consonants, Continous Speech data base [24]
3	Modeling syllable duration in Indian languages using neural networks	K S Rao *et al.*	Hindi,Telungu,Tamil	Syllable duration, using Neural Network [25]
4	Modeling syllable duration in Indian languages using support vector machines	K S Rao *et al.*	Hindi,Telungu,Tamil	Syllable duration, using Support Vector Machines [26]
5	Duration Modeling Of Indian Languages Hindi And Telugu	N.Sridhar Krishna *et al.*	Hindi and Telungu	Phoneme duration CART [27]
6	Bengali Diphone Duration Modeling for Bengali Text to Speech Synthesis System	Labiba Jahan *et al.*	Benglai	Diphone duration for speech synthesis [28]
7	Significance of Duration in the Prosodic Analysis of Assamese	D.Govind *et al.*	Asameese	Duration modification according to Prosody [29]

8	Duration Modeling in Hindi	Somnath Roy *et al.*	Hindi	Syllable and Phoneme analysis for implementing prosody in speech synthesis [30]
9	Duration Modeling For Telugu Language with Recurrent Neural Network	V.S RameshBonda *et al.*	Telungu	Syllable Modelling using recurrent Neural Network [31]
10	Durational Characteristics of Indian Phonemes for Language Discrimination	B.L.Kanth *et al.*	Hindi , Tamil and Telungu	Duration Analysis for Language discrimination [32]
11	Duration Analysis for Malayalam Text-To-Speech Systems	D.P.Gopindah *et al.*	Malayalam	Statistical analysis of Vowel and Consonant duration [33]
12	Modeling of Vowel Duration in Malayalam Speech using Probability Distribution	D.P.Gopindah *et al.*	Malayalam	A probabilistic Duration prediction system [34]

Audio visual database is the basic building block of audio visual speech processing applications. A standard isolated phoneme based and word based audio visual database is developed as part of this work. Table 2.2 lists the important audio visual speech corporas developed in different languages.

Table 2.2: Audio visual speech corpora developed for different languages

Database - Year	Speaker (Female, Male)	Corpus - Repetition	Video Parameters (Pixel size, FPS)	Audio Parameters (Sampling Frq., bit)	Special Features
TULIPS1[01] - 1995	12 (9,3)	First four Englishdigit – twice.	100 x 75, 30 fps. Mouth region.	Unknown	Isolated digits. [35]
DAVID [02] – 1996	124	Isolated digits. English-alphabet E-set. Video-conference control commands. 'VCVCV' nonsense utterances.	640 x 480, 30fps. Full face. Frontal View.	Unknown	Speech/Person recognition. Contain 4 corpus with different research theme. Complex background and variable illumination. Contain individual speaker and more than one speakers during recording. [36]
M2VTS [03] – 1997 Multi Modal Verification	37	Numbers (0 to 9) – 5 times.	286 x 350, 25fps.	48kHz, 16 bit.	Speech verification, face recognition.

for Teleservices and Security applications			Full face. Frontal view.		Mostly French Speakers. Head rotation (left, right, up and down). Presence of glasses and hats. [37]
XM2VTSDB [04] – 1999 Extended M2VTS Database	295	Three sentences (numbers and word) – twice.	720 x 576, 25fps. Full face. Frontal view. 2 Camera used.	32kHz, 16 bit.	Personal identity verification. Head rotation (left, right, up and down). Recorded in extremely controlled condition. Text dependent. [38]
AMP/CMU [05] – 2001 Advanced Multimedia Processing Lab	10 (3,7)	78 Isolated words (date and time, month, day and miscellaneous) – 10 times.	720 x 480, 25 fps. Full face. Frontal view.	16kHz, 16 bit.	Lip reading. Presence of glasses and hats. Recorded in controlled situation. [39]
AV Letters [06] – 2002	10 (5, 5)	Isolated letters (A to Z) – 3 times. Total 780 utterances.	376 x 288, 25fps. Full face. Frontal view.	22.05kHz, 16 bit.	Speech recognition. [40]
CUAVE [07] – 2002 Clemson University Audio Visual Experiments	36 (19, 17) Speaker pairs -20	Isolated digits. Connected digits. Total 7000 utterances.	720 x 480, 29.97 fps. Shoulder and head. Frontal view.	16kHz, 16 bit	Speaker independent digit recognition. Speaker independent database. Contain individual speaker & speaker pairs. Head movement in side-to-side, back-and-back.

					Presence of glasses, facial hairs and hats. Fit to one DVD-data disk. [41]
VidTIMIT [08] - 2002	43 (19, 24)	10 TIMIT sentences per speaker. First 2 sentences are same for all but remaining 8 are unique.	512 x 384, 25 fps. Frontal view.	32kHz, 16 bit.	Multi-modal person verification. Data acquisition with 3 session. Extended head rotation. Change in speaker appearance and voice during each session. Variability in camera zoom factor and background noise. [42]
BANCA [09] - 2003	52 (26,26) for each language class.	Numbers. Names. Addresses. Date of birth.	720 x 576, 25fps. Shoulder and head. Frontal view. 2 Camera used.	32kHz, 12& 16 bit. 2 Microphone used.	Multi-modal identity verification. 4 Languages-English, French, Italian and Spanish. Recorded in controlled, degraded and adverse condition. Text independent. [43]
AVICAR [10] – 2004 Audio-Visual Speech in a Car.	100 (50, 50)	Isolated digits. Isolated letters. Phone numbers. TIMIT sentences. Total 59,000 utterances.	720 x 480, 30fps. Full face. 4 Camera array.	48kHz, 16 bit. 8 Microphone array.	Speech recognition in car. 60% American English others Latin American, European, East Asian and South Asian.

					Recorded in 5 noisy condition (automotive noise). [44]
AV-TIMIT [11] - 2004	223 (106, 117)	450 TIMIT-SX sentences. Each speaker utter 20 sentences. First sentences is common and other 19 sentences are different.	720 x 480, 30 fps. Full face. Frontal view.	16kHz, 16 bit.	Speaker independent continuous speech recognition. Continuous phonetically balanced speech. Contain multiple speakers. Controlled office environment. Presence of facial hairs, glasses and hats. Recorded in different illumination condition. [45]
VALID [12] – 2005	106 (29, 77)	XM2VTS speech corpus.	576 x 720, 25 fps. Full face with shoulder. Frontal view.	32kHz, 16 bit.	Multi-modal speaker/speech recognition. 97 Europeans and 9 Asians. 5 Recording session – 1 controlled and 4 uncontrolled (varying noise, illumination). Presence of facial hairs. Text dependent. [46]
GRID [13] – 2006	34 (16, 18)	Command sentences. Each sentence contain six	720 x 576, 25 fps	25kHz, 16bit	Speech recognition. Mean age is 27.[47]

		word sequence. Total corpus size 34,000.			
AV Letters 2 [14] - 2008	5	26 Isolated letters – 7 times.	1920 x 1080, 50 fps. Full face. Frontal view.	48kHz, 16 bit.	Speech recognition. High-definition version of AV Letter database. [48]
UWB-07-ICAVR [15] – 2008 University of West Bohemia-2007-Impaired Conditions audio visual speech Recognition	50 (25, 25)	200 Sentences (50 shared and 150 unique). Total 10,000 continuous utterances.	720 x 576, 50 fps (high quality). 640 x 480, 30 fps (low quality). 2 Camera used.	44kHz, 16 bit. 2 Microphone used.	6 types of illumination condition Average age is 22. Czech language. [49]
IV2 [16] – 2008	300	15 French sentences.	780 x 576, 25 fps (high quality). 640 x 480, 25 fps (low quality). 2 Camera used. Full face. Frontal and profile view.	2 Microphone used.	Face recognition. Majority data acquisition within single session. Pose, expression, illumination and glass variability. Different illumination levels and orientations. Iris image, 3D laser scanner face data. [50]
DXM2VTS [17] – 2008 Damascened XM2VTS	295	XM2VTS database. Additional videos containing several degradation level of background noise.	720 x 576, 25 fps. Full face. Frontal view. 2 Camera used.	32kHz, 16 bit.	Face recognition. Internal video distortion (blur, salt and pepper and rotation). External video distortion (zooming and dynamic

					background noise). [51]
OuluVS [18] - 2009	20 (3, 17)	10 daily use short phrases – 9 times. Total 817 sequences.	720 x 576, 25 fps. Full face. Frontal view.	No audio	Visual only recognition. [52]
WAPUSK20 [19] - 2010	20 (9, 11)	100 GRID database sentence. Total 2000 sentences.	640 x 480, 32 fps. Full face. Frontal view.	16kHz, 32 bit. 4 audio channels	Stereoscopic video. Speakers – 2 England, 1 Greece, 1 Kazakhstan and 1 Spain. All other native German speakers. Mean age is 29.[53]

The following section describes a review of transcriptors developed for various languages.

2.3 Review on Transcripters in Various Languages

Grapheme-to-Phoneme transcription plays an important role in speech processing applications such as speech recognition, speech synthesis and speech database construction. This section presents a review of methods used for the development of grapheme to phoneme and grapheme to allophone transcripters in various languages. The methods can be classified as linguistic,dictionary based and data driven. Rule-based approaches uses expert linguistic and phonetic knowledge for transcription, whereas dictionary-based methods relies on storing maximum phonological information and using it for transcription. Data-driven methods process trained data for automatically deriving the converter predominantly using machine learning techniques.

Rule-based transcription system is applicable to languages having a defined correspondence between graphemes and phonemes. A regular correspondence exists for Arabic language and it is exploited for designing a transcription system in Arabic. Work by Yousif A. El-Imam consist of a text segmentation and pre-processing stage. Pre-processing is used for generating well-formed lexical items [54]. Bangla Grapheme-to-Phoneme (G2P) converter, developed by Ayesha *et al.* is based on some predefined rules for general cases and some specific rules to handle the exceptional case. Bangla pronunciation generator faces challenges like distinguishig the different vowel pronunciations. Heuristic assumptions are required in some situations [55]. Miohel Divay *et al.* designed a Grapheme to phoneme transcription for French based on letter-to-sound rules. In this system the left hand side of each rule and the right hand side of each rule specifies the corresponding phonemes with the preceding and succeeding grapheme context [56]. In a work for Lithuanian language, words grapheme to phoneme conversion take place based on encoded knowledge. The conversion problem is divided into three consecutive algorithmic steps, syllable boundary identification, accentuation, and transcription. Linguistics engineering approach is used for designing specific rule set [18].

Dictionary-based letter-to-phoneme transcription relies on storing maximum phonological knowledge in a lexicon. A comprehensive dictionary needs huge memory and tedious effort during creation. Jack Halpern on his study on Japanese speech technology discusses the complex allophonic variations based on a phonetic database. This database used directly in the development of acoustic models and can produce more accurate phonetic transcription [57]. Jozsef Domokos *et al.* presents a machine-readable language dictionary for Romanian language. The dictionary is available in a format, which is compatible to HTK and festival speech synthesis system. In this study parallel structure having Artificial Neural Network is used to design phonetic dictionary. This type of pronunciation dictionaries are very useful resources for spoken language technologies and used in ASR and TTS applications [58].

Attila Novak *et al.* in Hungarian used the dictionary based method by using database of Hungarian geographic terms into phonetic representation. The construction of this dictionary includes several steps. After collecting word forms from large written corpus the resulting words are cleaned and applied the transformation rules. Exceptions occurred during transformation process were corrected manually [59]. Researchers from Myanmar Ye Kyaw Thu *et al.* presented a grapheme to phoneme conversion using Myanmar dictionary [60]. Transcription for English and Thai developed by Aroonmanakun *et al.* implements a version of dictionary based approach. Transcriptions of Thai and English words are manually stored in the dictionary. Transcriptions of most of the words can be retrieved from the dictionary. The exceptional word set are generated by applying some special rules [61]. An English grapheme-to-phoneme system is developed to handle silent English words [62]. As part of the study carried out in Korean, Byeong chang kim developed a grapheme to phoneme converter using dictionary-based and rule-based hybrid method with phonetic pattern dictionary and letter-to-sound rules. The phonetic pattern dictionary, standing for the dictionary-based method, contains entries in the form of a morpheme pattern and its phonetic pattern. The conversion method consists of mainly two steps including morpheme to phoneme conversion and morphophonemic connectivity check [63].

In data-driven methods, the algorithm for transcription is learned automatically from data. Yousif A El. Imam explains different data driven approaches such as pronunciation by analogy (PbA), statistical methods based on stochastic theory and methods based on neural networks. PbA understands the pronunciation using similar parts of known words and their pronunciation. The statistical method sucessfully learns the grapheme–phoneme mapping with a certain probability. Trained neural networks using multilayer perceptrons

(MLP) with back-propagation training also gives promising results [1]. In Arabic, Khalid Nahar uses data driven methods to recognize and automatically transcribe Arabic phonemes. In their work, Ramya Rasipuram *et al.* developed an acoustic data driven G2P conversion method using KL-HMM and a well-trained multilayer perceptron (MLP) [64]. For Chinese language, Min-Siong Liang *et al.* developed a data-driven approach to phonetic transcription using text with speech augmentation, ASR with sausage searching net and pronunciation variation rules to convert Chinese text to Taiwanese. Data-driven rules are mainly used for developing pronunciation variation rules [65]. Christina Leitner developed data driven automatic phonetic transcripter for German language, by comparing with the samples in the database followed by appropriate concatenation [66].

Transcriptors are also developed for various Indian languages by employing different techniques. For Bangla, Joyanta Basu *et al.* designed G2P conversion algorithm based on orthographic information based rules. The part of speech (POS) and context information are included in the conversion to reduce exceptional cases. This conversion system contains verb, adjective and exception dictionary [67]. Shammur Absar Chowdhury *et al.* developed an experimental Grapheme-to-phoneme conversion method for bengla using conditional random field. This method adopts the G2P transcription by Mosaddeque *et al.* and and Alam *et al.* and additionally designed a sequence classification model using Conditional Random Fields (CRFs), which is a popular probabilistic graphical model widely used in Natural Language Processing (NLP) [68]. In their work, Ravi Kiran *et al.* developed two separate phonetic transcription system for Bangla and Odia using HMMs and MFCCs [69].

Grapheme to phone conversion for Hindi by C.S. Kumar *et al.* employed context sensitive rules for developing transcriptors. The context sensitive rules are designed to accurately predict the pronunciation using the contextual information. Exception cases are stored in exception lexicon. A decision tree approach is also implemented to generate the different pronunciation variation depending on the context [70]. Monojit Choudhary developed computational framework for rule-based grapheme-to-phoneme mapping for Hindi [71]. Sandeep Chaware *et al.* combined Hindi and Marathi for rule-based phonetic matching [72]. Grapheme-to-Phoneme conversion tool developed for Marathi by Sangramsing N Kayte *et al.* uses a language independent rule processing engine. The rule processing engine contains information in the form of lexicon, rules and mapping [73]. Grapheme to Phoneme Conversion for Punjabi Language by Ankita Goel *et al.* implemented transcription based on the rule set derived from the words in the dictionary [74]. In Urdu, Sarmad Hussain *et al.* converted the urdu letters to corresponding sound by applying letter-to-sound rules on a normalized text [75].

Neeshali R Nandarge *et al.* developed phonetic transcription system for Kannada language. In their method tokenization is used to produce the phonetic sound [76]. *A.G* Ramakrishnan *et al.* used rule based approach for G2P conversion in Tamil. The input normalized text is converted using the letter to sound rules for Tamil. These rules are based on pronunciation dictionary and a rule based approach for exceptional cases [77]. *N.* Udayakumar *et al.* followed decision tree learning technique for an automatic grapheme-to-phoneme conversion in Tamil. Based on the nature of the neighbouring phones the phonetic variations of a phoneme can be captured using Decision trees. Decision trees consist of nodes that contain the rules and leaves, which are labelled with target classes. When the object reaches a leaf, the class label of this leaf is used as the answer for the specific transcription. Rule-based system is used for bootstrapping the manually generated lexicon [78]. In Malayalam, Sumi S Nair *et al.* followed a rule-based method to convert grapheme form of a Malayalam word into phoneme form. G2P converter checks for suitable rules that matches the given grapheme [79]. Cini Kurian *et al.* build an automated transcription system for Malayalam by using HMMs with MFCC feature based on phonological rules for medium size vocabulary [80].

2.4 Review on Different Approaches Used in the Construction of Viseme Set

This section reviews different approaches used in literature for constructing the Viseme set in a language. Linguistic, perception based and data driven approaches are used for finding the Viseme set. Inlinguistic approach, viseme set is formed by exploiting the expert knowledge in the linguistics of the language. Phonetic properties, articulatory rules and visual intuition are used for classifying phonemes. In the perception based approach, experiments are conducted by human subjects to understand viseme classes. This method closely matches the way in which humans perceive visual speech. Data driven approach is used for automatically learning the natural division among phonemes in the parametric space. Visual features are extracted from the mouth region of talking faces and Viseme are formed by clustering in the feature space. Table 2.3 presents the summary of various works reported for the formation Viseme set for various languages.

Table 2.3: Summary of the Viseme set formation methods for different languages

Author – Year	Linguistic Information	Implementing Method	Special Features
Bozkrut– 2007	American English. 46 phonemes - 16 viseme classes.	Linguistics approach. HMM model was used for lip animation with audio speech as input. Contextual information was implemented using tri-phone model.	TIMIT speech database. Compared phone, tri-phone, viseme and tri-viseme based HMM structure for lip animation. Acoustic observation consists of 12 MFCC, energy, delta and acceleration coefficients resulting in 39 feature length. Applied for visual speech synthesis. [81]
Lander – 1999	35 phonemes - 12 classic Disney mouth position.	Linguistic approach	Facial Animation. [82]
Hazen – 2004	American English. 50 phonemes - 14 viseme classes. There are 54 phonemes but 4 phonemes were merged to get 50 phonemes.	Data-driven approach. Agglomerative hierarchical clustering algorithm. Bottom-up clustering using maximum Bhattacharyya distances. 96-dimension stacked PCA feature vectors.	AV-TIMIT speech database. Viseme was represented by three consecutives frames with middle frame describe the static viseme of each phoneme. Before clustering some phonemes were merged [45]
Lee – 2002	41 phonemes - 14 viseme classes. 7 vowel, 6 consonant and	Assumed to be linguistic approach. HMM modelling. Used context-independent recognition	TIMIT speech database. Two approaches in building viseme recognizer-viseme

	1 silence viseme.	units (phone model). Produced a sequence of viseme symbols from speech waveform.	HMMs and phoneme HMMs. [44]
Montgomery – 1983	American English. 15 vowels and diphthongs.	3 methods- Perceptual analysis using confusion matrix, Physical measurements (height, width, area, acoustical and visual duration) and Correlation between the two.	Study was restricted to vowels only. [84]
Neti– 2000	42 phonemes - 12 viseme classes (excluding silence).	Mixture of linguistic and data driven approach. Decision tree based HMM state clustering method. Models are trained using DCT visual features.	IBM ViaVoive database was used. [85]
Binnie– 1976	20 English consonants – 9 viseme classes.	Human testing. Confusion matrix.	Consonants are only studied. 20 English consonants were combined with the vowel /a/ to form 20 CV syllables. 34 female observers has participated in the testing process. Mapping is done subjectively. [86]
Fisher – 1968	23 initial consonants – 5 viseme classes. 20 final consonants – 5	Multiple-choice intelligibility test. Confusion matrix.	Studied visual perception of initial and final consonants. Mapping is done subjectively.

	viseme classes.		Fisher introduced the term viseme which is a compound word of visual and phoneme. [87]
Bear – 2015	46 phonemes - Visemes ranging from 2 to 45.	Viseme classes were obtained based upon the mapping of articulated phonemes, which was confused during phoneme recognition, into viseme groups.	Designed Speaker-dependent viseme classes. Studied on LiLIR dataset. 12 British speakers utter about 1000 words totally. [88]
Taylor – 2015	Created many-to-many mapping. Approximately 50000 visual speech gestures – 150 dynamic viseme classes.	Clustered the speech gestures identified by AAM (Active Appearance Model) of jaw and lips. 20 Dimension feature vector entirely describe the shape and appearance information. Dynamic visemes were learned entirely from visual data.	KB-2K database was used. A single actor recite 2542 phonetically balanced sentences from TIMIT database. Applied for automatic redubbing of video. [89]
Jeffers – 1980	American English. 43 phonemes -11 viseme classes.	Pure linguistic approach.	[90]
Setyati– 2015	Indonesian language. 49 phonemes – 12 viseme classes.	Linguistic approach.	Used Blend shape models for analysing the facial images. 10 speakers were used for this study. [91]
Mattheyses– 2013	Dutch language. Many-to-many phoneme-	Data driven approach. AAM (Active Appearance Model)-based	Coarticulation effect was studied.

	to- viseme mapping.	representation of mouth region. Tree- based and k-means clustering approach was used.	Applied for visual speech synthesis. [92]
Seko– 2013	Japanese language. 40 phonemes – 14 viseme classes (excluding silence).	HMM modelling.	CENSREC-1-AV database was used. [93]
Yu – 2010	50 words – 60 classes of visual speech units (VSU).	Data-driven approach. Used Expectation Maximization Principal Component Analysis (EM-PCA) as feature extraction method. Based on HMM classification.1	Introduced new term “Visual Speech Unit (VSU)” which include transition information between consecutive visemes. Two speakers utter a total of 50 words. [94]
Chitu– 2009	Dutch language. 40 phonemes – 18 viseme classes.	Confusion matrix.	[95]
Damien – 2009	Arabic language. 28 phonemes – 10 viseme classes.	Data-driven approach. Geometrical features used.	Four speakers utter four types of word sequences. [96]
Melenchón– 2007	Spanish language. 12 allophones – 6 viseme classes.	Data-driven approach. 12 PCA coefficients were used as feature vector.	Three speakers utter 12 Spanish sentences. [97]
Aschenberner– 2005	German language. 42 phonemes – 15 viseme classes.	Linguistic approach.	Applied for speech synthesis. [98]

In Indian languages, the concept of Viseme is mostly used with bimodal automatic speech recognition attempts [99-100]. S.Sandosh Kumar developed a viseme set in Malayalam mainly using visual observations [101].

2.5 Review on Facial Colour Analysis and Lip Tracking

Skin, lip, tongue and teeth are the different semantic regions around a talking mouth. There are a lot of investigations using different techniques and colour space to effectively segment these regions. RGB, normalised RGB, perceptual color spaces such as HSI, HSV, HSL, TSL, perceptually uniform colour spaces such as CIE Lab and CIE Luv, orthogonal colour spaces such as YCbCr,YIQ,YUV,YES. The range of algorithms employed includes simple thresholding, Bayees classification, Gaussian Mixture Models (GMM), Elliptical Boundary Models, self organising maps and neural networks. Table 2.4 summarises the approaches used for skin segmentation in different colour spaces.

Table 2.4: Details algorithms and techniques used for skin segmentation

No	Title of the Paper	Author	Colour Model	Technique used
1	Face-texture model based on SGLD and its application in face detection in a color scene	Dai and Nakano	YIQ	Thresholding [102]
2	Face detection in color images	R.L. Hsu, M. Abdel-Mottaleb, A.K. Jain	YCbCr	SGM [103]
3	Unsupervised and adaptive Gaussian skin-color model	L.M. Bergasa*, M. Mazo, A. Gardel, M.A. Sotelo, L. Boquete*	RGB	GMM [104]
4	A SOM based approach to skin detection with application in real time systems	D. Brown, I. Craw, J. Lewthwaite	RGB HSV XYZ TSL	SOM [105]
5	Performance evaluation of single and multiple-Gaussian models for skin-color Modelling	T.S. Caetano, S.D. Olabarriaga, D.A.C. Barone	RGB	SGM GMM [106]
6	Lafter: lips and face real time Tracke	N. Oliver, A. Pentland, F. Berard,	RGB	GMM [107]
7	Robust face tracking using color	K. Schwerdt, J.L. Crowely	RGB	The histogram-based Bayes classifier [108]
8	Skin detection, a Bayesian network approach	N. Sebe, T. Cohen, T.S. Huang, T. Gevers	RGB	BN [109]
9	Tracking regions	M. Störring, T. Koèka, H.J.	RGB	Thresholding

	of human skin through illumination changes	Anderson, E. Granum		[110]
10	Gaussian Mixture model for human skin color and its application in image and video databases	M.H. Yang, N. Ahuja	LUV	GMM(2) [111]
11	Detection of human faces in colour images,	C. Chen, S.P. Chiang	XYZ	Three layered feed forward NN [112]
12	Modeling facial colour and identity with Gaussian mixtures	S. McKenna, S. Gong, Y. Raja	HSV	GMM [113]
13	A novel approach for human face detection from color images under complex background	Y. Wang, B. Yuan	HSV	Thresholding [114]
14	A Bayesian approach to skin color classification in YCbCrcolor space	D. Chai, A. Bouzerdoum	YCbCr	Histogram+MLP [115]
15	Image chromatic adaptation using ANNs for skin color adaptation	P. Kakumanu	RGB	NN [116]
16	Statistical color models with application to skin detection	M.J. Jones, J.M. Rehg	RGB	Bayes GMM(16) [117]
17	Statistical models for skin Detection	B. Jedynak, H. Zheng, M. Daoudi	RGB	MaxEnt. model [118]
18	An elliptical boundary model for skin color Detection	J.Y. Lee, S.I. Yoo	XYZ	Elliptical boundary model [119]
19	Mixture clustering using multidimensional histograms for skin	Z. Fu, J. Yang, W. Hu, T. Tan	HSV	GMM(14)+Histogram merging[120]

	detection			
20	A probabilistic neural network for human face identification based on fuzzy logic chromatic rules	I. Anagnostopoulos, C. Anagnostopoulos, V. Loumos, E. Kayafas	RGB	Fuzzy rules+probabilistic neural network (PNN) [121]
21	Neural network-based skin color model for face detection	M.J. Seow	RGB	NN [122]
22	Mixture model for face-colormodeling and segmentation	H. Greenspan J. Goldberger	RGB	GMM(2) [123]
23	Skin-color extraction in images with complex background and varying illumination	Q.H. Thu, M. Meguro, M. Kaneko	HSV	GMM(4)+Multithresholding [124]
24	Improving face verification using skin color information	S. Marcel, and S. Bengio	rgb	Histogram+MLP [125]
25	Adaptive skin-color filter, Pattern Recognition	K.M. Cho, J.H. Jang, K.S. Hong	HSV	Thresholding [126]
26	Adaptive skin colormodeling using the skin locus for selecting training pixels	M. Soriano, J.B. MartinKauppi, S. Huovinen, M. Lääksonen	RGB	Skin locus thresholding [127]
27	Skin color-based video segmentation under time-varying illumination	L. Sigal, S. Sclaroff, V. Athitsos	HSV	Bayes [128]

Tongue segmentation is addressed recently in some works especially for applications related to traditional Chinese medicine [129-134]. Jian-qiang Du *et al.* uses the hue and intensity component based thresholding scheme for tongue segmentation [135].

Lip segmentation and tracking methods are classified as colour based, edge based, model based and hybrid approaches. Eveno *et al.* performed the most exhaustive survey for finding the appropriate colour space for lip – skin segmentation [136]. It compares 7 colour spaces and 12 additional special transforms and prepared a ranking of colour channels according to the suitability for segmentation. Axel *et al.* suggests a method for lip segmentation based on rgb-colour histogram [137]. Threshold-based segmentation strategies give a simple and effective approach to implement lip segmentation. Gritzman, Ashley D *et al.* propose a method called adaptive threshold optimisation (ATO) which selects the threshold by using feedback of shape information [138]. Statistical-colour model approach is used to automatically find and track the face and other facial features in the image [139- 140]. Shu-Hung *et al.* uses fuzzy clustering in the CIELab and CIELuv colour space clubbed with distance information for lip segmentation [141]. Similar approach by Simon Lucey *et al.* and Liew *et al.* uses connectivity information with fuzzy clustering [142]-[143]. Aashley *et al.* consolidate and compares the use of various colour models for lip segmentation purpose [144]. In the work of Menget *et al.* a 3 stage process is executed. In the first phase, CIE Lab and LUX colour space is used followed by a second phase using Gaussian model with hue and saturation values. The last phase uses a morphological filter for lip region extraction [145]. A markov random field based approach for lip segmentation using the same colour space is also reported [146]. Vladimir *et al.* uses a colour based lip tracking component in a general facial feature tracking system [147]. M. Uleeses *et al.* developed a statistical chromaticity model for lip segmentation and used it in personal identification system [148]. The low contrast between lip and skin for most ethnicities is the main concern in edge based lip tracking systems. To overcome this, edge is tracked in transformed space such as discrete Hartley transform (DHT) [149].

Nicolas *et al.* devised a jumping snake concept for developing a semi automatic lip tracking framework [150]. Localised colour active contours based approach implemented by Xin Liu *et al. is* found to be robust even frames with teeth and tongue [151]. Active contour or snake method in classical form with gradient vector flow is used for lip tracking in the work by Ghanshyam *et al.* [152]. In some other works variations of contour model is compared and the best is used for feature extraction in visual speech recognition applications[153][154]. Lip fitting for multimedia application with snakes using an adaptive color model is implemented by paul *et al.* The system is capable of learning and adapting from examples [155]. In their work, Hadi *et al.* uses different energy functions for outer and inner lip contours. Outer lip contours uses balloon functions and canny edge detectors, while inner lip employs image gradient and balloon energy [156]. The AAM based lip contour land mark localisation scheme implemented by Piotr *et al.* uses combined AAM, where a single model is used both for shape and texture in the implementation [157].

ASM has been employed in many works for segmentation and tracking [158-159]. Tanveer *et al.* used ASM based lip tracking for geometrical visual feature extraction [160]. Simon *et al.* used a linear regression based pre processing followed by ASM for lip tracking [161]. In the multiphase approach employed for lip reading by LEI XIE et.al colour ASM, 2- D mesh corner search and global search using M-estimators are combined [162]. *K.L.* Sum *et al.* used ASM for outer lip contour extraction in this work. 14 land mark points are used outer lip shape feature points and cost function is estimated using fuzzy clustering analysis [163]. In the work by juergen *et al.* two separate ASM models are used for inner and outer lip [164]. In their work, quoc *et al.* improves the performance of ASM by using multiple features for representing landmark neighbourhood. Both normal profiles and gray scale patches are used as texture feature for search. The approach has also incorporated the temporal information in to the model [165].

Recent attempts in lip tracking are mostly hybrid in their approach. Work by Brice *et al.* for analysis synthesis based lip animation framework starts with a low level segmentation in the LUX space and uses active contour estimation for inner and outer lip contour segmentation. Further the 3–D regularisation and repositioning is applied in the feedback loop for correction [166-167]. Shi-Lin Wang *et al.* uses shape guided fuzzy c-means clustering for lip tracking in challenging environment such as the presence of facial hair [168].

2.6 Review on Visual Speech Synthesis

Visual speech synthesis is an active research area with many potential applications. This section reviews the important works reported in the literature for synthesising visual speech in various languages. Eric *et al.* uses a unit concatenation framework for visual speech synthesis. Variable length video segments, which are optimally selected using Viterbi algorithm is used for concatenation. Factors such as co-articulation and temporal coherence are taken in to consideration while designing the system[169]. In the work, Gwenn *et al.* proposed a probabilistic model for generating lip movements corsponding to an audio. AAM is used for extracting features from video and HMM is used for aligning phonemes to speech signal using MFCC as a feature [170]. Oxana *et al.* uses a dendogram based method for clustering visemes. For synthesising facial movements a two phase approach is used. Trajectory planning in the parametric space is done using HMM, while the execution is done using conacatenation method [171]. In their unique approach Michel *et al.* analyses the relative influence of two sides of face in expressing emotions. They concluded that while happy emotion is symmetric, the sad emotion has an asymmetric nature [172]. Jie Yang *et al.* implemented an image based visual speech synthesis, in which visemes are the basic unit. The synthesis using visemes formed from automated data driven approaches is found to give minimum error, compared to synthesis using manual viseme set based method. The developed synthesis system is used as a communication agent in a bilingual translation framework [173]. Darren *et al.* developed a perceptual evaluation framework for using

"McGurk Effect" for talking head applications. Two psychology undergraduate students are used as human subjects in the work [174].

In their work, F. Elisei *et al.* aplied linear component analysis on an audio visual speech corpora to model and synthesis lip movements using a 3–D data driven approach. They found that six parameters are adequate to represent realistic lip movements. Coloured beads are used to mark land mark points in the face of human subjects [175]. Jonas Beskow*,* in his work used an opto-electronic motion tracking system to locate landmark points. The audio visual speech corpora consist of sentences uttered in different emotional states, during synthesis, the time aligned phonetic transcription is converted in to a trajectory in the control parameter space [176]. Olov Engwall describes visual speech synthesis system with an evaluation framework. The corpora consist of 460 phonetically balanced sentences, uttered by 40 speakers. The articulatory transitions between phonemes are labelled to be used for the concatenative visual speech synthesis system [177]. E.dey *et al.* uses a modified version of rhyme test for subjective evaluation of developed speech animation system, which is used in a pronunciation tutoring system [178]. Zhigang *et al.* used a concatenative approach for developing visual speech synthesis system. The video corpus is made from the utterances of a female actor. The corpora consist of 225 phoneme balanced sentence utterances with markers in the face. In this model users can set goals and constraings to produce emotionally varying visual speech [179]. In their work, Jose *et al.* employed a photogrammetric technique for visual speech synthesis. The main advantage of this work is that it considers both anticipatory and preservatory co-articualtion. In the work k-means algorithm is used for identifying articulatory patterns [180]. Jonas *et al.* describes ‘Furhat’ a robotic head, which uses facial animation frame work for human interaction. The four step process of using an animated face model, 3D mask printing, and distribution of colour uniformly and rigging with a projector is explained [181]. Eric *et al.* developed visual speech synthesis frameworks both using sample based and model based methods. Prosody analyser, audio renderer, visual prosody generator, co-articaulation engine and face rendering unit are the components of the frame work [182]. Ingmar *et al.* developed an animated tongue model using MRI data [183]. In their work Irene *et al.* developed a real time face animator from input text with options for representing non verbal cues. Linguistic information such as accent is used to improve the natural ness of visual speech. Wrinkles is introduced using vertex programme and register combiners [184]. Eric *et al.* reports the components used for developing a photorealistic talking head application from stored images. The facial parts are segmented and stored as bit maps for synthesis [185]. Michal *et al.* uses a production pipeline consist of analysis, expression classification, visual speech recognition and finally synthesis modules. The synthesiser uses a muscle model for generating animations for an MPEG-4 framework player [186]. In their work Richarod *et al.* uses Principal Component Analysis (PCA) for representing facial motion [187]. Darren *et al.* proposed a parametrically controlled deformable polynomial surface, for developing facial animation frameworks. Land mark points in face are captured using optical motion tracking system [174].

Tony *et al*. uses a generative model known as Multidimensional Morphable Model (MMM) to synthesis unknown face appearances from a training corpus. The recorded corpus consists of 30,000 images corresponding to 1- syllable and 2 – syllable words. The dimension of parametric vectors is 46 and the trajectory is obtained by treating it as a representation problem [188]. In concatenative synthesis appropriate segment selection in real time is the main challenge. Fu Jie *et al.* proposes a triphone based best unit selection method. The target cost for unit selection considers factors such as phonetic distance and co articulation [189]. 'SYNFACE' is a video telephony application for hearing impaired people. The main components of the application include phoneme recogniser and a 3-D synthetic face. An articulatory control model is used for face synthesis [190]. In their study, Ashish *et al.* established that a Viseme based acoustic model is sufficient for visual speech synthesis systems. The Viseme based model has the additional advantage of comparatively smaller training set and lower computational power demands [191]. The work by Salil *et al.* generates visual speech using a non parametric shifting state space model. The model is based on Gaussian processes. The system successfully models both preservatory and articulatorycoarticulation [192].

Thanveer *et al.* describes a multilingual visual speech synthesis framework for Indian languages. The speech recogniser in one language maps the incoming audio to a sequence of phonemes. This sequence is converted in to the viseme set of target language. Facial animations are generated for Hindi and Telungu languages using the proposed approach [193]. In the work proposed by Abdulrafay *et al.* lip is represented in a parametric space and this representation is used for synthesis in a animation framework developed using openGL [194]. V.AnandaNatarajan *et al.* describes the visual speech synthesis framework developed for Tamil. It is basically a mapping system from feature points in a 2- D images to 3-D images. The feature or landmark points in an image is automatically detected using a feature tracking algorithm. The computed feature points are mapped to a 3-D mesh of face by employing a 3 step process [195].

2.7 Conclusion

The thesis develops an integrated framework for Malayalam visual speech synthesis. Various components of the framework is reviewed in this chapter. A detailed discussion has been carried out about the durational models developed for speech synthesis applications. The detailed review of research works about transcriptors in various languages provided the insight for developing grapheme to allophone transcriptor in Malayalam. A summary of the strategies employed for viseme set formation in various languages is also presented. A review on the research works in segmentation of skin, lip and teeth are conducted with a special emphasis of lip segmentation and tracking. Finally various techniques for visual speech synthesis is reviewed for identifying relevant approaches in Malayalam visual speech synthesis.

CHAPTER 3

Allophone based Durational Analysis of Malayalam Vowels and Consonants for Visual Speech Synthesis

3.1 Introduction

In the widely used 35 languages of the world one third is from India. Malayalam, a south Indian language spoken by around 35 million people, is a classical Indian language and it is the official language of Kerala [196]. Malayalam, a low resource language, needs extensive studies to develop automatic language processing tools addressing its inherent peculiarities. This research work is an attempt towards this direction in the domain of visual speech analysis and synthesis. Speech being the most natural method of communication, interacting with a machine through speech is the most explored area in developing man – machine interfaces. Automatic speech recognition and synthesis are the main speech processing tasks. The study in one complements the developments in the other area. Speech is inherently bimodal, as a talking human being produces an audio signal and an image sequence. The facial movements while talking contributes both to the linguistic perception of the speech and to the perception of non-linguistic cues such as emotional state of the speaker. The accompanying image sequence is more important in noisy environment or where there is a hearing impairment [197]. This work performs visual speech analysis in various domains for developing an effective visual speech synthesis system in Malayalam with special emphasis on facial animation systems. Facial animation systems have reached the realm of reality by the effective convergence of computer vision and computer graphic domains [198].

Effectiveness of speech synthesis systems greatly depends on results from language specific explorations on many features. Duration and its modelling is an important cue affecting the intelligibility and naturality of synthesised audio visual speech. This chapter investigates the phoneme and allophone durational patterns of Malayalam. Phoneme, being the basic speech unit, its durational analysis is important both for developing successful speech recognition and synthesis systems. The durational information of the input text sequence can be shared by both audio and visual components of an audio visual speech synthesiser. Durational model capturing the dynamics of continuous speech is crucial in all three methods of speech synthesis including target based synthesis, model based synthesis and concatenative synthesis to generate intelligible and natural sound with corresponding image sequence [199]. In a text to speech synthesis system durational model predicts the duration sequence corresponding to the speech unit sequence and those values are predicted based on many factors effecting duration. The phoneme durational pattern exhibits wide variations across languages and hence language specific models need to be developed.

Phonetic identity of current segment, surrounding segments, positioning within the word and within the sentence, emotional state of the speaker *etc.* are factors affecting the duration of a segment [200]. It is possible to capture most of positional and contextual variability from the textual representation of the sentence. A phoneme can appear in the start, middle and end of a word creating positional variation. The change in duration due to the effect of surrounding speech units is called contextual variability. The effect due to neighbouring phonemes and position are generally known as co-articulation effects. Co-articulation effects in Malayalam are modelled by Malayalam linguists as allophonic characterisations for each phoneme. A well-defined allophone formation rule set exists for Malayalam [201]. So a durational model accommodating co-articulation effects due to positional and contextual variability of phonemes can be developed by understanding the durational pattern of allophones. Such rule based approach for phoneme duration modelling has been reported for many languages. Rule based segmental duration approaches have started with the work of Klatt for English [202] which is the basis for many Text to Speech systems. Rule based duration models has been established for many languages including French, Brazilian and Portuguese [203-204]. A combined rule based and statistics based prosody modelling is also applied for concatenative speech synthesis in Tamil and Hindi [205].

This chapter consolidates the findings of extensive investigation performed on the duration of allophonic variations of Malayalam Phonemes, which will be used as the durational model of a visual speech synthesis system in Malayalam to be discussed in chapter 7. The rest of the chapter is arranged as follows. Section 3. 2 describes the vowel and consonant phoneme sets of Malayalam. Section 3. 3

analyses the allophonic variations of Malayalam phonemes with the help of linguistic rule set. Section 3. 4 explains the creation of various audio visual speech corpora used in this work. Section 3. 5explains the phoneme and allophone duration pattern emerged based on the statistical analysis of speech corpora. Section 3. 6 discusses the modifications required for the durational model in the specific context of visual speech synthesis and section 3. 7 concludes the chapter.

3.2 Malayalam Phoneme Set

The term *phone* refers to the instances of phonemes, the smallest distinctive sound units in a language, in actual utterances. As they vary from one language to another, International Phonetic Alphabet (IPA) identified and defined 150 phones among all languages. British English has over 44 phonemes. Malayalam has a 51 member phoneme set including 11 vowels, 2 diphthongs and 38 consonants [206].

3. 2. 1 Malayalam Vowel Phonemes

In a system of language, the vowels are produced by comparatively open configuration of the vocal tract, with vibration of the vocal cords but without audible friction [207]. The difference is made by changing the relative positioning of active articulators. Diphthongs are vowel glides or combination of vowels. Malayalam has 11 monophthongs and 2 diphthongs [208]. The list of Malayalam vowel monophthongs is shown in table 3. 1. They are classified as 4 front vowels (ഇ /i/, ഈ /i: /, എ /e/, ഏ /e : /), three central vowels (അ

/a/, ആ /a: /, ˇ /ə/) and four back vowels (ഉ /u/, ഊ /u: /, ഒ /o/, ഓ /o:/). The frequently used central vowel ˇ /ə/ can either be treated as a separate phoneme or an allophonic variation of ഉ /u/. In this study we have considered ˇ /ə/ as an allophone of ഉ/u/. The vowels ഐ /ai / and ഔ /au/ are considered as the diphthongal vowel phonemes of Malayalam.

Table 3. 1 Monophthongal vowel phonemes in Malayalam

		Front	Central	Back
High	Short	i ഇ		u ഉ
	Long	i: ഈ		u: ഊ
Mid	Short	e എ	ˇə	o ഒ
	Long	e: ഏ		o: ഓ
Low	Short		a അ	
	Long		a: ആ	

3. 2. 2 Malayalam Consonant Phonemes

Consonants are speech sounds that are articulated with the complete or partial closure of the vocal tract [1]. Consonants can be classified based on place of articulation, manner of articulation and voicing. According to manner of articulation Malayalam Consonants can be broadly categorized as *plosives* or *oral stops*, *nasals* where the air flows through nose, *fricatives* where the tip of the tongue approaches alveolar ridge without actually contacting it, *trills, lateral, approximant and glide*. Malayalam has 38 consonants in which 21 of them are plosives. Plosives are again classified based on voice and aspiration. The consonant set can also be classified as *bilabial, labiodentals, dental, alveolar, retroflex, palatal, velar and glottal* based on the relative positioning of articulators while producing the phone. As this work explores the dynamics of visible articulators the consonant classification based on the place of articulation is elaborated below. The Malayalam consonant classes formed based on the place of articulation is listed in table 3. 2.

a. Bilabial

The lower and upper lip touches each other, resulting in the closure of the mouth during the utterance of phonemes of this class. There are 5 Malayalam phonemes in this category.

b. Labiodentals

The active articulator lower lip touches the upper teeth. /v വ/ is the only labio dental in Malayalam.

c. Dental

The active articulator tongue touches the upper teeth with its tip for producing a dental phoneme. There are 5 Malayalam phonemes in this category.

d. Alveolar

The tip of the tongue either touches or approaches the ridge behind the teeth for producing these classes of phonemes. There are 6 Malayalam phonemes in this category.

e. Retroflex

The active articulator tongue curves to touch or approaches the area behind alveolar ridge to create retroflex sounds. There are

8 Malayalam phonemes in this category, forming the largest consonant group.

f. Palatal

Here the body of the tongue approaches or touches the palate of the roof of the mouth. The 7 Malayalam phonemes in this category also includes the glide /y യ/in the language

g. Velar

Here the back of the active articulator tongue touches the soft palate or velum. There are 5 Malayalam phonemes in this category.

h. Glottal

The identity of glottal as a consonant itself is controversial in the absence of point of articulation. Glottis is the primary articulation producing the glottal sound.

Table 3. 2: Malayalam consonant classes based on place of articulation

Class Label	Phonemes	Number Of Phonemes
Bilabial	/P /പ, /p^h/ഫ, /b/ബ, /b^h/ഭ, /m /മ	5
Labiodental	/v/വ	1
Dental	/t /ത, /t^h /ഥ/d^h/ധ, /d/ദ, /n̪/ന	5
Alveolar	/r̠ /റ്റ, /n /ന, /s /സ, /r/ര/ṛ /റ, /l /ല/ൽ	6
Retroflex	/ʈ/ട, / ʈ^h/ഠ, /ɖ/ഡ, /ɖ^h/ഢ, /ɳ /ണ, /ʂ/ഷ, /ɭ/ള/ൾ, /ʐ /ഴ	8
Palatal	/c/ച, /c^{h}ഛ, / ɟ/ജ, /ɟ^h/ഝ, /ɲ/ഞ,	7

	/ʃ/ശ, /y /യ	
Velar	/k/ക, / k^h/ഖ, /g/ഗ, /g^h/ഘ, /ŋ/ങ	5
Glottal	/h /ഹ	1

3. 3. Allophonic Variations in Malayalam

As phoneme is an abstract cognitive concept, numerous realisations of any phoneme can be seen in real speech. We understand the sequence of sound as a sequence of phonemes. Allophone is a phonetic realisation of a phone and allophonic variations of a phoneme are caused by position, context and other non-linguistic reasons. Factors such as contextual and positional variability can be detected from the text while some others such as dialect cannot be detected from the text [14]. In their studies, Asher and V. R. Prabodhachandran Nair described the rules of Malayalam allophone formation [6, 15]. They have proposed linguistic descriptions for defining the allophones of each Malayalam phones. The linguistic descriptions can be converted to position and neighborhood-based rule set. Thus certain rule set for Malayalam vowel and consonant allophones based on position and neighbouring information's are created. As part of the study, 107 allophones in Malayalam which include 76 consonant allophones, 28 vowel allophones and 3 allophones corresponding to diphthongs are identified. Section 3. 3. 1 describes the rule set used for the formation of Malayalam vowel allophones and section 3. 3. 2 describes the same used for the formation of consonant allophones.

3. 3. 1 Rule based Malayalam Vowel Allophone Formation

A rule set for the formation of vowel allophones in Malayalam based on the position and neighbouring information are formed. The vowel allophone characterisation depends mainly on the relative positioning of phonemes. The three allophones of vowel ഇ/i/, [^{y}i], [I] and [i^{y}] correspond to initial, middle and final positions respectively. But its long vowel counter part ഈ [i:] has just two allophones corresponding to initial and middle positions. The position and neighbourhood based rule set used for formation of Malayalam vowel allophones is listed in table 3. 3.

Table 3. 3: Position and neighbourhood based rule set used for the formation of Malayalam vowel allophones

Sl. No.	Phoneme	Allophone	Position	Rule
1	ഇ /i/	[i]	Middle	*Metadata*: Low high front unrounded short vocoid. Neighbourhood: Any
		[^{y}i]	Initial	*Metadata*: High front unrounded long tense vocoid with onglide. *Neighbourhood*: Any
		[y^{i}]	Final	*Metadata*: High front unrounded short tense vocoid with offglide. *Neighbourhood*: Any
2	ഈ /i: /	[^{y}i:]	Initial	*Metadata*: High front unrounded long tense vocoid with onglide. Neighbourhood: Any
		[i:]	Middle	*Metadata*: High front unrounded tense long vocoid; medially. *Neighbourhood*: Any
3	എ /e/	[^{y}e]	Initial	*Metadata*: Higher mid front unrounded short tense vocoid with onglide. *Neighbourhood*: Any

		[^{e}y]	Initial	*Metadata*: Higher mid front unrounded short tense vocoid with offglide[j]. *Neighbourhood*: Any
		[E]	Middle	*Metadata*: Mean mid front unrounded short vocoid. *Neighbourhood*: Any
4	ഏ /e: /	[^{y}e:]	Initial	*Metadata*: Higher mid front unrounded long tense vocoid with onglide. *Neighbourhood*: Any
		[e^{r}:]	Final	*Metadata*: Higher mid front unrounded long tense vocoid with offglide [j]. *Neighbourhood*: Any
		[e:]	Middle	*Metadata*: Higher mid front unrounded long tense vocoid. *Neighbourhood*: Any
5	അ /a/	[ʌ]	Inital& Final	*Metadata*: Low mid back vocoid in the initial syllable and word. *Neighbourhood*: Any
		[A]	Middle	*Metadata*: Low mid central vocoid in the medial syllable. *Neighbourhood*: Any
6	ആ /a:/	[a:]	-	*Metadata*: Low back long tense vocoid after velar consonants. *Neighbourhood*: Left : Velar Consonants
		[a:]	-	*Metadata*: Low central long vocoid after all non-velaric consonants. *Neighbourhood*: Left : Non-velar Consonants
7	ഉ /u/	[^{w}u]	Initial	*Metadata*: High back rounded tense short vocoid with onglide [w]. *Neighbourhood*: Any
		[u^{w}]	Final	*Metadata*: Higher back rounded tense short vocoid with offglide [w]. *Neighbourhood*: Any

		[ɯ]	Middle	*Metadata*: High back unrounded short vociod in the medial syllable. *Neighbourhood*: Any
		[ə]	Final	*Metadata*: Higher mid central unrounded open vocoid, between consonant and vowel in the word boundary with open juncture. *Neighbourhood*: Any Other: Open juncture.
		[ə*]	-	*Metadata*: In open juncture 'ə' is in free variation after the following some phonemes. *Neighbourhood*: [ɳ] [ɭ] [ṛ] [l]]
		[ɯv]	-	Metadata*:* Low high back unrounded vocoid after some phonemes. *Neighbourhood:* Right : Labial, dental, palatal, and velar plosives and labial, and dental nasals and non-retroflex fricatives and labio-dental continuants. *Other:* In Sanskrit words
		[U]	-	*Metadata*: Low high back rounded tense vocoid, after word initial consonant. *Neighbourhood*: Right : Consonant as the first letter in a word
8	ഊ /u: /	[^{w}u:]	Initial	*Metadata*; High back rounded long tense vocoid with onglide [w]. *Neighbourhood*: Any
		[u]	Other than Initial	*Metadata*: High back rounded tense vocoid, elsewhere. *Neighbourhood*: Any
9	ഒ /o/	[^{w}O]	Initial	*Metadata*: Higher mid back rounded tense short vocoid with onglide [w] in the initial position. *Neighbourhood*: Any
		[O]	Middle	*Metadata*: Mean mid back tense rounded short vocoid. *Neighbourhood*: Any

10	ഓ /o: /	[^{w}O:]	Initial	*Metadata*: Higher mid back rounded tense long vocoid with onglide. *Neighbourhood*: Any
		[O]	Medial and Final	*Metadata*: Higher mid back tense long vocoid. *Neighbourhood*: Any

The following section describes the formation of rule set for Malayalam consonant allophones.

3. 3. 2. Rule Based Malayalam Consonant Allophone Formation

The rule set characterising allophones of 38 Malayalam consonants is listed in table 3. 4. While the rule set of vowel allophone formation mainly focuses on the position of the phoneme segment, consonant allophones varies both with position and neighbourhood phonemes. For example the first allophone of പ [P] is characterised by position, but the rule of formation of second allophone demands vowels on either side. The third and fourth allophones are medial but differ in the neighbouring phonemes or the context in which it appears. Nineteen consonants in Malayalam have just one allophone, *i. e.* these consonants are unaffected by neighbouring phonemes or position of occurrence. /P /പ, /t /ത/, /c/ ച, /k/ ക, /ŋ/ ങ *etc.* are the consonants with the largest number of allophones.

Table 3. 4: Rule set for the formation of Malayalam consonant allophones

Sl. No.	Phoneme	Allophone	Rule
1	പ /P/	[p]	Metadata: Voiceless tense bilabial stop contoid. Iinitially.
		[β]	Metadata: Voiced bilabial approximant. Voiced bilabial stop contoid with more lax quality . Intervocalically
		[b]	Metadata: Slightly voiced bilabial stop . Slightly voiced bilabial stop contoid In medial nasal – plosive cluster.
		[P]	Metadata: Voiceless most tense bilabial stop contoid. Medially singly or in a cluster except when preceded by nasal.
2	ഫ /p^h/	[p^h]	Metadata: -Voiceless aspirated bilabial stop. Occurs initially and medially in Sanskrit loans
3	ബ /b/	[B]	Metadata: Voiced tense bilabial stop contoid. In consonant clusters.
		[b]	Metadata: Voiced lax bilabial stop contoid. . occurs initially and intervocally.
4	ഭ/b^h /	[b^h]	Metadata: Voiced aspirated labio-labial stop. Occurs initially and medially in Sanskrit loans.
5	മ /m/	[m̪^{h}]	Metadata: Voiceless bilabial nasal contoid . 1. Before velar fricative
		[M]	Metadata : More tense bilabial nasal contoid . 1 In consonant clusters when preceded by alveolar flap
		[m]	Metadata: Labio – dental nasal. 1 Before labio – dental continuant
		[m]	Metadata : labial nasal . , Elsewhere
6	വ /v/	[w]	Metadata: Voiced bilabial continuant. Preceded by consonants except flapped in consonant clusters.
		[v]	Metadata: Voiced labiodental continuant. 1 Initially 2 In the clusters in medial position, where [w] does not appear. 3. Mostly short and rarely long in intervocalic position.
7	ത /t/	[t]	Metadata: Voiceless lamino-dental stop. , Voiceless tense dental stop contoid, Occurs initially.

		$[t^{i}]$	Metadata: Voiceless most tense dental stop contoid . In clusters except when not preceded by nasals and [j], medially when geminated.
		[ð]	Metadata: Voiced lamino dental approximant possibly with slight friction. More voiced dental stop contoid with more lax quality. Intervocalically or preceded by [j]
		[d̪]	Metadata: voiced lamino-dental stop. Slightly voiced dental stop contoid with lax quality. Medially preceded by nasal.
8	ഥ $/t^{h}/$	$[t^{h}]$	Metadata: Voiceless aspirated dental stop. Voiceless aspirated lamino-dental stop . Medially in Sanskrit loans.
9	ദ /d/	[ḍ]	Metadata: voiced tense dental stop contoid . medially in consonant clusters
		[d]	Metadata: voiced dental stop contoid with lax quality. Initially, intervocalic, and in clusters.
10	ധ $/d^{h}$ /	$[d^{h}$]	Metadata: Voiced aspirated lamio dental stop. Occurs initially and medially in Sanskrit loans.
11	ന /n̪/	[n̪]	Metadata: More tense dental nasal contoid. 1. Preceded by alveolar flap.
		[n]	Metadata: Less tense dental nasal contoid. 1 Elsewhere, i. e. short initially and before the other dental consonants 2. Long in intervocalic position
12	റ/r̠/	[d]	Metadata: Voiced alveolar stop. Voiced lax alveolar stop contoid. 1 In a medial homorganic nasal stop [In such sequence the -n is a stem-final consonant and [r̠] is the first consonant of the genitive case suffix] 2. After a nasal
		[t]	Metadata: Voiceless tense alveolar stop contoid. Voiceless apico-alveolar stop with, for some speakers, a slight palatal quality and/ or a hint of affrication. 1 As identical consonant cluster in intervocalic position. . 2 Medially when it is long . .
13	ന/n/	$[n^{h}]$	Metadata: Voiceless alveolar nasal contoid. When preceded by velar fricative /h/

		[n]	Metadata: Voiced alveolar nasal contoid, elsewhere. . 1. Word final position 2. Short before other alveolar consonant 3 Short or long in intervocalic position
14	സ/s/	[s]	Metadata: voiceless apico alveolar fricative. Occurs initial medial and final position. In the case of final /s/ there is an alternative pronounciation with an added enunciativevowel'. , Voiceless denti – alveolar sibilant slit fricative . . Singly in initial position and in clusters, long intervocally and medially
15. a	ര/r/	[r]	Metadata: voiced apicodenti alveolar tap . 1. Word initially 2. The second consonant in some initial consonant sequence. 3. Intervocalically. 4. In a number of medial clusters. Voiced palatalized denti alveolar flap contoid . 1. Word initially 2. Intervocalically 3. before [j] 4. after [b] [d] or [g]
15. b	റ /ṛ/	[ṛ]	Metadata: - Voiced apico-alveolar tap or trill . 1. Word initially 2. Second consonant in initial consonant sequences 3. Intervocalically 4. In a number of medial sequences Voiced velorized alveolar flap. 1. Rarely in initial position 2. Intervocalically 3. Finally 4. Followed by consonants except [j] 5. After consonants except [b] [d] [g]
16	ല/ൽ /l/	[l]	Metadata: Voiced apico – alveolar lateral . 1 Word initially 2 The second element in some word initial clusters 3. Intervocalically 4 Medially as a geminate consonant. 5. Word-finally. Voiced frictionless alveolar lateral contoid . 1. Rarely in word initial position 2. Short or long intervocalically 3. Finally 4. In clusters
17	ട/ʈ/	[ɖ]	Metadata: Slightly voiced and laxed retroflex stop contoid. . Voiced sub laminopostalveolar (Retroflex) . After homorganic nasal.
		[ɽ]	Metadata: More voiced and lax retroflex plosive contoidintervocalically. Voiced sublamino post alveolar flap. .

		[ʈ]	Metadata: Voiceless tense retroflex plosive contoid. voicelesssublamino post alveolar (Retroflex). In word initial position in loan words.
		[T]	Metadata: Voiceless retroflex plosive contoid with more tense quality. . In consonant cluster; not preceded by nasal.
18	ഠ/ʈʰ/	[ʈʰ]	Metadata: Voiceless aspirated tense retroflex. . Intervocalically and preceded by nasal
		[Tʰ]	Metadata: Voiceless aspirated more tense retroflex. . Elsewhere; not after nasal in clusters.
19	ഡ/ɖ/	[ɖ]	Metadata: Voiced sublamino postal velar stop . Voiced retroflex stop. Occurs initially and medially. Both intervocalically and in the sequence / d/ lax when short and tense when long .
20	ഢ/ɖʰ /	[ɖʰ]	Metadata: Voiced aspirated sublamino-postalveolar stop. Medially in a small number of Sanskrit loans .
21	ണ/ɳ/	[ɳ]	Metadata: Voiced sub lamino – palatal (retroflex) nasal. Retroflex nasal contoid. Occurs medially ie, 1. Intervocalically 2. Medially as a geminate 3. In the following medial clusters ɳʈ, ɳɖ , ṛɳ, ɳj, ʂɳ. Short in word final position [tu: .]
22	ഷ /ʂ/	[ʂ]	Metadata: Voiceless apico alveolar fricative. Voiceless retroflex more long, tense sibilant groove fricative . In initial, medial and final positions in loan words.
23	ള/ൾ /ɭ/	[ɭ]	Metadata - Voiced sublamino palatal (retroflex) lateral . 1 The second element in some word initial clusters in loans 2. Intervocali ally 3. Medially as a geminate consonant. 4. FinallyVoiced frictionless retroflex lateral contoid . 1. Finally 2. In clusters 3 Short or long intervocalically. 4. Rarely in initial position.

24	ഴ/ʐ/	[ʐ]	Metadata : Voiced sublamino palatal approximant. 1 Intervocalically 2. First element in medial consonant clusters 3. Finally (in which, there is alternative with vocalic release) Voiced retroflex continuant. 1 Medially, intervocalically 2. In consonant clusters. VRP posits that the concerned Malayalam sound is without even a trace of friction and employs a new symbol [ỵ] instead of (ʐ).
25	ച/c/	[c]	-
		[ç̬]	Metadata : Voiceless aspirated tense velar plosive; initially .
		[ɟ]	Metadata: Voiced lamino – palato- alveolar stop, slightly affricated. Occurring medially after / ɲ /.
		[C]	Metadata: Voiced palatal affricate with maximum tense quality. In clusters; not preceded by nasals. .
26	ഛ/cʰ/	[cʰ]	Metadata: Voiceless aspirated tense palatal affricate, occurs initially .
		[Cʰ]	Metadata: In consonant clusters.
27	ജ/ɟ/	[J]	Metadata: Voiced tense palatal affricate in consonant clusters.
		[j]	Metadata: Voiced palatal lax affricate. Initially before a vowel, and intervocalically.
28	ഝ/ɟʰ /	[ɟʰ]	Metadata: Voiced aspirated lamino-palatoalveolar affricate. Found only in Sanskrit loans Initially and medially
29	ഞ/ɲ /	[ɲ]	Metadata: Voiced lamino – palatal nasal. 1 Word initially 2. In the sequence -ɲj - in Sanskrit loans, where there is alternative of a long palatal nasal. 3. In the initial and medial sequence - ɟɲ 4. In the sequence - ɲc - in a small number of native words forms. 5. as a geminate consonant in native words. Palatal nasal contoid. 1 Word initially (short) 2. Short when followed by other palatal consonant 3. Long in intervocalic position.

30	ശ/ʃ /	[ʃ]	Metadata: Voiceless lamino palatal alveolar (retroflex) . Voiceless palatal sibilant slit fricative. Occurs initially and medially in loans. In initial position followed by a vowel in clusters, intervocally. [long/short].
31	യ/y /	[y]	Metadata: - Voiced close front dorso – palatal semivowel . 1. occurs word initially: 2. Intervocalically 3. Medially as a geminated consonant 4. Finally as a variant of / j / 5. In word initial and medial clusters. Voiced palatal continuant . occursshort in initial and final position 2. short and long in medical position, in consonant clusters, and intervocalically position.
32	ക/k/	[k]	Metadata: Voiceless dorso velar plosive in initial position, and medially when doubled. Voiceless tense velar contoid; in initial position.
		[kj]	Metadata : Voicelesspalataliseddorso velar plosive in the environment of preceding front vowel. Voiceless, most tense palatalised velar plosive contoid in the environment of preceding front vowel/in the sandhi environment where [j] precedes.
		[ɣ]	Metadata: Voiced dorso velar approximant in intervocalic position. Variant realistions in this environment are [h] and [ɦ]. Not palatalised, not following high front vowel.
		[ɡ]	Metadata : Voiceddorso velar stop when preceeded by a nasal. Not palatalised, velar stop contoid with a little voiced and lax quality in the environment preceded by a nasal.
		[ʈ]	Metadata : The sequence [kʂ] [ക്ഷ] is pronounced with retroflexion.
		[K]	Metadata : Voiced velar contoid with most tense quality, in clusters except the contoid occurs after [j] or a nasal.
33	ഖ/k^h/	[k^h]	Metadata: Voiceless aspirated tense velar plosive contoid. Intervocalicallyand when preceded by a nasal.
		[K^h]	Metadata: Voiceless aspirated tense velar plosive. Initially.
		[K^h]	Metadata: Voiceless aspirated, more tense velar plosive. Elsewhere .

34	ഗ/g /	[G]	-
		[g]	Metadata : Voiced and lax elsewhere ie, intervocalically and initially .
35	ഘ/g^h/	[g^h]	Metadata: In borrowing from Sanskrit, occurs in initial and medial position. VRP opines that it is represented in orthography only, not realized in speech. However, he considers it as an allophone of /K^h /]
36	ങ/ŋ/	[ŋ]	Metadata: Voiced dorso – velar nasal. 1. In the sequence - ŋg - in Sanskrit loans. 2. In the Sequence - ŋk - bridging a morpheme juncture in a small number of word forms in the native lexicon. 3. As a geminated consonant in native words. 4. In the sequence - ŋk - in English loans.
		[ŋj]	Metadata: Voiced dorso – palato velar nasal. 1. In native lexicon, where geminate [ŋ] follows a front vowel.
		[ŋ$^<$]	Metadata: Pre velaric nasal contoid with clear palatalization and tense quality. 1 In the names of fruits and plant except the one which occurs after a long low vowel.
		[ŋ$^>$]	Metadata: Post velaric nasal contoid. 1 Long and tense when after a vowel in low-back region 2 Short before homorganic plosive.
		[ŋ']	Metadata : Tense mid - velaric nasal contoid . Else where
37	ഹ /h/	[H]	Metadata : Voiceless extremely short velar fricative . – Finally.
		[h]	Metadata: Voiceless velaric or glottal fricative. 1. Initially 2. After vowel 3. In clusters 4. Intervocalically

A rigid rule based allophone characterisation framework is one of the peculiarities of Malayalam language. This work has utilised the rule set for developing the basic components of an allophone based speech synthesis system. The implementation of a grapheme to allophone transcription system using this rule set is described in chapter 4.

3.4 Creation of Audiovisual Malayalam Speech Data Set

A phoneme based audiovisual speech corpora considering the allophonic variability is the primary requirement for carrying out research and development in audiovisual speech recognition and synthesis. The database creation is executed as a two phase process. In phase 1 the phoneme set, allophone set and inventory of words in each allophonic category with male and female utterances are developed. This work has been carried out as part of the Malayalam phonetic archive project owned by Thunchath Ezhuthachan Malayalam University (TEMU), Kerala, India. In the second phase, an audio visual data set of isolated phoneme utterances and continuous word utterances is created. The audio visual speech corpora is based on the allophone based specifications and listings defined by TEMU project.

3. 4. 1 TEMU Malayalam Phonetic Archive

The TEMU data set is a comprehensive corpora created based on a carefully compiled inventory of phones which are currently employed in the Malayalam language. The author of this work is also involved directly in the project by providing proper directions for content listing and technical support for digitization and web publishing. Malayalam phoneme segments are recorded in its standardized orthography followed by a number of examples of its occurrence in phonologically relevant different positions. Allophones are listed together and pronunciation of each example recorded from the natural speech is demonstrated in both male and female voices. The data comprises of 11 vowels, 2 diphthongs and 38 consonants, and its allophonic variation with 900 spoken words as examples. This archive is presently available in public domain under creative commons license. The dataset is archived and published in web portal [15]. The following section describes the process of developing a comprehensive audio visual speech dataset in detail.

3. 4. 2 Comprehensive Malayalam Audio Visual Speech Dataset

Audio visual speech corpora of 10 different speakers based on the specifications and listings defined by TEMU repository is developed for the research purpose. The Malayalam Audio Visual Speech Corpus – Isolated Phoneme (MAVSC-IP) consisting of data procured from 10 female speakers is developed. This dataset is further used for the purpose of mouth region colour analysis, lip tracking and lip motion synthesis. Each speaker uttered 51 Malayalam phonemes of Malayalam in a silence-phoneme-silence mode. A Malayalam Audio Visual Speech Corpus – Isolated Word (MAVSC-IW), consist of 214 word utterances procured from 10 female speakers is also developed. Each speaker uttered 214 words (consist of 2 words from each allophone category) accommodating the allophonic variability in Malayalam. The list of words is spoken by reading continuously without break to bring the co articulation effect of continuous speech. The same set of 10 speakers is involved in the creation of both MAVSC-IP and MAVSC-IW corpora. Five frames with maximum phoneme or allophone visual presents is manually selected for each utterance. In all these data set creations, the data procurement is restricted to female speakers only, so as to avoid the complexity that may arise due to facial hairs in further processing. All the female speakers belong to the age group of 20 to 45 years and cover the skin tone variability in Indian sub-continent. The videos are captured using a HDR-CX405 Sony Handycam having frame rate of 25fps with a resolution of 1280 x 720 in MP4 format and audio sampled at 44100Hz. The recording is done in an ordinary office room with normal lighting condition. All the spoken video samples are taken in the frontal face talking mode. MAVSC-IP data set consists of around 2500 frames and MAVSC-IW consists of around 10, 500 frames.

The audio only data set provided by TEMU is used for estimating the average phoneme and allophone based duration in Malayalam. The durational model formed from the average behaviour is adjusted based on the audio visual asynchrony analysis performed on the audiovisual data samples taken from MAVSC-IP and MAVSC-IW corpora. The following section describes the durational properties of the Malayalam vowel and consonant allophones derived on the basis of the detailed analysis conducted on the TEMU dataset.

3.5 Durational Analysis of Malayalam Vowel and Consonant Allophones

Duration is one of the most important cues deciding the intelligibility and naturality of synthesised speech. Malayalam phonemes exhibit wide variability in duration during continuous speech. The duration of a phoneme varies depending on the context and position it appears. Allophone characterisation in Malayalam captures the positional and contextual variability of Malayalam Phonemes. So an allophone based durational pattern analysis can model the contextual and positional variability in duration. Section 3. 5. 1 analyses the durational variation of vowel allophones and section 3. 5. 2 analyse the durational variations in consonant allophone of Malayalam.

3. 5. 1 Durational Analysis of Malayalam Vowel Allophones

This section consolidates the detailed investigation performed on the duration of Malayalam vowel speech segments. The word samples uttered by both male and female are taken from the TEMU data sets are used for the durational analysis. An allophone centric durational analysis is performed, instead of the conventional phoneme based durational model. The duration of isolated phoneme utterances and the duration of allophone in sample word utterances are computed. The mean duration of each allophone is reported. The mean duration of allophones each phonemes is compared with isolated phoneme utterances. Table 3.5 consolidates the durational statistics of Malayalam vowel allophones separately for male and female speakers.

Table 3. 5: Durational statistics of Malayalam vowel allophones

Sl. No	Vowel	Allophone	Average Duration (in sec)	
			Male	Female
1	ഇ /i/	[i]	0. 08871	0. 12866
		[^{y}i]	0. 11778	0. 14847
		[y^{i}]	0. 09356	0. 10311
2	ഈ /i: /	[^{y}i:]	0. 20628	0. 22192
		[i:]	0. 20229	0. 22763
3	എ /e/	[^{y}e]	0. 12792	0. 14015
		[E]	0. 08989	0. 10110
4	ഏ /e: /	[^{y}e:]	0. 24340	0. 25133
		[e^{r}:]	0. 20149	0. 11787
		[e:]	0. 20317	0. 24064
5	അ /a/	[ʌ]	0. 11452	0. 15527
		[A]	0. 08414	0. 08588
6	ആ /a: /	[a:]	0. 22137	0. 26522
		[a]	0. 21536	0. 27507
	ഉ /u/	[^{w}u]	0. 09610	0. 10346
		[u^{w}]	0. 10917	0. 08431

7		[ɯ]	0. 07510	0. 07979
		[ə]	0. 15100	0. 08847
		[ə*]	0. 13990	0. 08253
		[ɯ^{v}]	0. 06245	0. 07106
		[U]	0. 08139	0. 07943
8	ഊ /u: /	[^{w}u:]	0. 19784	0. 24334
		[u]	0. 19789	0. 21948
9	ഒ /o/	[^{w}O]	0. 10789	0. 11639
		[O]	0. 09402	0. 13223
10	ഓ /o: /	[^{w}O:]	0. 24083	0. 26733
		[O]	0. 20402	0. 23338

The average of Malayalam vowel allophone durations is 142. 84ms for males and 155. 63 ms for females. The range of vowel allophone duration is from 41. 55 ms to 289. 36 ms for male and 43. 16 ms to 330. 0 ms for female. The difference between phoneme duration in isolated phoneme utterances and actual duration during continuous speech in different context (manifested as allophones) is an important cue for speech synthesis application. A graph depicting the difference is shown in figure 3. 1.

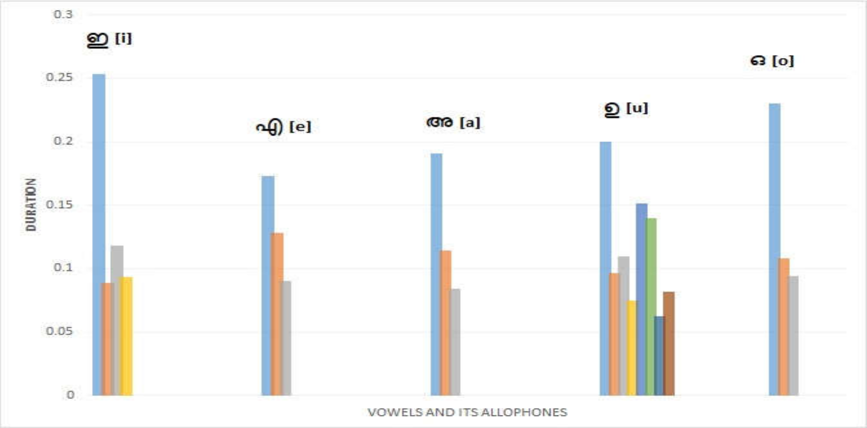


Figure 3. 1 Comparison of duration of isolated vowel phonemes and vowel allophones

The following section describes the duration alanalysis performed on Malayalam consonant allophones.

3.5.2 Duration Analysis of Malayalam Consonant Allophones

This section consolidates the detailed investigation performed on the duration of Malayalam consonant speech segments. The speech corpora and the methodology of analysis is the same as employed for duration analysis of vowel allophones. Table 3.6 consolidates the durational statistics of Malayalam consonant allophones.

Table 3.6: Durational statistics of Malayalam consonant allophones

Sl. No.	Consonants	Allophone	Average Duration in Seconds	
			Male	Female
1	P പ	[p]	0. 01340	0. 03024
		[β]	0. 02416	0. 02316
		[b]	0. 03509	0. 03309
		[P]	0. 03216	0. 03016
2	p^h ഫ	[p^h]	0. 05521	0. 05321
3	b ബ	[B]	0. 08418	0. 08218
		[b]	0. 07514	0. 07714
4	b^h ഭ	[b^h]	0. 07674	0. 09839
5	m മ	[m̪^{h}]	0. 07099	0. 03903
		[M]	0. 09796	0. 06659
		[m]	0. 09531	0. 09075
		[m]	0. 08339	0. 08211
6	v വ	[w]	0. 03924	0. 03214
		[v]	0. 08288	0. 07236
7	t ത	[t]	0. 01341	0. 02404
		[t$^{'}$']	0. 03007	0. 03109
		[ð]	0. 02965	0. 02846
		[d̪]	0. 01431	0. 02676
8	t^h ഥ	[t^h]	0. 06104	0. 06526

9	d ദ	[ḍ]	0. 04851	0. 02513
		[d]	0. 02243	0. 02335
10	dʰ ധ	[dʰ]	0. 04455	0. 03448
11	n̪ ന	[n̪]	0. 08948	0. 07447
		N	0. 10796	0. 15118
12	r̠ റ്റ	[d]	0. 02843	0. 02232
		[t]	0. 01488	0. 01821
13	n ന	[n]	0. 16886	0. 15322
		[n]	0. 12224	0. 14245
14	s സ	[s]	0. 11396	0. 13874
15	r ര ṛ ൃ	[r]	0. 10939	0. 12834
16	l ല/ൽ	[l]	0. 14006	0. 12209
17	ʈ S	[ɖ]	0. 02822	0. 01282
		[ɽ]	0. 01426	0. 01430
		[ʈ]	0. 10374	0. 10111
		[T]	0. 02006	0. 01256
18	ʈʰ O	[ʈ]	0. 02513	0. 03250
		[T]	0. 03658	0. 03232
19	ɖ ഡ	[d]	0. 01608	0. 02910
20	ɖʰ ഢ	[dʰ]	0. 04557	0. 05465
21	ɳ ണ	[ɳ]	0. 08460	0. 09675
22	ʂ ഷ	[sʰ]	0. 14175	0. 15000
23	ɭ ള/ൾ	[ɭ]	0. 06460	0. 04564
24	ʐ ഴ	[ʐ]	0. 08493	0. 09225
24	c ച	[c]	0. 04773	0. 05961
		[c̬]	0. 03856	0. 04505
		[ɟ]	0. 04609	0. 05339
		[C]	0. 08531	0. 06377
25	cʰ ഛ	[cʰ]	0. 09809	0. 08155
		[Cʰ]	0. 09645	0. 08601
26	ɟ ജ	[J]	0. 06379	0. 07052
		[j]	0. 05652	0. 05273
27	ɟʰ ഝ	[jʰ]	0. 08089	0. 09455

28	ɲ ഞ	[ɲ]	0. 09071	0. 09066
29	ʃ ശ	[ʃ]	0. 11473	0. 16944
30	y യ	[y]	0. 06059	0. 09967
31	k ക	[k]	0. 02758	0. 02120
		[kj]	0. 04233	0. 02900
		[ɣ̥]	0. 05296	0. 04231
		[ɠ]	0. 03121	0. 02515
		[ʈ]	0. 02236	0. 02452
		[K]	0. 02247	0. 02217
32	k^h ഖ	[k^h]	0. 05651	0. 05491
		[K^h]	0. 06391	0. 05405
		[K^h]	0. 07139	0. 05729
33	g ഗ	[G]	0. 02534	0. 05021
		[g]	0. 05686	0. 08251
34	g^h ഘ	[g^h]	0. 07507	0. 07607
35	ŋ ങ	[ŋ]	0. 13133	0. 13329
		[ŋj]	0. 10551	0. 11969
		[ŋ$^<$]	0. 09054	0. 15796
		[ŋ>]	0. 14165	0. 15324
		[ŋ']	0. 15364	0. 18484
36	h ഹ	[H]	0. 0857	0. 0854
		[h]	0. 08746	0. 08921

The range of consanant duration varies from 13. 3 ms to 168. 4 for male and 12. 4 ms to 184. 3 ms for females. The average duration of consonants is much smaller compared to that vowels. A *vargga* classification also exists in Malayalam for consonant phonemes. There are 5 vargga classes in Malayalam. *Kavarggam /(*k/ ക, /k^h /ഖ, /g/ ഗ, /g^h /ഘ, /ŋ /ങ*)* , *chavarggam (*/c/ ച, /c^{h}ഛ, / ɟ/ ജ, /ɟʰ /ഝ, /ɲ/ ഞ*)* , *tavarggam (*/ʈ/ ട, / ʈh$^/$ഠ, /ɖ/ ഡ, /ɖʰ/ഢ, /ɳ /ണ/*)* , *thavarggam (*/t /ത, /t^h /ഥ, /d^h/ധ, /d/ദ, /n̪/ന) and *and pavarggam (*/P /പ, /p^h/ഫ, /b/ബ,

/b^h/ഭ, /m /മ) each consisting of 5 consonants. The 5 consonants in each vargga are characterised linguistically as –voiceless unaspirated, -aspirted, +voiced unaspirated, +aspirated and Nasal *(eg.* k ക: -voiceless unaspirated, k^h ഖ-: -aspirted, g ഗ: +voiced unaspirated, g^h ഘ: +aspirated, ŋ ങ : Nasal*).* A similar durational pattern exists across voiceless unaspirated, -aspirted, +voiced unaspirated, +aspirated and nasal consonants in a vargga. Figure 3. 2 shows the durational pattern of *ka-vargga* class.

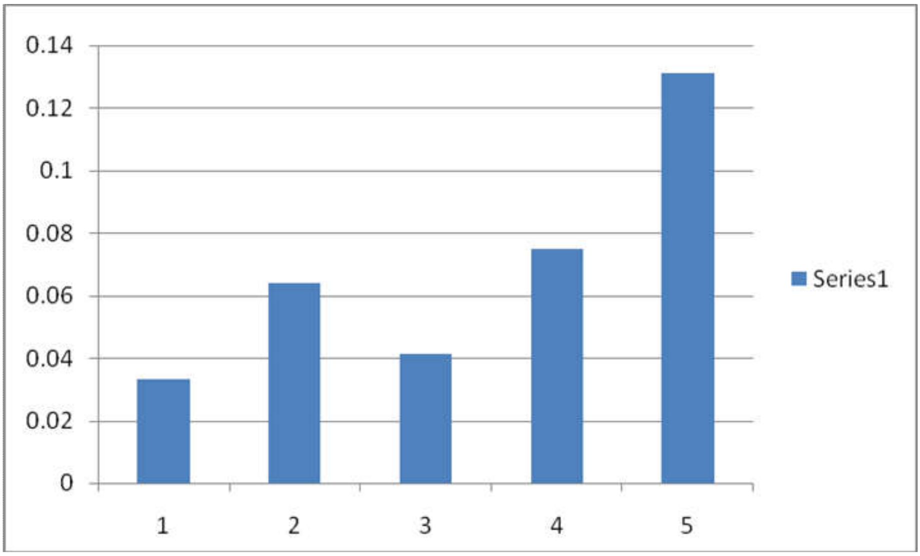


Figure 3. 2: The durational pattern among consonants in a vargga class, 1. –voiceless unaspirated, 2. –aspirted, 3. +voiced unaspirated, 4. +aspirated, 5. Nasal

The first 4 phonemes in each vargga class (expect nasals) and alveolar /r̠/ റcombine to form the plosive set in Malayalam. The nasals (/m /മ, , /n̪/ന, /n /ന, /ɲ/ ഞ/ɳ /ണ, /ŋ/ ങ), fricatives (/s /സ, /ʂ/ ഷ, , /ʃ

/ശ, /h /ഹ), trills (/r/ര/ṛ /റ), laterals (/l /ല/ൽ, , /ḷ/ ള/ൾ), approximants (/ʐ /ഴ) and Glides (/v/വ, /y /യ) are the remaining linguistic consonant classes in Malayalam. The durational analysis based on this classification is performed. The average duration of plosive is around 40 ms. Table 3.7 shows the average duration of each class of consonants. From the table it is evident that plosives have the smallest duration compared to other classes of consonants.

Table 3. 7: Average duration of each class of consonants

Consonant Class	Average Duration in second (male)	Average Duration in second (female)
Plosive	0. 03959	0. 04004
Nasal	0. 10894	0. 11574
Fricative	0. 10872	0. 12656
Trill	0. 10939	0. 12834
Lateral	0. 10233	0. 08386
Approximant	0. 08493	0. 09225
Glide	0. 060903	0. 06805

From the experimental results, it is evident that the Malayalam consonant and vowel allophone duration information together with the allophone transcriptor can be used for the development of visual speech synthesis systems which is detailed in the following chapters. The duration analysis performed so far is solely based on the audio speech signals. The following section explains the findings of some preliminary investigations on the audio visual asynchrony in Malayalam.

3.6 Audiovisual Asynchrony Model for Malayalam

Co-articulation reveals itself differently on two speech modalities and causes asynchrony between them [9]. Investigating and modelling the audiovisual asynchrony in a language is important for visual speech synthesis. Modelling this asynchrony is a key aspect for improving the naturality and intelligibility of audiovisual speech synthesis systems. The audiovisual speech asynchrony is observed to be different for different languages. An almost perfect audio visual synchrony has been reported for Japanese language [210]. Speech synthesis system has been developed in French which addresses the asynchrony in the language [211]. An elaborate scheme with separate audio -only and video–only HMMs has been employed for developing an asynchrony modelling frame work for Russian language [212].

This work performed a preliminary analysis to understand the audio visual asynchrony of Malayalam speech based on the MAVSC-IW corpora. Audio and image sequence of selected Malayalam words are labelled as phoneme sequences in both audio and visual domain. The audio also contains portions labelled as silence. The detailed durational information in the audio and visual domain for the word തർക്കം/tharkam/is given in table 3.8. Silence is the principal origin of asynchrony. It is evident that a significant audio visual synchrony exists between the audio and video manifestations in portions without silence.

Table 3.8: Durational information in the audio and visual domain for the word തർക്കം /tharkam/

Phoneme/ Silence	**Audio (start)**	**Audio (end)**	**Video (start)**	**Video (end)**
Silence	0. 68	0. 8		
Th	0. 86476	0. 88491	0. 68	0. 88
A	0. 88894	0. 98966	0. 92	0. 96
r	0. 99772	1. 13067	1	1. 08
silence	1. 13067	1. 20722		
K	1. 20722	1. 23945	1. 12	1. 24
a	1. 23945	1. 40060	1. 28	1. 32
m	1. 39658	1. 53356	1. 36	1. 52

A similar phenomenon where the silence creates an audiovisual asynchrony, is evident in most utterances. So the most crucial aspect in modelling the audio visual asynchrony in Malayalam is incorporating silence in the visual durational information. As the first step, the work performed a detailed analysis of interphone silence on sample words to model it. Out of the 900 words in the CMLTEMU dataset, 300 words are found to have silence between phonemes. Inter phoneme silence is preceded by plosive consonants in almost all occurrences. The silence is attributed to the stop phase or obstruction phase of plosive formation. The silence corresponding to the stop phase of a plosive is a property of the plosives or stop consonants [213].

It is also observed that, in the visual domain the articulatory positioning during silence (stop phase) has the maximum visual phoneme presence. For example consider the frames of the absolute closure corresponding to the articulatory preparation forthe consonant phoneme /P /പ, in the Malayalam word കറുപ്പ് /karuppu/, which is shown in figure 3.3.

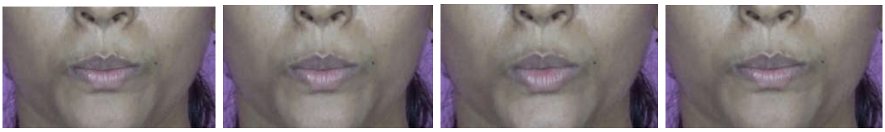


Figure 3.3: The frames corresponding to to the articulatory preparation for the consonant phoneme പ/P/, in the word കറുപ്പ് /karuppu/.

The phoneme /pa/ is visually more evident during silence created by complete block of air flow. Hence the duration of silence corresponding to the articulatory formation stage for the anticipatory plosive should be added to the plosive duration. Based on the observation a strategy is formulated for incorporating silence to the durational model to get rid of audio visual asynchrony in visual speech synthesis. Text to visual speech synthesis systems and audio to visual speech synthesis systems are addressed separately in the following way.

i. For an audio to visual speech synthesis system silence can easily be segmented from the audio signal. The duration of the silence is added to the next consonant phoneme duration.

ii. In a text to visual speech synthesiser duration is computed using information obtainable from text. In this context, Pre computed silence duration models are required. The solution adopted in this work is to compute the mean silence duration before each plosives from sample utterances. The average silence duration before plosives extracted from the sample utterances is summarised in table 3.9. Plosives with similar silence behaviour are grouped together and average silence duration across each group is listed in the table.

Table 3.9: The average silence duration before plosives

Class of Plosive	Isolated		Consonant Clusters	
	Male	Female	Male	Female
Voiceless Unaspirated Expect bilabial	0. 05741	0. 06502	0. 1363	0. 1517
Aspirated Expect Bilabial	0. 0843	0. 0919	0. 140722	0. 1523
Voiced Un aspirated Expect bilabial	0. 0818	0. 0858		
Aspirated Expect Bilabial	0. 0787	0. 0708		
P പ	0. 0968	0. 1208	0. 1394	0. 15203
p^h ഫ	0. 0922	0. 1133		
b ബ	0. 07347	0. 07345		

Naturally the silence duration for consonant clusters is higher compared to isolated consonants.

3.7 Conclusion

A detailed investigation is performed on the duration of allophonic variations of Malayalam Phonemes in this chapter. The durational properties obtained as part of the experiments can be effectively used for developing visual speech synthesis system in Malayalam. Rule set for the formation of Malayalam vowel and consonant allophones are derived as part of this study. The process involved in the creation of audio visual dataset is also explained in detail. Durational properties of the Malayalam vowel and consonant allophones are analysed. The phoneme and allophone duration patterns emerged from the statistical analysis of speech corpora are presented. It is found that the range of vowel duration varies from 13. 3 ms to 168. 4ms for male and varies from 12. 4 ms to 184. 3 ms for female. The average duration of consonants is much smaller compared to those vowels. Plosives have the smallest duration compared to other classes of consonants. The average duration of plosive is 40ms. An investigation is performed which reveals the basic natute of audio visual asynchrony in Malayalam and the findings can be effectively used for the development of Malayalam visual speech synthesis system. It is also observed that the inter phoneme silence in Malayalam is mainly attributed to the stop phase or obstruction phase of plosive formation.

Chapter 4

Rule based Grpaheme to Phoneme and allophone transcripter in Malayalam

4. 1 Introduction

Developing audiovisual speech synthesisers in mother tongue will be a mile stone activity in bridging the digital divide, a growing concern of governments and other organisations towards ensuring equality. Language specific explorations are required for discovering the components such as phoneme set, allophonic variations, viseme set and co-articulation effects in the audio and visual domain. Audio-visual speech synthesis systems often consider phoneme as the basic unit of synthesis. A phoneme is actually a mental abstraction forming in the brain of the listener. There might be infinite variations corresponding to a phoneme in the actual utterance. The variation happens due to many factors such as dialectics variation, emotional state of the speaker and co-articulation effects. The intelligibility and naturality of synthesised audiovisual speech depends on the ability of the system in incorporating these factors. While designing a text to audio visual speech synthesis system, conversion of the set of graphemes to a sequence of phonemes to get a phonemic transcription is the first stage. To accomplish a true phonetic representation additional information cues representing emotional, dialectical and co-articulation effects need to be incorporated into the phoneme sequence.

The allophonic variations of Malayalam phonemes explained in the last chapter accommodate most of the co articulation effects.

Co-articulation refers to changes in the articulation of a speech segment depending on preceding (backward or carry over co-articulation) and upcoming segments (forward or anticipatory co-articulation). Forward co-articulation is due to phonemes yet to be released. Carry over articulation is due to some phonemes happened at a previous time instant [199, 214]. Co-articulation happens either due to activities in brain or due to the inertia of motor systems. Anticipatory co-articulation is attributed to neuronal activities, but both motor system inertia and neuronal activities are responsible for carry over co-articulation. Allophone formation rule set in Malayalam characterises and enumerates the possible variations a phoneme can attain due to positional and contextual influences. So this work assumes that allophone characterisation in Malayalam encodes the co-articulaion effects on a phoneme. Additional information cues modelling continuous speech is visible in an allophone centric approach, compared to the phoneme based approach. This aspect is extremely significant in automatic speech processing, especially in developing audiovisual speech synthesis systems. A grapheme to allophone transcripter is an indispensable prerequisite for the allophone centric paradigm shift in Malayalam speech processing. This chapter explains the three stage development process of a grapheme to phoneme and allophone transcripter in Malayalam which can be used for speech synthesis and recognition applications.

Pre-processing, grapheme to phoneme conversion and phoneme to allophone mapping are the implementation stages of the proposed allophone transcripter. Firstly, the given grapheme sequence, after appropriate pre-processing in the grapheme domain is converted first in to a sequence of phonemes, and this sequence is converted to the corresponding sequence of allophones. Grapheme to Phoneme transcripter is reported for many languages. The underlying implementation strategies can be classified into dictionary based, data driven and rule based approaches. Dictionary based approach uses a phonetic dictionary with a transcribed phoneme sequence entry for each word. Storing and accessing this huge lexicon creates real time access issues. The huge manual labour of linguistic experts required for data preparation is another demerit of dictionary based approach [215]. Data driven approach uses machine learning algorithms to understand the behaviour of grapheme to phoneme transcription to automate the conversion. Hidden Markov Models (HMM), Artificial Neural Network (ANN) and Classification and Regression Trees (CART) using inductive learning technique *etc.* are the prominent techniques reported in the literature for this purpose [216-217]. Linguistic rule set can be formulated for phonetically perfect languages such as Sanskrit with written and spoken form correspondence. Successful implementations are also emerging in many languages which effectively combine different methods. Grapheme to Allophone transcriptors are reported for many languages such as Slovain and Arabic [218-219]. The phonetic or broad transcriptors whichuse allophone characterisation for encoding contextual information are reported from Spanish and Polish languages [54, 220].

The proposed work uses a rule based approach both for grapheme to phoneme transcription and phoneme to allophone transcription. This chapter explains the algorithm and implementation stages of a rule based Malayalam grapheme to phoneme and allophone transcripter. Pre-processing brings different class of graphemes to a unified framework required for the phoneme and allophone transcription. A comprehensive statistical and probabilistic analysis is performed, based on the novel phoneme and allophone converter developed as part of this study on standard word and sentence corpora. Statistical properties of phonemes and allophones thus derived can be used for the development of various language computing tools. Section 4.2 elaborates the processing steps employed for the development of grapheme to phoneme and allophone transcription. Section 4.3 sketches the steps in grapheme to phoneme transcription followed by the detailed explanation of phoneme to allophone transcription algorithm. Section 4.4 analyses and interprets the frequency of occurrence of phoneme and allophone using the newly developed transcripter and Section 4.5 concludes the chapter with relevant future directions.

4. 2 Pre-processing for the Development of Malayalam Graphemes to Allophone Transcription

Pre-processing brings different classes of graphemes to a unified framework required for the grapheme to allophone transcription. This Section explains the classification inherent in Malayalam orthography and the pre-processing performed on each class of graphemes. Malayalam is considered to be having an alphasyllabary writing system, which combines the features of an alphabet based and syllable based orthographic structure [201]. Even though the writing system of Malayalam language is based on the set of alphabet known as *'aksharamala'* (the garland of letters), the vowel signs in Malayalam makes it possible to write consonant vowel sequences as a single unit. The consonant vowel sign combinations are valid syllables in Malayalam. The Malayalam orthographic symbols are categorised as vowels, vowel signs, diphthongs, consonants, consonant compounds, *anusvaram, chandrakala and chillukal*. The following sections describe the pre-processing strategies required for each class of graphemes. A symbol array corresponding to the sequence of graphemes is the output of the pre-processing stage. As part of pre-processing some graphemes are replaced with other graphemes, while some others change its relative positioning.

4.2.1 Pre-processing for Vowels and Dependent Vowel Signs

Table 4.1 presents the grapheme symbols corresponding to Malayalam vowels and diphthongs with IPA symbols. Pre-processing causes no changes to these symbols or its positioning. They are mapped as such to the symbol array. Hence the vowels and diphthong graphemes in Malayalam are listed in Table 4.1, remains the same after pre-processing.

Table 4.1: Malayalam Vowels and Diphthongs with IPA Symbols

അ	ഇ	ഉ	എ	ഒ	ഋ	ഐ
/a/	/i/	/u/	/e/	/o/	/ṛ/	/ai/
ആ	ഈ	ഊ	ഏ	ഓ	-	ഔ
/a: /	/i: /	/u: /	/e: /	/o: /	-	/au/

Table 4.2 lists the dependent vowel signs of Malayalam language with examples and adopted pre-processing strategy for each. A vowel sign always occurs with a consonant known as the effected consonant. The vowel signs in Malayalam generally occur either to the left or right of the effected consonant. But the symbols /െ◌ാ/ and േ◌ാ has two components, one each in the left and right of effected consonant. The pre-processing unifies the positioning by placing all vowel signs to the right of the effected consonant.

Table 4. 2 Vowel signs in Malayalam with example usage

Vowel Sign	Corresponding Vowel	Example	Relative Position With the effected consonant
ി	ഇ/i/	കി/ki/	Right
ു	ഉ/u/	കു/ku/	Right
െ	എ/e/	കെ/ke/	Left
ൊ	ഒ/o/	കൊ/ko/	Left and Right
ൃ	ഋ/ṛ/	കൃ/kṛ/	Right
ാ	ആ/a: /	കാ/ka: /	Right
ീ	ഈ/i: /	കീ/ki: /	Right
ൂ	ഊ/u: /	കൂ/ku: /	Right
േ	ഏ/e: /	കേ/ke: /	Left
ോ	ഓ/o: /	കൊ /ko: /	Left and Right

4. 2. 2 Pre-processing for Consonant and Consonant Compounds

In Malayalam, a consonant grapheme symbol is associated for each consonant phoneme introduced in chapter 2. Each consonant grapheme symbol is retained as such after pre-processing. Consonant compounds form a prominent class of grapheme symbols in Malayalam. Consonant compounds are formed as a combination of more than one consonant in Malayalam. Based on the difference in the required pre-processing, consonant compounds are classified into nine categories [222]. Pre-processing replaces each consonant compound with its constituent consonants. For example ½ is replaced as മ+മ<m>+<m>, its constituent consonants and ണ്ട is replaced as (ണ+ട<na>+<ṭa>), consonants forming - . The work employed a look-up table strategy for replacing consonant compounds with its constituent consonants. Table 4.3 shows a sample look up table depicting the conversion of each class of compound consonants. The look-up table consists of a total of 93 entries which maps the entire Malayalam consonant compounds.

Table 4. 3: Portions of look-up table for replacing consonant compounds with constituent phonemes

Compound Consonant	Constituent Consonants
ത്ത	ത, ത(<t>, <t>)
. . .	. . .
പ്പ	പ, പ(<pa>, <pa>)
. . .	. . .
ഞ്ച	ഞ, ച(<ɲa>, <ca>)
. . .	. . .
ന്ത	ന, ത(<na>, <ta>)
. . .	. . .
ട്ട	ട, ട(<ʈa>, <ʈa>(
. . .	. . .
ഹ്ന	ഹ, ന(<ha>, <na>)
. . .	. . .
ക്ത്ര	ക, ത, ര(<ka>, <ta>, <ṛa>)
. . .	. . .
വ്വ	വ, വ(<va>, <va>)
. . .	. . .
ങ്ക	ങ, ക(<ŋ><ka>)

4. 2. 3. Pre-processing applied for Special grapheme symbols

In Malayalam *Anusvaram, chandrakala, chillukal*and consonant diacritics are treated as special symbols. *Anusvaram* represented by grapheme symbol / ◌ം/ is actually a consonant മ/m / after a vowel. *Anusvaram* is usually considered as a vowel in grapheme categorisation. *Chandrakala* ◌് /ə/, a diacritic, occurs in the word ending attached to a consonant without a natural vowel ending. *Chillu* is special class of consonants which is characterised by the absence of an inherent vowel following the consonant. The 5 grapheme symbols under this class are ൺ, ൻ, ർ, ൽ, ൾ corresponding to the base characters <ṇa> ണ, <na> ന, <ra> ര, <la> ല and <ḷa> ള. Pre-processing retaines *Anusvaram, chandrakala* and *chillu* as such for further processing.

Malayalam has 4 special grapheme symbols for representing the consonant–approximant combination, known as consonant diacritic. The occurrence of the symbol is replaced by corresponding approximant to the right of the effected consonant. Table 4. 4 shows the usage of different approximant signs with ക /ka/ as the effected consonant**.** The table also shows the portion of the signs with respect to the effected consonant. Diacritics are replaced by the consonant approximant combination as per the usage given in Table 4. 4. For example, l y<*kya*>is replaced by **‘**ക്+യ’ combination.

Table 4.4: Use of approximant signs based on the corresponding approximant withക /ka/ as the effected consonant

Approximant Sign with consonant ക /ka/	Corresponding Approximant	Position of the sign With respect to the effected consonant
ക്യ <kya>	യ	Right
ക്ര<kra>	ര	Left
ക്ല <kla>	ല	Under
ക്വ<kva>	വ	Right

As part of pre-processing the input text is converted into a symbol array based on the rules described in section 4. 2. 1 to 4. 2. 5. The symbol array formed after applying processing the word sXm¸n is shown in Table 4. 5.

Table 4.5: The symbol array after the pre-processing on the example Malayalam word sXm¸n

തൊപ്പി				
ത	ൊ	പ	പ	ഇ

The following section describes the process involved in the development of Malayalam text to allophone transcritper in detail.

4.3 Comprehensive Malayalam Grapheme to Allophone Transcripter

The grapheme to allophone transcripter is implemented in two phase frame work. Initially, the sequence of graphemes corresponding to a given word, $\alpha_1, \alpha_2 . . . \alpha_m$ is initially mapped to a sequence of phonemes $\beta_1, \beta_2 . . . \beta_n$

$$f(\alpha_1, \alpha_1 . . . \alpha_m) = \beta_1, \beta_2 . . . \beta_n$$

In the second stage, each phoneme β is mapped to one of its allophones γ_i’s

$$g(\beta_1, \beta_2 . . . \beta_n) = \gamma_1, \gamma_2 . . . \gamma_n$$

The set of graphemes, phonemes and allophones in Malayalam is described in chapter 3. The mapping functions f and g receives the current instance, position and neighbouring instances as inputs. Section 4. 3. 1 discusses the proposed algorithm for grapheme to phoneme transcription and section 4. 3. 2 elaborates the phoneme to allophone transcription process proposed as part of the work.

4. 3. 1 Grapheme to Phoneme Transcription Algorithm

The grapheme to phoneme mapping for Malayalam is implemented using a rule based approach. The output obtained from the previous pre-processing step, symbol_array, is the input to this algorithm. Phonemic_Out_Array is the output of this transcription algorithm, which is an array containing the sequence of phonemes corresponding to the grapheme sequence.

The detailed subroutine based algorithm used for grapheme to phoneme transcription is available in the work by Vivek [12]. The transcription rules employed in each subroutine is summarised as 15 groups in table 4. 6. The second column of the table displays the instance of the symbol array which acts as a rule set. The current symbol is highlighted in a different colour. One left neighbour and two right neighbours are also considered for making decisions. Don’t care conditions are indicated by a /-/. The framework uses the look up tables, for vowel signs and long vowels as shown in Table 4. 7 and for plosive aspirated consonants as shows in table 4. 8.

Table 4. 6: Summary of the grapheme to phoneme transcription rules

Sl. No.	Description				Action
Group 1					
1.	<ൻ>	<ു>	<െ>	-	Insert /ഺ/ to Phonemic_Out_Array
2.	<ൻ>	<ു>	്>	<ു>	Insert / ്ു/ to Phonemic_Out_Array and advance symbol_Array by two positions
Group 2					
3.	vowel or vowel symbol	◌ം	-	-	Insert / മ് / to Phonemic_Out_Array
4.	Any other symbol	◌ം	-	-	Insert / അ + മ് / to Phonemic_Out_Array
Group 3					
5.	<ന>	<ഡ>	-		Insert /ന്/to Phonemic_Out_Array
6.	<ന>	NOT <ഡ>	-		Insert /ന്1/to Phonemic_Out_Array
7.	ര<r>, ു/ർ< ṛ>	<ന>	-		Insert /ന്1/to Phonemic_Out_Array
8	-	<ന>	**Any Dental**		Insert /ന്1/to Phonemic_Out_Array
9.	vowel	<ന>	<ന>	**vowel**	Insert /ന്1/to Phonemic_Out_Array and advance Symbol_Array by one position
10.	Other	<ന>	others	**other**	Insert /ന്/to Phonemic_Out_Array
Group4					
11.	-	Aspirated Plosive	-	-	Insert corresponding entry from Look_Up_Table2 in to Phonemic_Out_Array

12.	-	Un Aspirated Plosive	**-**	**-**	Insert plosive to Phonemic_Out_Array
Group 5					
13.	-	Long vowels or Vowel symbols	**-**	**-**	Insert corresponding entry from Look_Up_Table1 in to Phonemic_Out_Array
Group 6					
14.	-	All remaining cases	**Vowel**	**-**	Insert / symbol + /മ്/(CHANDRAKALA) / to Phonemic_Out_Array
15	-	All remaining cases	**NOT Vowel**	**-**	Insert / symbol + /മ്/ + /അ/ to Phonemic_Out_Array

Table 4. 7: Look-up table for vowel signs and long vowels

Sl. No:	Grapheme	Phoneme
1	ി	ഇ/i/
2	ു	ഉ /u/
3	െ	എ /e/
4	ൊ	ഒ /o/
5	ൃ	ഋ /ṛ/
6	ാ or ആ /a: /	അഅ /a: /
7	ീ or ഈ /i: /	ഇഇ/i: /
8	ൂ or ഊ /u: /	ഉഉ /u: /
9	േ or ഏ /e: /	എഎ /e: /
10	ോ or ഓ /o: /	ഒഒ /o: /

Table 4. 8: Lookup table for plosive aspirated

Sl. No:	Grapheme	Phoneme
1	ഫ് /p^h/	പ്ഹ്/Ph/
2	ഥ് /t^h/	ത്ഹ്/t h/
3	ഠ്/ʈ^h/	ട്ഹ്/ʈh/
4	ഛ്/c^h/	ച്ഹ്/t h/
5	ഖ് /k^h/	ക്ഹ്/k h/
6	ഭ് /b^h/	ബ്ഹ്/b h/
7	ധ് /d^h/	ദ്ഹ്/d h/
8	ഢ് /ɖ^h/	ഡ്ഹ്/ɖ h/
9	ഝ് /ɟ^h/	ജ്ഹ്/ɟ h/
10	ഘ് /g^h/	ഗ്ഹ്/g h/

The Phonemic_Out_Array, an array of Malayalam phonemes corresponding to the input grapheme sequence obtained from the stage, is the output of grapheme to phoneme transcripter in Malayalam. The same Phonemic_Out_Array wil also act as the input to the phoneme allophone converter which is described in the following section.

4. 3. 2 Phoneme to Allophone Transcripter

The allophone characterisation of Malayalam phonemes is described in section 2. 3. The Malayalam allophone set consists of 107 elements. The 51 Malayalam phonemes and its corresponding allophones are depicted in table 4. 9.

Table 4. 9: Malayalam Phonemes with Allophone

Phoneme	Allophone Symbols	Phoneme	Allophone Symbols	Phoneme	Allophone Symbols
അ/a/	[ʌ]	ങ/ŋ/	[ŋ]	ഫ/ph/	[pha]
	[A]		[ŋ^{j}]	ബ/b/	[B]
ആ/a: /	[a:]		[ŋ$^{<}$]		[b]
	[a]		[ŋ$^{>}$]	ഭ/b^{h} /	[bha]
ഇ/i/	[i]		[ŋ']	മ/m/	[m̪h]
	[^{y}i]				[M]
	[y^{i}]				[m]
ഈ/i: /	[^{y}i:]	ച/c/	[c]		[m]
	[i:]		[ç̬]		
ഉ/u/	[^{w}u]		[ɟ]	യ/y /	[ya]
	[u^{w}]		[C]	ര/r/	[ra]
	[ɯ]	ഛ/ch /	[c^{h}]	ല/l/	[la]
	[ə]		[C^{h}]	വ/v/	[w]
	[ə*]	ജ/ɟ /	[J]		[v]
	[ɯv]		[j]	ശ/ʃ /	[Sa]
	[U]	ഝ/ɟh /	[jha]	ഷ/ʂ/	[sha]
ഊ/u: /	[^{w}u:]	ഞ/ɲ /	[ɲa]	സ/s/	[sa]
	[u]	ത/t/	[t]	ഹ/h/	[H]
എ/e/	[^{y}e]		[t"]		[h]
	[^{e}y]		[ð]	ള/ɭ /	[ɭa]
	[E]		[d̪]	ഴ/ʐ/	[ỵa]
ഏ/e: /	[^{y}e:]	ഥ/th/	[tha]	റ/ṛ/	[ṛə]
	[e^{r}:]	ദ/d/	[ḍ]	ന/n/	[nh]
	[e:]		[d]		[n]
ഒ/o/	[^{w}O]	ധ/dh /	[dha]		[T^{h}]

	[O]	ന/n̪/	[n̪]		
ഓ/o: /	[ʷo:]		N		
	O	ട/ʈ/	[ɖ]		
ഐ /ai/	[ai]		[ɽ]		
	[ei]		[ʈ]		
ഔ /au/	[au]		[T]		
		റ്റ/r̠/	[d]		
ക/k/	[k]		[t]		
	[kj]	ഠ/ʈh/	[ʈʰ]		
	[ɣ]				
	[ɠ]	ഡ/ɖ/	[da]		
	[ʈ]	ഢ/ɖʰ /	[dha]		
	[K]	ണ/ɳ/	[ɳa]		
ഖ/kh /	[kʰ]				
	[Kʰ]	പ/P/	[p]		
	[Kʰ]		[β]		
ഗ/g /	[G]		[b]		
	[g]		[P]		
ഘ/gh/	[gʰ]				

The phoneme to allophone converter maps the phoneme sequence to its corresponding allophone sequence where each phoneme β is mapped to one of its allophones γ_i's

$$g(\beta_1, \beta_2 . . . \beta_n) = \gamma_1, \gamma_2 . . . \gamma_n$$

For a grapheme to phoneme transcripter the number of allophones in the output sequence is always equal to the number of phonemes in the input sequence.

In the rule based phoneme to allophone converter, each phoneme is mapped to one of its allophones based on the allophone formation rule. The mapping rules for a phoneme is frames basedon the positioning and left-right neighbours of phoneme. The three positioning options are characterised as word initial, word final and middle. All phonemes occurring in positions other than word beginning and word end are treated as middle phonemes. Most of the vowel allophone mappings are based solely on positioning. In Malayalam all vowels apart from ഔ/<au>/ has got more than one allophone characterisations, while there are 19 consonant phonemes with just one allophone. The pseudo code for the proposed phoneme to allophone algorithm is given bellow.

Phoneme to Allophone Conversion Algorithm for Malayalam – Pseudo- code

```
// Input : Phonemic_Out_Array, the sequence of phonemes obtained from grapheme to phoneme transcripter

//Output : Allophonic_Out_Array, the sequence of allophones corresponding to the input

    Len_Array=length(Phonemic_Out_Array)
    Current_Pos = 1
    While(Curent_Pos <= Len_Array)
```

```
        {
                Curent_phone = Phonemic_Out_Array[Curent_Pos]
                Left_Phone = Phonemic_Out_Array[Curent_Pos -1]
                Right_Phone = Phonemic_Out_Array[Curent_Pos +1]
                if (Curent_Pos =1) then
                        Position=Initial
                else if (Curent_Pos = Len_Array) then
                        Position=Final
                else
                        Position=Middle
                end if
Allophonic_Out_Array[Curent_Pos]=
Phone_to_Allophone(Phonemic_Out_Array, . . .
Curent _phone, Position, Left_Phone, Right_Phone)
                Curent_Pos = Curent_Pos + 1
        } // end while
```

Phone_to_Allophone is a subroutine which map the current phoneme to its allophone. The mapping requires current phoneme, its position with the neighbouring phonemes. Hence the Phonemic_Out_Array and position are passed as arguments to the subroutine in addition to the current phoneme. The pseudo code for Phone_to_Allophone conversion is given bellow.

```
Phone_to_Allophone(Phonemic_Out_Array, Curent_Phone, Position,
Left_Phone, Right_Phone)
{
        If( isVowel(Curent_Phone)) then
                Vowel_to_Allophone (Curent_Phone, Position,
                Phonemic_Out_Array, Left_Phone, Right_Phone)
        else
                Consonant_to_Allophone (Curent_Phone, Position,
                Phonemic_Out_Array, Left_Phone, Right_Phone)
        End if
}
```

Phone_to_Allophone subroutine just checks whether the current phone is a vowel orconsonant. Vowel_to_Allophone subroutine will be called for vowels based on the isBoolean check operation which returns TRUE if the current phoneme is a vowel.

```
isVowel(Current_Phone)
{
        if(Curent_Phone==’അ’ or Curent_Phone==’ആ’ or
        Curent_Phone==’ഇ’ or Curent_Phone==’ഈ’ or
        Curent_Phone==’എ’ or Curent_Phone==’ഏ’ or
        Curent_Phone==’ഉ’ or Curent_Phone==’ഊ’ or
        Curent_Phone==’ഒ’ or Curent_Phone==’ഓ’ or
        Curent_Phone==’ഐ’ or Curent_Phone==’ഔ’) then
```

```
                return TRUE
        else
                return FALSE
        endif
}
```

The Vowel_to_Allophone conversion routine converts the vowel phoneme to corresponding to vowel allophone.

```
Vowel_to_Allophone (Curent_Phone, Position, Phonemic_Out_Array,
Left_Phone, Right_Phone)
{
        If (Curent_Phone==’അ’ and position==initial and
position==final ) then
                Return (‘[ʌ]’)
        Else if (Curent_Phone==’അ’ )then
                Return(‘[A]’)
        Else if (Curent_Phone==’ആ’ )then
                res= call check_Velar()
                if res== true then
                        return (‘[a: ]’)
                else
                        return false
        else if (Curent_Phone==’ആ’ )then
                return(‘[a: ]’)
```

```
else if (Curent_Phone==’ഇ’ and position==final) then
        return(‘[y i ]’)
else if (Curent_Phone==’ഇ’ and position==initial) then
        return(‘[ y i]’)
else if (Curent_Phone==’ഇ’ and position==middle) then
        return(‘[i]’)
else if (Curent_Phone==’ ഈ’ and position==initial) then
        return(‘[ y i: ]’)
else if (Curent_Phone==’ ഈ’ and position==middle) then
        return(‘[i: ]’)
else if (Curent_Phone==’എ’ and position==initial) then
        return(‘[ y e]’)
else if (Curent_Phone==’എ’ and position==final) then
        return(‘[e y]’)
else if (Curent_Phone==’എ’ and position==middle) then
        return(‘[E]’)
else if (Curent_Phone==’ഏ’ and position==initial) then
        return(‘[ y e: ]’)
else if (Curent_Phone==’ഏ’ and position==final) then
        return(‘[e r : ]’)
else if (Curent_Phone==’ഏ’ and position==middle) then
        return(‘[e: ]’)
```

```
else if (Curent_Phone==’ഉ’ and position==initial) then
          return(‘[ w u]’)
else if (Curent_Phone==’ഉ’ and position==final) then
          return(‘[u ]’)
else if (Curent_Phone==’ഉ്’ and position==final) then
          return(‘[ə]’)
else if (Curent_Phone==’ഉ’ )then
          res= call check_Lateral()
          if res== true then
                    return (‘[ə*]’)
          else
                    return false
else if (Curent_Phone==’ഉ’)then
          return(‘[ɯ]’)
Else if (Curent_Phone==’ഊ’ and position==initial) then
          Return(‘[ w u: ]’)
Else if (Curent_Phone==’ഊ’) then
          return(‘[u]’)
else if (Curent_Phone==’ഒ’ and position==initial) then
          return(‘[ w O]’)
else if (Curent_Phone==’ഒ’ and position==middle) then
          return(‘[O]’)
```

```
        else if (Curent_Phone==’ഓ’ and position==initial) then
                return(‘[ W. O: ]’)
        else if (Curent_Phone==’ഓ’) then
                return(‘[O]’)
        else if (Curent_Phone==’ഐ’ and position==initial) then
                return(‘[ai]’)
        else if (Curent_Phone==’ഏ’ and Right_Phone==’യ’ )then
                return(‘[ei]’)
}
```

The Vowel_to_Allophone subroutine uses check_Velar and check_Lateral routines to check corresponding class of consonants in the neighbourhoods.

```
Boolean check_Velar(curent_phone, Left_Phone)
{
        If(Left_Phone==’ക’ or Left_Phone==’ഖ’ or
Left_Phone==’ഗ’ or Left_Phone==’ഘ’ or Left_Phone==’ങ’) then

                Return          True
        Else
                Return          False
        End if
}
Boolean check_Lateral(curent_phone){
        If(Left_Phone==’ൺ’ or Left_Phone==’ർ ’ or
Left_Phone==’ൾ’ or                    Left_Phone==’ൽ’){
                Return True
        }
        Else
                Return False
        End if
}
```

The subroutine Consonant_to_Allophone processes consonant phoneme to return corresponding allophone. The python implementation of Consonant_to_Allophone employes separate functions for each consonant phoneme. As the number of rules characterising consonant allophones is comparatively large and each comprises complex logic, presenting each sub routine with if-else sequence will obscure the description. Hence a table based description, as shown in table 4. 10 with allophone and corresponding rule set is adopted for Consonant_to_Allophone subroutine. For example the first entry in the table depicts the subroutine for പ/pa/ and can be interpreted as

```
--
if(Position==initial) then
        return [p]
else if(Check_Intervocalic())
```

```
        return [β]
else if (Position!=initial AND Left_Phone==’മ’)
        return [b]
else if(NOT check_Nasal())
        return [P]
-----
```

Table 4. 10: Allophone and rule set used for the implementation of Consonant_to_Allophone subroutine

Phone	Allophone	Mapping logic
പ/P/	[p]	If(Position==initial) then
	[β]	p=Check_Intervocalic() If (p ==TRUE) then
	[b]	If (Position!=initial AND Left_Phone==’മ’)then

	[P]	p=check_Nasal()
		if(p==false) then

ബ/b/	[b]	p=Check_Intervocalic() If (p ==TRUE) then
	[B]	Else
ത/t/	[t]	If(Position==initial) then
	[ð]	p=Check_Intervocalic() If (p ==TRUE and Left_Phone==’മ’) then

	[d̪]	p=check_Nasal()
		if(p== TRUE and position==middle) then

	[t̍']	p=check_Nasal()
		if(p==false) then

ദ/d̲/	[ḍ]	p=Check_Cluster() If (p ==TRUE) then
	[d]	Else

റ/ṟ/	[t]	If(Position!=initial) then

	[d]	p=check_Nasal()
		if(p==true) then

ട/ʈ/	[ʈ]	If(Position==initial) then
	[ɖ]	If(Position!=initial AND Left_Phone==’ണ’) then
	[ɽ]	p=Check_Intervocalic() If (p ==TRUE) then
	[T]	If(Position!=initial) then

ഠ/ʈh/	[ʈ^{h}]	p=Check_Intervocalic()
		q=check_Nasal()
		If (p ==TRUE and
		q==TRUE) then

	[T^{h}]	Else
ച/c/	[c]	If(Position==initial) then
	[ɟ]	If(Position!=initial AND Left_Phone==’ഞ’) then

	[C]	p=check_Nasal()
		if(postion!=initial and
		Right_Phone===’ച’ and
		p==false) then

	[c̬]	Else
ജ/ɟ /	[J]	Else
	[j]	If(Position==initial) then
ക/k/	[ʈ]	If(Right_Phone==’ഷ’) then
	[ɠ]	If(Left_Phone==’ങ്’) then
	[ɣ]	p=Check_Intervocalic() If (p ==TRUE and left==’ ധ’ and Right_Phone!=’ ക’) then

	[kj]	p=check_LeftVowel()
		if(p==TRUE and
		Right_Phone===’ക’)then

	[k]	If(Position==initial and Right_Phone===’ക’) then
	[K]	Not coded

ഖ/kʰ /	[K^h]	If(Position==initial) then

	[k^h]	p=Check_Intervocalic()
		q=check_Nasal()
		If (p ==TRUE and q==TRUE) then

	[K^h]	Else
ഗ/g /	[g]	If(Check_Intervocalic() and position==initial) then
	[G]	Else
ഘ/gʰ /	[gʰ]	Curent_phone
മ/m /	[M]	If(check_alveolarFlap()) then
	[m]	If(Right_Phone==’വ’)then
	[m̪ʰ]	If(Left_Phone==’ഹ’) then
	[m]	Else
ന/n̪ /	[n̪]	If(check_alveolarFlap()) then
	N	Else
ന/n/	[n^h]	If(Left_Phone==’ഹ’) then
	[n]	Else
ങ/ŋ/	[ŋ]	If(Right_Phone==’ക’)then
	[ŋʲ]	If(check_FrontVowel())then
	[ŋ$^>$]	If(check_BackVowel())then
	[ŋ$^<$]	If(check_VowelDiphthongs()) then
	[ŋ']	Else
ഹ/h/	[H]	Curent_phone
വ/v/	[v]	If(Right_Phone==’ഉ’)then
	[w]	Else

The conversion of consonant phonemes to its allophones uses the following 8 subroutines.

```
Boolean Check_Intervocalic (curent_phone)
{
        If(Left_Phone==’അ’ or Left_Phone==’ഇ’ or Left_Phone==’ഉ’ or Left_Phone==’ഒ’ or Left_Phone==’ഏ’)
        {
                If(Right_Phone==’അ’ or Right_Phone==’ഇ’ or Right_Phone==’ഉ’ or Right_Phone==’ഒ’ or Right_Phone==’ഏ’){
                        Return True
                Else
                        Return False
                Endif
}
Else
        Return False
End if
}
Boolean check_Nasal (curent_phone)
{
        If(position==middle){
                If(Left_Phone==’മ’ or Left_Phone==’ണ’ or Left_Phone==’ങ’ or  Left_Phone==’ഞ’ or Left_Phone==’ന’)
                        Return True
                Else
                        Return False
                End if
        }
        Else
                Return False
        End if
}
Boolean check_leftVowel (curent_phone)
{
        If(Left_Phone==’അ’ or Left_Phone==’ഇ’ or Left_Phone==’ഉ’
or Left_Phone==’ഒ’ or Left_Phone==’എ’){
                Return True
        Else
                Return False
        End if
}
Boolean check_alveolarFlap (curent_phone)
{
        If(Left_Phone==’ർ ‘ or Left_Phone==’ര’ or Left_Phone==’റ’)
                Return True
        Else
                Return False
        End if
}
Boolean check_frontVowel(curent_phone){
        If(Left_Phone==’ഇ’ or Left_Phone==’ ഈ’ or
Left_Phone==’എ’ or Left_Phone==’ഏ’){
        Return True
    }
    Else
        Return False
    End if
}
Boolean check_Voweldiphthongs(curent_phone){
    if(check_Vowel())
    {
        if(Left_Phone==’ഐ’ or Left_Phone==’ഔ’){
            return True
        }
        Else
            Return False
        End if
    Else
        Return False
    End if
}

Boolean check_BackVowel(curent_phone){
    If(Left_Phone==’ഉ’ or Left_Phone==’ ഊ’ or Left_Phone==‘ഒ’
or Left_Phone==’ഓ’)
        Return True
        else
                Return False
        End if
}
Boolean check_Lateral (curent_phone)
{
        If(Left_Phone==’ഞ്’ or Left_Phone==’ ര് ’ or Left_Phone==’
ത്’ or  Left_Phone==’ ഗ്’){
                Return True
        }
        Else
                Return False
        End if
}
```

The next section discusses the usage of the proposed grapheme to allophone transcripter to derive the phoneme and allophone statistics of Malayalam language with atext corpora consisting of 82, 324 words.

4.4. Phoneme and Allophone Statistical Analysis Based on the Proposed Grapheme to Allophone Transcripter

Statistical linguistics at the phoneme, word and sentence level is part of the general language modelling frame work [223-227]. The developed grapheme to allophone transcripter in Malayalam is used for statistical analysis at the phoneme and allophone level. The results obtained in the analysis can be effectively integrated in the various phases of development of automatic speech recognition, automatic translation, audiovisual speech synthesis, spell checking systems *etc.* Knowledge about the permissible phoneme and allophone combinations in a language is decisive in language identification systems [228-232]. The pattern analysis of phoneme and allophone combinations, which can be obtained by applying transcription on word and sentence corpora can improve the performance of spell checking systems [233-234]. Frequency of occurrence of Japanese and Russian phonemes and diphones based on a respective corporas are reported inliterature [235-236]. Frequency of occurrence analysis of phonemes is used for concatenative speech synthesis in Catalan language [237]. Initiatives in Polish language are instrumental in developing corpora, grapheme to phoneme transcripter and in performing statistical analysis at grapheme, phoneme and word level [228]. An attempt in Malayalam is made by N. Sreedevi *et al.* based on manual selection and conversion on a limited word data set [239].

This section reports the estimated phoneme and allophone frequency of occurrence on a word and sentence corpora in Malayalam to verify the effectiveness of the proposed transcripter in the above mentioned applications. The phonemic corpora for the analysis is obtained from grapeme to phoneme transcripter. The allophone corpora, obtained by applying phoneme to allophone transcripter, can act as a phonetic data base which can model the positional and neighbourhood information of phoneme occurrence. The word corpora is taken from the online Malayalam-English dictionary Olam [32]. The words and phoneme counts in the corpora are given in table 4.11.

Table 4. 11: Word and phoneme counts of text corpora

Source	Number Of Words	Number Of Phonemes
Oalam Online dictionary	82, 324	7, 63, 392

A detailed analysis of frequency of occurrence of Malayalam allophones is carried out against the text corpora. The frequency of occurrence of each allophone with its percentage of occurrence is given in Table 4. 12.

Table 4.12: The frequency of occurrence of each allophone with its percentage of occurrence

Phoneme	Allophone	Frequency	Percentage
അ/a/	[ʌ]	2512	0. 721347818
	[A]	53902	15. 47853904
ആ/a: /	[a:]	3784	1. 086616299
	[a]	14565	4. 182496403
ഇ/i/	[i]	24816	7. 126181308
	[^{y}i]	781	0. 22427255
	i^{y}	1698	0. 487598963
ഈ[i:]	[^{y}i:]	27	0. 00775334
	[i:]	2026	0. 581787691
ഉ[u]	[^{w}u]	684	0. 196417957
	[u^{w}]	2095	0. 601601783
	[ɯ]	17215	4. 943472405
	[ə]	5791	1. 662947935
	[ə]	0	0
	[ə]	0	0
	[U]	1704	0. 489321927
ഊ/u: /	[^{w}u:]	1659	0. 476399693
	[u]	0	0

എ/e/	[^{y}e]	695	0. 199576725
	[e^{y}]	3212	0. 922360347
	E	5711	1. 639975074
ഏ/e: /	[^{y}e:]	168	0. 048243007
	e^{r}:	1332	0. 382498126
	[e:]	3502	1. 005636966
ഒ/o/	[^{w}O]	360	0. 103377872
	[O]	1401	0. 402312218
ഓ/o: /	[^{w}o:]	173	0. 049678811
	O	3688	1. 059048866
ഐ /ai/	[ai]	499	0. 143293217
	[ei]	0	0
പ/P/	[p]	3621	1. 039809096
	[β]	966	0. 27739729
	[b]	519	0. 149036432
	[P]	5867	1. 684772152
ബ/b/	[B]	797	0. 228867122
	[b]	208	0. 059729437
മ/m/	[m̪^{h}]	2	0. 000574322
	[M]	191	0. 054847704
	[m]	662	0. 19010042
	[m]	16747	4. 809081172
ത/t/	[t]	1441	0. 413798649
	[t̍']	13597	3. 904524792
	[ð]	3666	1. 05273133
	[d̪]	692	0. 198715243
ദ/d/	[ḍ]	545	0. 156502612
	[d]	1661	0. 476974015
ന/n̪/	[n̪]	184	0. 052837579
	N	2187	0. 628020572
ന/n/	[n^{h}]	6	0. 001722965
	[n]	20572	5. 907471061
റ/r̠/	[d]	716	0. 205607101
	[t]	911	0. 261603448
ട/ʈ/	[ɖ]	2053	0. 589541031
	[ɽ]	3592	1. 031481434

	[ʈ]	111	0. 031874844
	[T]	6021	1. 728994909
ഠ/ʈʰ/	[ʈʰ]	63	0. 018091128
	[Tʰ]	44	0. 012635073
ച/c/	[c]	880	0. 252701465
	[ç̬]	2581	0. 74116191
	[ɟ]	160	0. 045945721
	[C]	1933	0. 55508174
ഛ/ cʰ/	[cʰ]	45	0. 012922234
	[Cʰ]	35	0. 010050626
ജ/ɟ /	[J]	599	0. 172009293
	[j]	782	0. 224559711
ക/k/	[k]	10254	2. 944546387
	[kj]	6	0. 001722965
	[ɣ]	5147	1. 478016408
	[ɠ]	574	0. 164830274
	[ʈ]	749	0. 215083406
	[K]	0	0
ഖ/kʰ/	[kʰ]	173	0. 049678811
	[Kʰ]	27	0. 00775334
	[Kʰ]	203	0. 058293633
ഗ/g /	[G]	1099	0. 31558967
	[g]	501	0. 143867538
ഘ/gʰ/	[gʰ]	276	0. 079256369
ങ/ŋ/	[ŋ]	2571	0. 738290302
	[ŋj]	122	0. 035033612
	[ŋ<]	0	0
	[ŋ>]	1784	0. 512294788
	[ŋ']	0	0
ഹ/h/	[H]	0	0
	[h]	1078	0. 309559294
വ/v/	[w]	6090	1. 748809001
	[v]	2131	0. 61193957

ഫ/pʰ/	[pʰ]	384	0. 11026973
ഭ/bʰ /	[bʰ]	1272	0. 365268481
ഥ/tʰ/	[tʰ]	795	0. 228292801
ധ/dʰ /	[dʰ]	1756	0. 504254287
സ/s/	[s]	5877	1. 68764376
റ/ṛ/	[ṛ]	5564	1. 597762443
ല/l/	[l]	8456	2. 428231348
ഡ/ɖ/	[ɖ]	647	0. 185793009
ണ/ɳ/	[ɳ]	6559	1. 883487395
ഷ/ʂ/	[ʂ]	2193	0. 629743537
ഢ/ɖʰ /	[ɖʰ]	19	0. 005456054
ള/ɭ /	[ɭ]	5981	1. 717508478
ഴ/ʐ/	[ʐ]	1052	0. 302093115
ഝ/ɟʰ /	[ɟʰ]	0	0
ഞ/ɲ /	[ɲ]	1083	0. 310995098
ശ/ʃ /	[ʃ]	1978	0. 568003974
യ/y /	[y]	13865	3. 981483874
ഔ/au/	[au]	221	0. 063462527
ര/r/	[r]	9093	2. 611152749

From the table 4.12, it is observed that അ/a/ is the most frequently occurring phoneme and ത/t/ is the top among consonants. The probability distribution of Malayalam phonemes in decreasing order is shown in figure 4. 1. Phoneme ഝ/ɲ / is the least frequently occurring phoneme in Malayalam.

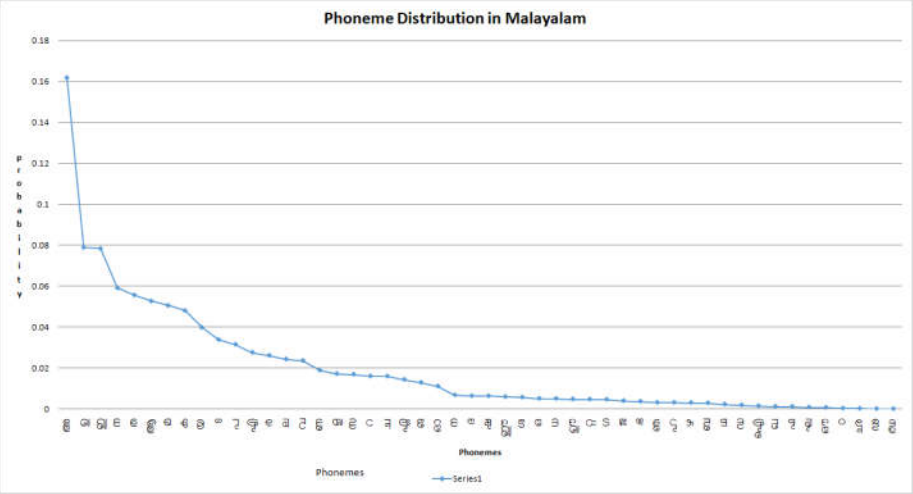


Figure 4.1: The probability distribution of Malayalam phonemes in decreasing order

Table 4.12 also shows the non-uniform pattern in the frequency of occurrence of allophones of the same phoneme. Figure 4.2 depicts the pattern in the frequency of occurrence of vowel allophones grouped in to phonemes.

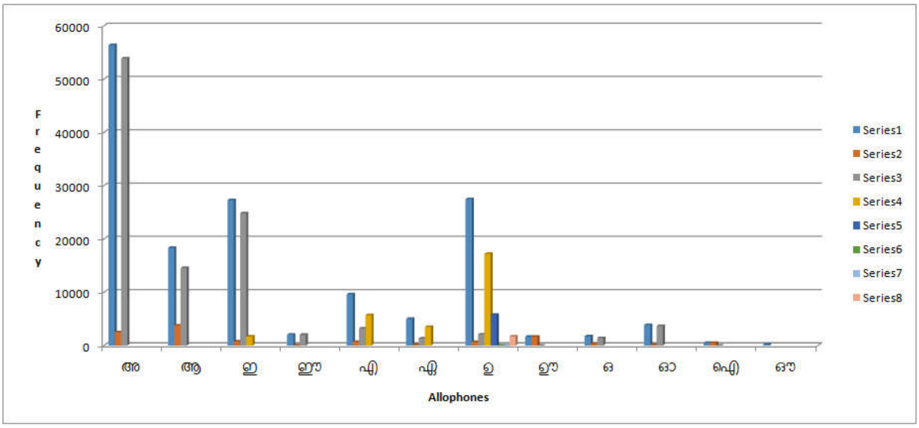


Figure 4. 2: Frequency of occurrence of vowel allophones grouped in to phonemes

The vowel allophone characterisation mainly depends on the positioning of the phoneme. Naturally the most occurring allophones will be the ones corresponding to the middle positioning. Figure 4.3 shows the allophone based frequency of occurrence distribution for 5 consonant phonemes with maximum number of allophones.

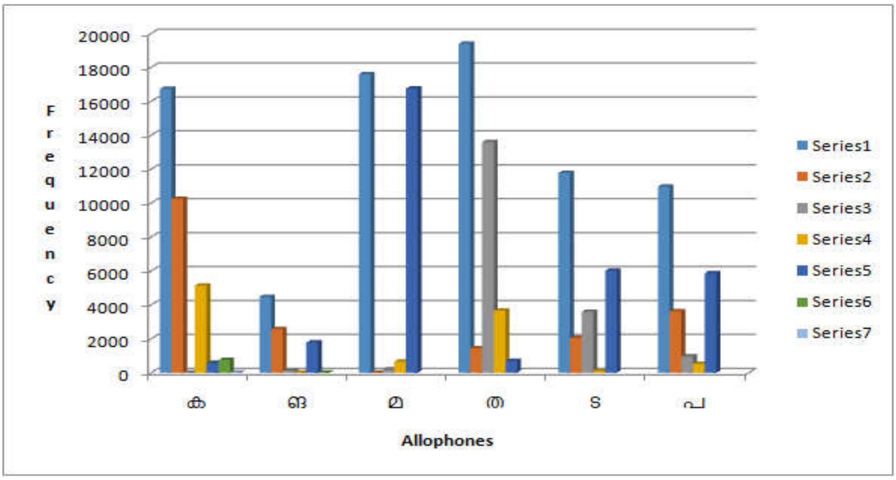


Figure 4. 3: Allophone based frequency of occurrence distribution for 5 consonant phonemes with maximum number of allophones

4. 5 Conclusion

This chapter explains the proposed implementation constituents of Malayalam grapheme to phoneme and allophone transcripters. The rule based approach gives satisfactory results in Malayalam transcription. Pre-processing brings different class of graphemes to a unified framework required for the phoneme and allophone transcription. A comprehensive statistical and probabilistic analysis based on the developed phoneme and allophone converter is performed on standard Malayalam word corpora. അ/a/ is the most frequently occurring phoneme and ത/t/ is the top among consonants. Phoneme

ഞ/ɲ / is the least frequently occurring phoneme in Malayalam. From the experiments it is evident that, the frequency of occurrence of allophones of the same phoneme varies significantly. The allophones positioned in the middle are found to be the most frequently occurring. The results of this analysis can be used for developing various language computing tools in Malayalam.

Chapter 5

Malayalam Viseme Set Formation based on linguistic knowledge, perception experiments and parametric approaches

5.1 Introduction

The characteristics of phoneme set, allophonic variations and the method of construction of transcripters in Malayalam are discussed in the previous chapters. The centre of discussion, so far has been on the auditory and textual domain representations of language. The human perception of speech is bimodal in nature and brain combines audio and visual cues through a complex process of cortical integration [240]. This chapter performs studies to understand the relevant characterisations of visual cues in Malayalam speech to be used for audiovisual speech synthesis and recognition applications. Identifying visually separable atomic units of Malayalam speech is one of the prime tasks realised in this work. The atomic unit of visual speech is termed as visual phoneme or viseme in the literature. Speech is produced through a complex process starting with phonemic conversion of utterance happening in the brain and its realisation by movements of articulatory organs such as jaw, lips, teeth, tongue, velum, larynx, nasal cavity and oral cavity. The relative positioning of visible articulators such as lip, tongue, teeth and jaw contributes to the characterisation of viseme set. The identification and synthesis of visually discerning phonemes in a language is fundamental in developing visual speech synthesisers. The Viseme set quantifies the capability of visual cues in bimodal speech recognition. This chapter explains the classification of phonemes based on visual cues and establishes a phoneme to viseme mapping for Malayalam. The aim of the work is to express speech as a sequence of viseme which has to be generated either from a sequence of phonemes or from a sequence of allophones.

It can be easily observed that all phonemes are not visually separable. Consider the example of four dental consonants ത്/t/,ഥ്/t^h/,ദ്/d/,ധ്/d^h/ in Malayalam, these consonants are formed by the active articulator tongue touching the invisible alveolar ridge inside the mouth. So it is not possible to visually discern these four consonants. The articulatory processes differentiating voiced and voiceless consonants are also not visible. Generally, for any language, we can identify group of phonemes with indistinguishable appearance on the visible articulators, which lead to a many to one phoneme - viseme mapping. The main intend of this study is to produce phoneme to viseme maps in a many to one fashion. But the visual manifestation of phonemes varies for different instances due to co-articulation effects. The visible articulator configuration of a phoneme is effected by the inertia of the articulatory organs and due to the anticipatory preparation for the future phonemes. The alteration due to impending phonemes is particularly evident in consonant vowel combinations such as പ്/p/+ഉ/u/. So the same phoneme exhibits context dependent differences in visual manifestations. Malayalam is one of the few languages in which allophone formations happening due to contextual and positional variability can be explained in a rigorous rule based approach. So the co-articulation effects can be modelled in Malayalam speech using an allophone centric approach.

Phoneme to viseme maps are developed from linguistic knowledge, perception experiments conducted using visual speech segments without audio and by clustering performed in the parametric space based on different visual features. Many to one phoneme to viseme maps are developed for Malayalam using linguistic knowledge, perception testing and data driven approaches. Geometric features of lips and Discrete Cosine Transform (DCT) based features are used for data driven clustering. The benefit of having an allophone set in Malayalam for modelling co articulation effects and to design a many to many phoneme to viseme map is exploited as part of this work. Allophone to viseme map is generated using data driven approach. Section 5.2 discusses various viseme set formation strategies. Section 5.3 explains the different phoneme to viseme mappings in Malayalam derived using linguistic, perception based and data driven approaches. Section 5.4 discusses the data driven allophone to viseme mapping performed to obtain a viseme set which incorporates co-articulation effect. Section5.5 concludes the chapter with future directions.

5.2 Viseme Set Formation Strategies

A viseme is the visual speech equivalent of a phoneme or the set of visually separable phonemes. Many researchers have analysed the importance of the phoneme to viseme mapping in the context of speech processing. The number and nature of viseme are language dependent. Hence a language specific exploration is needed for establishing the viseme set in a language. Among all the realms of language processing establishing the atomic unit for representing the language is considered as the first step towards automisation. The studies on Viseme set for visual speech representation and processing starts with the work of Woodward and Barber in 1960.They have proposed a hierarchy of visual contrasts in speech segments for lip reading [241]. The paper 'Hearing Lips and Seeing Voices' by Mcgurk *et al.* [242] is a land mark work in understanding the influence of visual modality in speech perception. Many approaches are used in literature for defining a viseme. Fisher in 1968 has coined the term viseme by concatenating visual and phoneme [87]. Homophone is an alternate term used in some works for representing the visually contrasting speech segments [243]. Auer *et al.* use the concept of Phonemic Equivalent Classes (PEC) to group visually similar phonemes[244].Saenko *et al.* conceptualise visemes as facial and oral articulatory gestures formed for uttering a phoneme[45]. Most approaches conceptualized viseme as a static mouth shape in 2D or 3D [245,188,246,247]. But recent practices rely more on using the realistic concept of dynamic Viseme where an image sequence is used instead of a static frame approximation [248]. The most popular approach is defining a viseme as a group of phonemes with similar facial and oral appearance. Practically the approach can be viewed as grouping of phonemes which are visually identical with regard to the mouth area, predominantly on the lips. The viseme set created as part of MPEG-4 facial animation framework [249] is shown in table 5.1.

Table 5.1: The Viseme set created as part of MPEG-4 facial animation framework

Viseme No.	Phonemes	Example	Viseme No.	Phonemes	Example
1	p, b, m	put, bed, mill	8	n, l	not, lot
2	f, v	far, voice	9	R	Red
3	T,D	think, that	10	A:	Car
4	t, d	tip, doll	11	E	Bed
5	k, g	call, gas	12	I	Tip
6	tS,dZ,S	chair,join, she	13	Q	Top
7	s, z	sir, zeal	14	U	Book

Basically there are three approaches for obtaining visemes from a many to one mapping:

i. Linguistic knowledge based approaches

ii. Perception experiments with human subjects

iii. Data-driven approach.

Some authors blend linguistic and perception experiments based approaches and name them as subjective assessments[250]. Section 5.2.1 discusses these three approaches used for viseme set formation in Malayalam.

Different phonemes can have similar visual mouth appearance. The phonemes which have almost same visual mouth appearance or common visible articulatory dynamics are grouped to a single viseme class. It can be observed from the literature that the size of the viseme set varies with in the range of 10 to 20. This shows that the number of viseme classes is highly language dependent.

5.2.1 Approaches to Viseme set Formation

Linguistic knowledge, perception experiments and data driven methods are used for the construction of Malayalam phoneme to viseme maps. The following section explains these three methods in detail.

i**. Linguistic Knowledge based Approach**

Viseme set is formed by exploiting the expert knowledge in the linguistics of the language. Phonetic properties, articulatory rules and visual intuition are used for classifying phonemes [251]. For English language Jeffers & Barley [90] mapped 43 phonemes into 11 visemes, which are shown in table 5.2. Another linguistic map prepared by Bozkurt *et al.* [81] consisting of 45 phonemes mapped to 15 visemes is shown in table 5.3.

Table 5.2: English viseme classes prepared by Jeffers & Barley constructed by mapping 43 phonemes into 11 visemes

Viseme Class	Phonemes Set	Viseme Class	Phonemes Set
/A	/f/ /v/	/G	/oy/ /ao/
/B	/er/ /ow/ /r/ /q/ /w/	/H	/s/ /z/
/C	/b/ /p/ /m/ /em/	/I	/aa/ /ae/ /ah/ /ay/ /eh/ /ey/ /ih/ /iy/ /y/ /ao/ /ax-h/ /ax/ /ix/
/D	/aw/	/J	/d/ /l/ /n/ /t/ /el/ /nx/ /en/ /dx/
/E	/dh/ /th/	/K	/g/ /k/ /ng/ /eng/
/F	/ch/ /jh/ /sh/ /zh/	/S	/sil

Table 5.3: English viseme classes prepared by Bozkurt *et al.* constructed by mapping 43 phonemes into 15 visemes

Viseme Class	Phonemes Set	Viseme Class	Phonemes Set
S	Silence	V9	/g/,/ hh/, /k/,/ ng/
V2	/ay/,/ ah/	V10	/R/
V3	/ey/, /eh/, /ae/	V11	/l/, /d/,/ n/,/ en/, /el/, /t/
V4	/Er/	V12	/s/, /z/
V5	/ix/, /iy/, /ih/, /ax/, /axr/,/y/	V13	/ch/, /sh/, /jh/, /zh/
V6	/uw/,/ uh/,/ w/	V14	/th/, /dh/
V7	/ao/, /aa/, /oy/, /ow/	V15	/f/, /v/
V8	/Aw/	V16	/m/,/ em/, /b/,/ p/

ii. Perception Experiments

The perception based approach focuses on conducting perception experiments on human subjects by matching visual speech to audio segments. The works using this approach generally create a confusion matrix as a first step for arriving at the viseme set. Confusion matrix helps to find the set of phonemes which are visually confused by the human subjects and such phonemes are mapped to form a viseme [252-253, 92]. The viseme set formed through such experiments can model the human perception of visual speech exactly. Hence the visual speech synthesis systems using this viseme set can produce convincing lip synching experience. The difficulty with this approach is the time and the manual labour required for the conduct of perception experiments. Creating confusion matrix based on HMM based automatic speech recognisers is an emerging alternative [254,255].

iii. Data Driven Approach

This approach is used for automatically learning the natural division among phonemes in the parametric space. Visual features are extracted from the mouth region of talking faces and viseme are formed by clustering it in the feature space [252]. Various visual features have been reported in literature. Based on the type of information embedded in the features, they are broadly classified into three groups: appearance-based features, geometry-based features and hybrid features [255]. For appearance-based approach, entire mouth region is assumed to be carrying relevant information. But it is computationally infeasible to use entire pixels in the Region of Interest (ROI) for clustering or similar processing. The usual practice is to use selected coefficients from any one of the transformed domains. Principle Component Analysis (PCA),Discrete Cosine Transform

(DCT), Discrete Wavelet Transform (DWT), Linear Discriminant Analysis (LDA),optical flow based features and Active Appearance Model (AAM) coefficients are the the widely used transforms used for viseme set formation [97,255,256,89]. Geometric visual features explicitly model speaker's lip contour and extracts geometrical features from it. Height and width and area of mouth, nose to chin distance, inner width and height of lip, area of inside mouth, dark region inside the mouth etc. are the geometrical visual features that are reported in the literature [96]. The visual speech recognisers also operate in one of the parametric spaces. Hence viseme set formed through data driven approaches are more relevant in visual speech recognition applications.

5.3 Malayalam Phoneme to Viseme Mapping

Malayalam language consists of 51 phonemes. The following section demonstrate the experimental studies conducted to perform phoneme to viseme mappings based on linguistic knowledge, perception experiments and data driven methods. The following section describes the pre-processing performed on the MAVSC – IP and MAVSC – IW images for various phoneme to viseme mapping experiments.

5.3.1 Visual Speech Data Set Preparation

The images in the MAVSC – IP and MAVSC – IW data set which are introduced in chapter 3 are used for various viseme set formation experiments explained in this chapter. The lip region of the subset of images from MAVSC – IP and MAVSC – IW data set is manually landmarked for feature extraction. To extract geometrical features the lip contour is needed to be accurately marked. The lip contour is obtained from the landmark points marked manually on each image. Lip contour is defined using 36 shape feature points on the lip. 20 points are used for marking the outer lip, while 16 land mark points defines the inner lip. The image with lip contour marked manually using 36 land mark points is shown in figure 5.1.

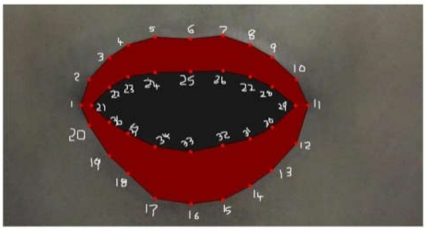


Figure 5.1. Manual labelling of landmark points in the inner and outer lip

The outer lip is represented by 20 points. Assuming an oval shape for lip, points 1,6,11 and 16 corresponds to 4 points, which meets two axes. These points are characterised by the presence of corners. Four almost equidistant points are marked between each pair(1,6),(6,11),(11,16) and (16,1). Similarly for inner lip points 21,25,29 and 33 are the land mark corner points. Three almost equidistant points are marked between each pair(21,25),(25,29) and (29,33).

Manual land marking is performed on a subset of the data set corresponding to 5 speakers from MAVSC-IP and MAVSC-IW datasets. Speakers selected from both data sets are made mutually exclusive to obtain manually processed data for maximum number of speakers. The image sequences from 10 speakers are used for clustering experiments in the parametric space.

5.3.2 Linguistic Knowledge based viseme mapping

Malayalam Phonemes are categorised based on the manner, organs of articulation and place of articulation. The visual speech appearance depends primarily on lip and lower jaw movements. Visibility of teeth and tongue is also a contributing factor. This section explores the possibilities of forming a viseme set from the linguistic knowledge about the language and its phoneme set. There are many phoneme to viseme maps reported for English, and two examples of the same are presented in table 5.2 and 5.3. But there are significant differences in the phoneme set of English and Malayalam. English language consists of 12 vowels, 8 diphthongs and 25 consonants. English phoneme set has two additional vowel phonemes compared to the 10 vowels in Malayalam, and 6 additional diphthongs compared to the 2 diphthongs in Malayalam. Malayalam language has 38 member consonant set, of which approximately 20 phonemes can only find a matching pair in English phoneme set. The viseme set obtained through Malayalam linguistic knowledge and comparison with existing phoneme to viseme linguistic map in English is explained bellow.

Initially, each vowel is assigned to a separate viseme class. But monophthongal short and long phonemes of the same vowel is placed in the same class, as phoneme duration is considered as the major difference between long and short vowels. The two diphthongs are assigned to separate viseme classes. The viseme set for vowels and diphthongs in Malayalam based n linguistic understanding, is given in table 5.4.

Table 5.4: Malayalam viseme set of vowels and diphthongs generated using linguistic knowledge

Viseme	Viseme class description	Phoneme set
Viseme 1	Front, High – Vowel	ഇ/i/ , ഈ/i:/
Viseme 2	Front, Mid – Vowel	എ/e/ , ഏ/e:/
Viseme 3	Central, Low – Vowel	അ/a/ , ആ/a:/
Viseme 4	Back, High – Vowel	ഉ/u/ , ഊ/u:/
Viseme 5	Back, Mid – Vowel	ഒ/o/ , ഓ/o:/
Viseme 6	Diphthong 1	ഐ/ai/
Viseme 7	Diphthong 2	ഔ/au/

The consonant viseme set is constructed based on the existing place of articulation based classification in the language. One of the major diffrence in malayalam consonant phonemes with its English counterpart is the presence of consonatns characterised as *athigharam* (voice less aspirated) and *ghosham* (voiced aspirated plosieves). In real life situations, voiced aspirated phonemes like ഭ്-/b^{h}/, ധ്-/d^{h}/, ഢ്/ɖ^{h}/, ഝ്-/ɟ^{h}/ and ഘ്-/g^{h}/ are often misinterpreted with the corresponding voiceless unaspirated phonemes and some voiceless aspirated phonemes (ഥ/ t^{h} / ,ഠ/ʈ^{h}/ , ഛ/c^{h} /, ഖ/ k^{h} /) are often confused with coresponding voiceless unaspirated phonemes. This observation is considered while assigning viseme class for voice less aspirated and voiced aspirated phonemes.

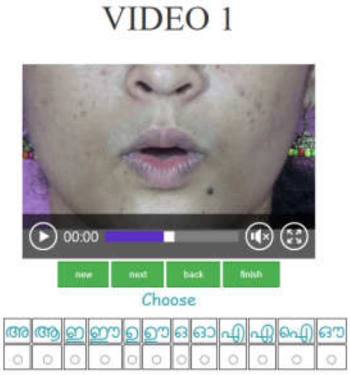
Viseme 8, the first consonant viseme class is formed from bilabial plosives expect ഫ്/p^h/ .വ-va, the only true labiodentals in the Malayalam and the bilabial - plosive-voiceless aspirated ഫ്/p^h/ are placed in the next viseme class with resect to the linguistic classifications given in table 5.1 and 5.2. Viseme 10 consists of dental consonants which have got the maximum teeth visibility. The velar consonants and the only glottal phoneme ഹ-ha are placed in the next class based on the linguistic classification strategies adopted in other languages. Viseme set 12, 13 and 14 are linguistically characterised as alveolar, retroflex and palatal consonants respectively in Malayalam. Due to the significant differences in the phoneme set in these categories a comparison with English viseme set is found irrelevant for these 3 classes of viseme. For example, many phonemes such as ര്/r/,റ്/ṛ/ and ഴ്/ʐ/ have no comparable counterparts in English. Many phonemes such as സ്/s/(to s as in sir) and ല്/l/ (to l as in lot)which are in the same linguistic class in Malayalam are treated differently in the construction of English viseme set. The Malayalam Viseme set corresponding to consonant phonemes generated using linguistic knowledge is given in table 5.5.

Table 5.5: Malayalam viseme set corresponding to consonants generated from linguistic knowledge

Viseme	Viseme class description	Phoneme set
Viseme 8	Bilabial - Plosive-voiced and voice less unaspirated, Nasal	പ്/p/ , മ്/m/, ബ്/b/ , ഭ്/b^h/
Viseme 9	Bilabial - Plosive-voiceless aspirated And Labiodental	ഫ്/p^h/, വ്/v/
Viseme 10	Dental	ത്/t/ , ഥ്/t^h/ , ദ്/d/ , ധ്/d^h/ , ന്/n̪/
Viseme 11	Velar Glottal	ക്/k/ , ഖ്/k^h/ , ഗ്/g/ , ഘ്/g^h/ , ങ്/ŋ/ ഹ്/h/
Viseme 12	Alveolar	റ്റ്/r̠/ , ന്/n/ , സ്/s/ , ര്/r/ , റ്/ṛ/ , ല്/l/
Viseme 13	Retroflex	ട്/ʈ/ , ഠ്/ʈʰ/ , ഡ്/ɖ/ , ഢ്/ɖʰ/ , ണ്/ɳ/ , ഷ്/ʂ/ , ള്/ɭ/,ഴ്/ʐ/
Viseme 14	Palatal	ച്/c/ , ഛ്/c^h/ , ജ്/ɟ/ , ഝ്/ɟh/ , ഞ്/ɲ/ , ശ്/ʃ / , യ്/y/

5.3.3 Perception Experiments Based Viseme Mapping

This is an attempt carried out to model the visual perception of Malayalam phonemes by native speakers. Viseme set is formed by identifying how individuals with normal hearing ability perceive phonemes [29]. The participants of this experiments include equal number of males and females with in the age group of 20 to 35 with normal hearing abilities. Visuals of isolated Malayalam utterances are shown to such participants without audio. The participants are requested to record the phoneme as perceived by them. A web based frame work is developed to record the responses of participants and to conduct subsequent analysis. The web interface of the developed framework is shown in figure 5.2. The results obtained with perception experiments conducted with a combined vowel and consonant phoneme set are found to be of poor performance. Hence perception experiments are conducted separately for vowels and consonants. The phoneme to viseme mapping is generated finally by computing the average behaviour of responses recorded by the participants.



അ	ആ	ഇ	ഈ	ഉ	ഊ	ഒ	ഓ	എ	ഏ	ഐ	ഔ
○	○	○	○	○	○	○	○	○	○	○	○

Figure 5.2: Interface of the web frame work developed for the conduct of perception experiments

The conclusions derived from the perception experiments are given bellow.

i. Out of the total 400 responses from 40 participants, vowels recorded a recognition accuracy of 78%.Most of the wrong identifications are between short and long phonemes of the same vowel. 95% recognition accuracy is recorded for the two diphones in Malayalam. So according to the perception experiments, each vowel phoneme and diphone can be considerd as a separate viseme

ii. The analysis cannot derive a reliable viseme mapping (as in the case of vowels) from perception experiments. A variant of confusion matrix based approach is used for finding the mapping[258]. Confusion matrix finds out the set of phonemes which are visually confused by the participants and such phonemes are grouped to form a viseme. Mutual confusion in this context is defined as phoneme 'i' assigned to phoneme 'j', and phoneme 'j' assigned phoneme 'i'. But due to the way native speakers of the Malayalam language learns the alphabet, the participants have a tendency of mapping to certain phonemes in a group. In a '*varggam*', the users show a tendency to choose the first phoneme in the 5 member group. So phonemes which are correctly identified by more than 75% of participants are assigned separate viseme class. Phonemes, which are confused to one of these viseme classes in more than 75% cases is assigned to the corresponding class. Some phonemes shows a totally fractured mapping and cannot be assigned to any specific class. The results of the perception experiments based phoneme to viseme map for Malayalam consonants is shown in table 5. 6.

Table 5.6: Perception experiment based phoneme to viseme map for Malayalam consonants

Viseme	Phoneme set
Viseme 1	ങ്/ŋ/
Viseme 2	ച്/c/ , ഛ്/c^{h}/ , ജ്/ɟ/
Viseme 3	ഞ്
Viseme 4	ത്/t/ , ഥ്/t^{h}/ , ദ്/d/ , ധ്/d^{h}/ , ന്/n̪/
Viseme 5	പ്/p/ , ബ്/b/
Viseme 6	മ്/m/,
Viseme 7	ഭ്/b^{h}/
Viseme 8	യ്/y/
Viseme 9	ഫ്/p^{h}/,
Viseme 10	വ്/v/
Viseme 11	ല്/l/,ര്/r/ ,ള്/ɭ/,ഺ്/r̠/,റ്/ṛ/
Unassigned Phonemes	ഝ്/ɟh/ , ശ്/ʃ / , ക്/k/ , ഖ്/k^{h}/ , ഗ്/g/ , ഘ്/g^{h}/ , ഹ്/h/, ന്/n/ , സ്/s/ , ട്/ʈ/ , ഠ്/ʈʰ/ , ഡ്/ɖ/ , ഢ്/ɖʰ/ , ണ്/ɳ/ , ഷ്/ʂ/ , ഴ്/ʐ/

From table 5.6 it can be seen that the perception experiment based approach fails to assign viseme classes to 16 phonemes in Malayalam which are listed in the last row of the table.

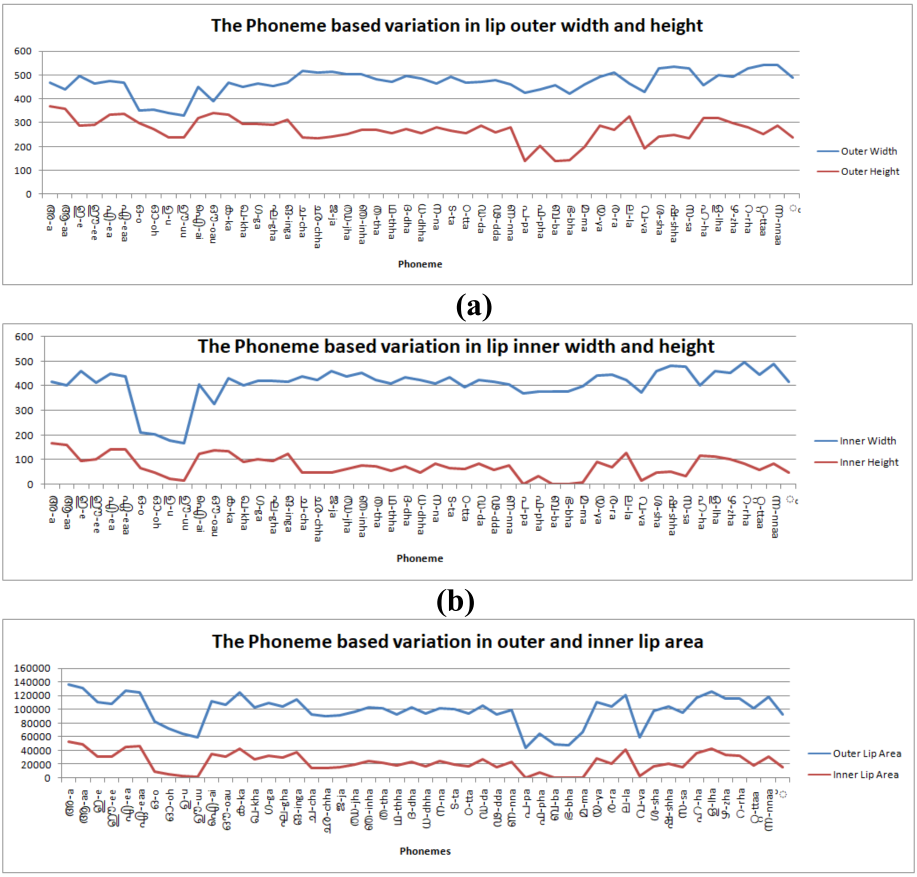
5.3.4 Data Driven Approach based Phoneme to Viseme Mapping

This section describes the attempts performed to compare the viseme set formed using linguistic and perception based approaches with viseme set based on the proposed data driven approach. In data driven approach visual features are extracted from the mouth region of talking faces and viseme are formed by clustering in the feature space. Both geometric features and appearance-based features are used as visual cues. The geometrical feature used in this study is computed as the distances from centroid to land mark points .The method used for labelling the 36 landmark points is already explained in section 5.3. Centroid is computed as the mean of 36 landmark points.

DCT of the mouth region is the appearance based feature used in this work. Hierarchical agglomerative clustering is used to find the viseme set by performing clustering in the feature space. The centroid based features are extracted from landmarked images. DCT is applied on a rectangular area enclosing the lip defined by the centroid. The following section presents a brief explanation of the visual features used for data driven approach based viseme set formation.

a. Centroid to Landmark Point Distance (CLPD) features

Centroid to Landmark Point Distance (CLPD) feature is the geometrical feature used in the study. In general the usual geometrical feature set consist of outer lip width,outer lip height,inner lip width,inner lip height, outer lip area and inner lip area. Figure 5.3 demonstrates the phoneme based variability of these geometric features.



(a)

(b)

(c)

Figure 5.3: The variation of geometrical features with respect to the 51 phonemes are represented along the x-axis and the y-axis shows the measurements in pixels (a). The phoneme based variation in lip outer width and height; (b). The phoneme based variation in lip inner width and height; (c). The phoneme based variation in outer and inner lip area.

The CPLD feature vector consist of 36 values, which are the Euclidian distances measured from centroid to the 36 land mark points characterising lip boundary. This feature, compared to conventional width and height information carries much more relevant information on the shape of the lip region during utterance. Euclidean distance between centroid and i^{th} cardinal point in each frame computed based on the equation 5.1

$$\text{Centroid Distance } (d_i) = \sqrt{(Xcentroid - Xcardinal_i)^2 + (Ycentroid - Ycardinal_i)^2} \quad (5.1)$$

Figure 5.4 presents a a box plot that depicts the variation found with respect to all 36 values in the CPLD feature vector used for visual representation of Malayalam phonemes. In the box plot central line indicates the median with edges that shows 25% and 75% of the distance variations. Whiskers show the extremes except some outliers which are represented by plus signs.

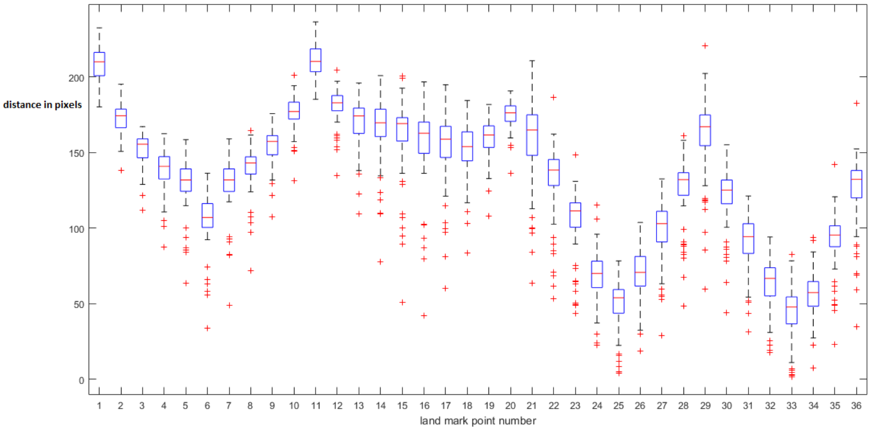


Figure 5.4: Box plot of landmark point number versus distance (in pixels) between centroid and the land mark points corresponding to Malayalam Phonemes.

b. DCT based Feature

Discrete Cosine Transform (DCT) is the most commonly accepted visual feature extraction method proposed by various researchers [259]. A two-dimensional DCT of an M-by-N image is represented as

$$D\ (i, j) = \sum_{i=1}^{M} \sum_{j=1}^{N} I(i,j) \cos\left(\frac{(2i+1)\pi i}{2M}\right) \cos\left(\frac{(2j+1)\pi J}{2N}\right) \qquad (5.2)$$

where I(i, j) is the grey-scale image of the Region Of Interest(ROI). The DCT returns a 2-dimensional matrix having M*N coefficients. Most of the visually significant information and hence energy is concentrated in a few coefficients of DCT, which represent the low frequency aspect of an image. To avoid the curse of dimensionality, these low frequency components are selected in a zig-zag manner starting from DCT component D (1, 1). In this study, the 25 DCT coefficients extracted from a rectangular area around the centroid are used as the visual feature.

5.3.5 Hierarchical Agglomerative Clustering (HAC) for Viseme set Generation

Viseme set is formed by performing clustering in the feature vector space. The four different feature vectors used for the clustering are CPLD features extracted from a single frame representation of a phoneme, CPLD features extracted from five frame representation of a phoneme , DCT based features extracted from a single frame representation of a phoneme, DCT based features extracted from five frame representation of a phoneme. Clustering experiments are also conducted using the combination of these features. Hierarchical agglomerative clustering [260], where initially each instance is treated as a separate cluster to iteratively form a single cluster accommodating all instances, is used for the conduct of clustering experiments. Here , single linkage clustering is used for merging clusters and Euclidian distance measure is used for computing the distance [261]. Single link distance between C_i and C_j is the minimum distance between any instances in C_i and C_j

$$dist(Ci, Cj)= min_{x,y}\{d(x, y), x \in Ci \wedge y \in Cj\} \tag{5.3}$$

The hierarchical agglomerative clustering algorithm used for the generation of viseme set is given bellow [260].

```
Hierarchical agglomerative clustering Algorithm

Input : Set of Instances $X_1$, X2.....$X_n$

Output : The set of clusters in C

Step1: Initialise C such that {C1, C2.....$C_n$} = { $X_1$,$X_2$.....$X_n$} // each instance in a separate    cluster

Step 2: While (size(C)>1) do

        Step 3 : (min1,min2) = $min_{i,j}\left(dist\left(C_i, C_j\right)\right)$ // min1 and min2 are the indexes of minimum distance cluster pairs and $dist(Ci, Cj)$ between cluster $C_i$ and $C_j$ computed as given in equation 5.3

        C = C – {$C_{min1,}$ $C_{min2}$}
```

$C = C \cup \{C_{min1}C_{min2}\}$

Step 4: Update the distance measures to reflect new clustering

Step 5: Repeat step 2 to 4

5.3.5.1 Experimental Results and Analysis

Clustering experiments are conducted using HAC algorithm with different features. Clustering experiments are conducted by varying the number of clusters from 10 to 20. When the number of clusters are set below 12, all the front vowels show a tendency to cluster in to a single group. When it is set above 14, the grouping deviates from the pattern followed in linguistic and perception based clustering. The viseme set arrived out of linguistic knowledge and perception based clustering as detailed in section 5.4.1 and section 5.4.2 are compared with viseme set generated based on the data driven approach. Clustering experiments are also conducted by combining DCT and CPLD features. The clustering with CLPD and DCT combined feature set extracted from 5 frames is found to be closest to the visme set formed form linguistic knowledge. The data driven viseme set with number of clusters fixed as 14 and 20 for combined CLPD and DCT features are given in table 5.7 and table 5.8 respectively.

Table 5.7: The Malayalam viseme set generated out of data driven approach by HAC using CPLD and DCT combined features,with number of clusters 14.

Viseme Number	**Phonemes**
Viseme-1	ഷ്-/ʂ/, റ്റ്-/r̠/
Viseme-2	ക്-/k/, ഖ്-/kh/, ഗ്-/g/, ഘ്-/gh/, ങ്-/ŋ/, ച്-/c/, ഛ്-/ch/, ജ്-/ɟ/, ഝ്-/ɟh/, ഞ്-/ɲ/, ത്-/t/, ഥ്-/th/, ദ്-/d/, ധ്-/dh/, ട്-/ʈ/, ഠ്-/ʈh/, ഡ്-/ɖ/, ഢ്-/ɖh/, ണ്-/ɳ/, ശ്-/ʃ /, സ്-/s/
Viseme-3	യ്-/y/, ര്-/r/, ല്-/l/, ഹ്-/h/, ള്-/ɭ/, ഴ്-/ʐ/, റ്-/ṛ/, ന്-/n/
Viseme-4	ന്-/n̠/
Viseme-5	ഫ്-/ph/
Viseme-6	വ്-/v/
Viseme-7	ഇ-/i/, ഈ-/i:/
Viseme-8	മ്-/m/
Viseme-9	പ്-/P/, ബ്-/b/, ഭ്-/bh/
Viseme-10	ഐ-/ai/
Viseme-11	അ-/a/, ആ-/a:/
Viseme-12	എ-/e/, ഏ-/e:/
Viseme-13	ഉ-/u/, ഊ-/u:/, ഒ-/o/, ഓ-/o:/
Viseme-14	ഔ-/au/

Table 5.8: The Malayalam viseme set generated out of data driven approach by HAC using CPLD and DCT combined features,with number of clusters 20.

Viseme Number	Phoneme
Viseme-1	ഇ-/i/
Viseme-2	ഈ-/i:/
Viseme-3	ഢ്-/ɖh/
Viseme-4	ഖ്-/kh/, ഗ്-/g/, ഘ്-/gh/, ങ്-/ŋ/, ച്-/c/, ഛ്-/ch/, ജ്-/ɟ/, ഝ്-/ɟh/, ഞ്-/ɲ/, ത്-/t/, ഥ്-/th/, ദ്-/d/, ധ്-/dh/, ട്-/ʈ/, ഠ്-/ʈh/, ഡ്-/ɖ/, ണ്-/ɳ/, ശ്-/ʃ /, സ്-/s/
Viseme-5	ഷ്-/ʂ/
Viseme-6	റ്റ്-/r̠/
Viseme-7	ഹ്-/h/
Viseme-8	യ്-/y/, ര്-/r/, ല്-/l/, ള്-/ɭ/, ഴ്-/ʐ/, റ്-/ṛ/, ന്-/n/
Viseme-9	ക്-/k/
Viseme-10	ഭ്-/bh/
Viseme-11	പ്-/P/, ബ്-/b/
Viseme-12	ന്-/n̪/
Viseme-13	ഫ്-/ph/
Viseme-14	വ്-/v/
Viseme-15	മ്-/m/
Viseme-16	ഐ-/ai/
Viseme-17	അ-/a/, ആ-/a:/
Viseme-18	എ-/e/, ഏ-/e:/
Viseme-19	ഉ-/u/, ഊ-/u:/, ഒ-/o/, ഓ-/o:/
Viseme-20	ഔ-/au/

Clustering is also performed separately for vowel and consonant phonemes with CPLD and DCT combined features. When the number of clusters is assigned value 7, the vowel phonemes found an exact match with viseme set formed based on the linguistic knowledge. The consonant alone viseme set (the number clusters is fixed as 10) using combined features(CPLD and DCT) is given in table 5.9.

Table 5.9: The Malayalam viseme set formed from consonant phonemes generated out of data driven approach by HAC using CPLD and DCT combined features,with number of clusters 10

Viseme	**Phonemes**
Viseme-1	ക്-/k/,ഖ്-/kh/, ഗ്-/g/, ഘ്-/gh/, ങ്-/ŋ/
Viseme-2	ച്-/c/, ഛ്-/ch/, ജ്-/ɟ/, ഝ്-/ɟh/, ഞ്-/ɲ/,
Viseme-3	ത്-/t/, ഥ്-/th/, ദ്-/d/, ധ്-/dh/, ട്-/ʈ/, ഠ്-/ʈh/, ഡ്-/ɖ/, ഢ്-/ɖh/, ണ്-/ɳ/, ശ്-/ʃ /, സ്-/s/
Viseme-4	ഭ്-/bh/
Viseme-5	പ്-/P/, ബ്-/b/
Viseme-6	ഷ്-/ʂ/, റ്റ്-/r̠/
Viseme-7	യ്-/y/, ര്-/r/, ല്-/l/, ഹ്-/h/, ള്-/ɭ/, ഴ്-/ʐ/, റ്-/ṛ/, ന്-/n/,ന്-/n̪/
Viseme-8	ഫ്-/ph/
Viseme-9	വ്-/v/
Viseme-10	മ്-/m/

The viseme sets obtained using linguistic knowledge (Table 5.5), perception experiments (Table 5.6) and clustering experiments are analysed in detail to frame a reliable phoneme to viseme map in Malayalam, which can be used for visual speech synthesis applications. The final Malayalam viseme set thus obtained is given in table 5.10. The vowel and diphone visme set obtained from linguistic and data driven approaches is exactly same. Most of the users participated in perception experiments can discern between long and short vowels of the same vowel phoneme. This fact has not been considered in arriving at the final vowel viseme set, as duration is observed to be the main factor which helped users to identify between long and short vowels.

Table 5.10: The final viseme set for vowels and diphthongs in Malayalam for the use of visual speech synthesis applications

Viseme	Viseme Class	Malayalam Phoneme
Viseme 1	Front, High – Vowel	ഇ/i/ , ഈ/i:/
Viseme 2	Front, Mid – Vowel	എ/e/ , ഏ/e:/
Viseme 3	Central, Low – Vowel	അ/a/ , ആ/a:/
Viseme 4	Back, High – Vowel	ഉ/u/ , ഊ/u:/
Viseme 5	Back, Mid – Vowel	ഒ/o/ , ഓ/o:/
Viseme 6	Diphthong 1	ഐ/ai/
Viseme 7	Diphthong 2	ഔ/au/

The final consonant viseme set is obtained by updating viseme set obtained using linguistic knowledge (Table 5.5) using relevant observations from viseme set obtained using other two approaches. The consonants മ്/m/, ഭ്/b^h/ , ഫ്/p^h/ and വ്/v/ are assigned to separate viseme classes, both in perception experiments and data driven clustering. Hence they are assigned separate viseme class in the final viseme set. This is a significant change from the viseme set formed from linguistic knowledge. No other significant change from linguistic knowledge based mapping shows a similar uniform behaviour in perception based and data driven approaches. Hence linguistic knowledge based mapping is adopted for remaining phonemes. The final consonant viseme set with 12 members is given in table 5.11.

Table 5.11: The final viseme set for consonants in Malayalam for the use of visual speech synthesis applications

Viseme Set	Viseme Class	Phoneme
Viseme 8	Bilabial - Plosive-voiced and voice less unaspirated, Nasal	പ്/p/ , ബ്/b/ ,
Viseme 9	Bilabial - Plosive-voiced and voice less unaspirated, Nasal	മ്/m/,
Viseme 10	Bilabial - Plosive-voiced and voice less unaspirated, Nasal	ഭ്/b^h/
Viseme 11	Bilabial - Plosive-voiceless aspirated And Labiodental	ഫ്/p^h/
Viseme 12	Bilabial - Plosive-voiceless aspirated and Labiodental	വ്/v/
Viseme 13	Dental	ത്/t/ , ഥ്/t^h/ , ദ്/d/ , ധ്/d^h/ , ന്/n̪/
Viseme 14	Velar Glottal	ക്/k/ , ഖ്/k^h/ , ഗ്/g/ , ഘ്/g^h/ , ങ്/ŋ/ ഹ്/h/
Viseme 15	Alveolar	റ്റ്/r̠/ , ന്/n/ , സ്/s/ , ര്/r/ , റ്/ṛ/ , ല്/l/
Viseme 16	Retroflex	ട്/ʈ/ , ഠ്/ʈh/ , ഡ്/ɖ/ , ഢ്/ɖh/ , ണ്/ɳ/ , ഷ്/ʂ/ , ള്/ɭ/, ഴ്/ʐ/
Viseme 17	Palatal	ച്/c/ , ഛ്/c^h/ , ജ്/ɟ/ , ഝ്/ɟh/ , ഞ്/ɲ/ , ശ്/ʃ / , യ്/y/

The next section describes the results of allophone to viseme mapping using data driven methods with combined feature set.

5.4 Allophone to Viseme Mapping

Defining the viseme set in a language is the principal foundation in visual speech synthesis and recognition applications. Malayalam viseme set generated by performing phoneme to viseme mapping using linguistic knowledge, perception based and data driven techniques are discussed in the previous sections. But the Phenomenon of visual co-articulation creates visible differences while uttering the same phoneme in different contexts. Both backward and forward co-articulation are observed in visual domain. The lip rounding for final phonemes for the word ‘boots’ is an example of backward visual co-articulation, while the lip rounding happening for the initial phonemes during the utterance of the word ‘school’ is an example of forward or anticipatory co-articulation. The existence of a phoneme in various visual representations is an established fact [250]. This contextual variability demands a many to many phoneme to Viseme map. Many attempts have been reported in literature to theoretically model visual co articulation effects [262-263,214]. Pichara *et al.* in a study using Vowel Consonant Vowel (VCV) utterances of hearing impaired and normal human subjects established visual co-articulation effects in identifying consonants [264]. A. P. Breen *et.al* used visual database of tri-viseme to model visual co-articulation for mouth shape generation[265].

The contextual phoneme variations in Malayalam are modelled using allophonic characterisations. The allophonic characterisations and allophonic transciptors of the language is already discussed in Chapter 3. Allophone based many to many phoneme to Viseme mapping is attempted for many languages. Many works considers an allophone set as a finer subdivision on phoneme set. Some consonant phonemes are further categorised as rounded consonants or widened consonants according to the vowel context [11-12]. There are methods of using machine learning approaches to compute the allophonic variations in the visual domain [266].Most of the contextual variability in Malayalam is accommodated in the allophone characterisation which are unique to the language. Hence a many to many phoneme to viseme mapping is also attempted in this work using the allophone set defined in chapter 3. In this study, a novel attempt is made to perform allophone to viseme mapping using data driven approach with combined DCT and CPLD feature set extracted from 5 frames for a phoneme. Hierarchical Agglomerative Clustering (HAC) is used to derive the mapping. Instead of fixing the number of viseme (as in the phoneme to viseme mapping), the number of visemes is decided based on the natural clustering in the feature space.

Dendogram analysis performed on the hierarchical clustering tree is the method used for understanding the clustering pattern in the parametric space. A hierarchical clustering tree is visualised using a dendogram, where the instances are represented as leaf nodes. Moving up from the leaf to root node the sequence of merges leading towards a single cluster has been traced. The intermediate nodes express the proximity between data instances. The height quantifies the distance between instances and formed clusters. Dendogram analysis can bring out the natural clustering within the data. The height difference between two successive layers can be taken as a clue for natural division. If the height difference between two layers is small it indicates the absence of natural division between joined sets at this level. If the height difference is significant between successive levels the scenario pinpoints a natural division. In a dendogram X – axis represents objects or clusters and Y- axis represents distance or dissimilarity between objects and clusters. The natural division is performed based on the computation of inconsistency coefficient, which quantifies the height difference between successive levels in a hierarchy. The value of inconsistence coefficient of a link in a tree measures the difference in height of the link with the average height of the links bellow it [261]. A high value of inconsistency coefficient indicates a natural division at that level. The natural clusters obtained after clustering using inconsistency coefficients for vowel allophones is shown in table 5.12 and table 5.13 shows similar results obtained for consonant allophones. Dendogram analysis is implemented using Matlab toolbox.

Table 5.12: Malayalam viseme set formed from the vowel allophones based on natural clustering with CPLD and DCT combined features

Viseme set	Allophones
Viseme-1	ഇ-ഇന്ന്-[in̠n̠ə], ഇ-വടി-[vaʈi], ഈ-ഈണം-[i:ɳam], ഈ-ചീര-[ci:ra]
Viseme-2	ഇ- ചിത്രം -[citram], ഉ-അത്-[atə], ഉ-കൃഷി-[kr̩ɯvʂi]
Viseme-3	എ- പിന്നെ-[pinne]
Viseme-4	അ-അല-[ala], ആ- നാളെ-[na:ɭe], എ- എവിടെ-[eviʈe]
Viseme-5	ഏ- വേണം -[ve:ɳam]
Viseme-6	ആ-കാലം-[ka:lam]
Viseme-7	ഉ-പാല്-[pa:l]
Viseme-8	ഉ- ഉച്ച-[ucca], ഉ- നിന്നു-[nin̠n̠u], ഉ- കറുപ്പ്-[kaṛuppə], ഉ-കുട്ടി -[kuʈʈi], ഊ - ഊമ-[u:ma], ഊ- കൂമൻ-[ku:man], ഒ-ഒരുമ-[oɾuma], ഒ- കൊടി -[koʈi], ഓ- ഓരത്ത് [o:ṛttə], ഓ-പോയി-[po:ji], ഔ-ഗൗരവം-[gauravam]
Viseme-9	ഏ-ഏട്ടൻ-[e:ʈʈan], ഏ-പുറകേ-[puṛake:], ഐ-തൈ-[tai], ഐ-നെയ്-[daja]
Viseme-10	അ-വരച്ചു-[varaccu], എ-വെളുത്ത-[veɭutta]

From table 5.12, it can be seen that 10 viseme classes are formed from the vowel allophones after natural clustering using combined CPLD and DCT features. It can also be observed that allophones of same phoneme are present in different clusters for many vowels.

Table 5.13: Malayalam viseme set formed from the consonant allophones based on natural clustering with CPLD and DCT combined features

Viseme Set	Allophones
Viseme-1	ഖ്-രേഖ-[re:Kha], ങ് – മാങ്ങ-[ma:ŋŋa], ങ്–പെങ്ങൾ-[peŋŋal, ഠ്-കഠിനം-[kaʈʰinam]
Viseme-2	ശ്-ശേഷം-[ʃe:ʂəm]
Viseme-3	ത്-ചന്ത-[canta]
Viseme-4	ന്-അവന്-[avən]
Viseme-5	ട്-വണ്ടി-[vaɳʈi]
Viseme-6	ക്-തർക്കം-[taṛkkam]
Viseme-7	ന്-ചിഹ്നം-[cihnam]
Viseme-8	ക്-കച്ചവടം-[kaccavaʈam], ഖ്-ഖേദം-[Khe:dam], ച്-ചായ-[ca:ja], ഛ്-ഛായ-[cʰa:ja], ജ്-ജാതി-[ɟa:ti], ഝ്-ഝഷം-[ɟʰaʂəm], ഞ്-ഞങ്ങൾ -[ɲaɲɲaɭ], ട്-കുട-[kuʈa], ത്-തിര-[tiɾa], ദ്-മുദ്ര-[mudɾa], ധ്-ധാരാളം-[dha:ra:ɭəm]
Viseme-9	ഠ്-കുഷ്ഠം-[kuʂTʰam]
Viseme-10	ഡ്-പണ്ഡിതന്-[pʌɳɖiðan]
Viseme-11	ക്-എങ്കില്-[eŋgil]
Viseme-12	ത്-നെയ്തു-[nejtu]
Viseme-13	ഥ്-അർത്ഥം-[aṛtham]
Viseme-14	ര്-രാത്രി-[ɾa:tṛi]
Viseme-15	ട്-വട്ടം-[vaʈʈam]
Viseme-16	ദ്-ദയ-[daja]
Viseme-17	യ്-യാത്ര-[ja:tṛa]
Viseme-18	ഘ്-ആഘോഷം-[a:gho:ʂam], ന്-ചേർന്നു-[ce:ṛnnu]
Viseme-19	ക്-ക്ഷീണം-[kʂi:ɳam], ഗ്-ഭംഗി-[bhaŋgi]
Viseme-20	ഹ്-സഹായം-[saha:jam]
Viseme-21	ട്-ടിപ്പു-[ʈippu]
Viseme-22	ഖ്-മുഖം-[mukham]
Viseme-23	ന്-നല്ല-[n̪alla]

Viseme-24	ക്-ചിരിക്കു-[ciɾikkju:],
Viseme-25	ണ്-പണം-[paɳam]
Viseme-26	ങ്-തേങ്ങ-[te:ŋŋa]
Viseme-27	ഹ്-നമഃ-[namaH]
Viseme-28	ഗ്-ഗാനം-[ga:nam], റ്-പ്രശ്നം-[pṛaʃnam]
Viseme-29	ല്-ലോകം-[lo:kam], വ്-സർവം-[saṛvam
Viseme-30	വ്-ജ്വാല-[ɟwa:la]
Viseme-31	ഢ്-ഗൂഢം-[gu:ɖʰʌm]
Viseme-32	മ്-സംവരണം-[samwaɾaɳam]
Viseme-33	ക്-പകല്-[paɣal], ങ്-പടവലങ്ങ-[paʈavalaŋŋa], ച്-വാചകം-[va:ç̬akam], ച്-സഞ്ചി-[saɲɟi], ച്-വളർച്ച-[vaɭaṛCa], ഛ്-അച്ഛന്-[aCʰacn], ജ്-രാജ്യം-[ɾa:ɟjam], ത്-സത്യം-[satjam], ഷ്-ഭാഷ-[bha:ʂa], സ്-മനസ്സ്-[manassə], ള്-മകള്-[maɣaɭ], ഴ്-മഴ-[maỵa], റ്റ്-എന്റെ-[enr̲e], റ്റ-തെറ്റ്-[ter̲r̲ə]
Viseme-34	ങ്-പൊങ്ങും-[po:ŋŋum], പ്-പകുതി-[pakuti], പ്-ഉപകാരം-[uβaka:ram], പ്-തുമ്പ-[tumpa], പ്-തീർപ്പ്-[ti:ṛppə], ഫ്-ഫലം-[phaləm], ബ്-അർബുദം-[aṛbudam], ബ്-ബാക്കി-[ba:kki], ഭ്-ഭാര്യ-[bha:ɾja], മ്-ബ്രാഹ്മണൻ-[bɾa:mhaɳan], മ്-ഓർമ്മ-[o:ṛma], മ്-മരം-[maɾam]

The Malayalam consonant allophones are mapped to 34 viseme classes after performing natural clustering. Most of the allophones of the labial consonants are accommodated in a single viseme class. This is in contrast to the clustering behaviour of isolated phoneme utterances of labial consonants. As in the case of vowels, the allophones of many consonant phonemes are mapped to different viseme classes. This shows the relevance of allophone based mapping in the visual speech processing of Malayalam language.

5.5 Conclusion

This chapter derived the process of development of Malayalam viseme sets using different approaches. Phoneme to Viseme maps are developed based on linguistic knowledge, perception experiments and data driven clustering methods. Centroid to land mark points distances based geometric features of lips and Discrete Cosine Transform(DCT) coefficients of lip area are the visual features used for clustering. The final phoneme to viseme map, with 17 members is formed by comparing viseme sets obtained using three approaches. More research in all the three methods have potential scope for improvement in the visme set. Many to many phoneme to viseme maps are generated using data driven approach by exploiting the well-defined allophone set in Malayalam. Designing perception experiments for allophone to viseme mapping is a possible future work in the area.

Chapter 6

Mouth region Analysis and Lip Segmentation from Talking Frontal Face Images

6.1 Introduction

Facial feature extraction and analysis is an active research area with potential applications in computer vision, visual speech recognition, data driven animation and automatic sign language processing. Mouth motion accounts for the prominent non-rigid facial motion, especially during talking. Synthesis of head, eyes and mouth movements in a realistic and expressive manner is the most appealing aspect of animated characters and virtual agents. These movements reflect the internal state of characters, which decides the level of engagement with the audience. Earlier approaches used markers and sensors for tracking target regions, which is retargeted to a synthesis framework. Automatic video analysis frameworks are employed for face synthesis in many applications [166, 267, 268]. In these models the targeted region is extracted and its features are rendered to 2-d or 3- d animated models. In their work, Hadi Seyedarabi *et al.* track facial characteristic points to generate facial emotions [269]. Systems with low cost sensors for facial tracking and subsequent retargeting to animated figures are also coming up [270, 271]. Development of personalized character animations, mimicking the minute features of individuals with real time tracking techniques is an emerging area of application[270]. An analysis synthesis framework for facial motion animation, especially lip motion synthesis for speech animation is an emerging research topic [272,166,267,273]. Attempts are also made to track facial expression for imparting expressions to animation characters [269].

Nowadays Automatic Speech Recognition (ASR) has changed its mode of operation from audio-only speech recognition System to Audio-Visual Speech Recognition System (AV-ASR). The performance of Audio-only ASR degrades drastically in the noisy environment [274-275]. In the light of this fact researchers in this field have incorporated the visual information with its audio counterpart which improves the robustness by providing complementary information [276-277]. Human's major visual speech information is provided by the lower part of the face, especially the mouth region. The reliability of a visual speech recognition system deeply depends on the accurate tracking of mouth. The statistical modeling of mouth region pixels is important both for visual speech analysis and speech synthesis. Such an analysis is not performed adequately in the context of the color tone of Indian subcontinent. This chapter discusses detailed statistical and probabilistic analysis of the pixel intensities in the mouth region comprising of skin, lip, teeth, tongue and inner mouth cavity. The analysis is performed in RGB, Normalized RGB, HSV, CIE La*b*, and YCbCr colour models. A 5 class Bayesian classifier is designed to test the distinguishing power of mouth region pixels in different colour spaces. The 15 color components are ranked based on the performance of Bayesian classifier on each colour component.

Results of ranking based on overall classification performance and region wise classification performance are reported separately.

Lips are the most dynamic visible articualtory organs of human speech productions system. Lip based features are used most frequently for visual speech recognition applications. The idea used in this work for lip segmentation is to represent inner and outer lip contours using 36 shape feature points and train the system to find these landmark points. Active Shape Model (ASM) and Convolution Neural Network (CNNs) are used to perform lip segmentation using this approach. Training and testing are executed on MAVSC- IP and MAVSC-IW data sets. ASM is usually performed on the gray scale images. In this work ASM training and searching are carried separately for top performing color components identified in the Bayesian classifier based study. H of HSI color model is found to be performing best in finding the landmark points. The segmented images using ASM in a single speaker mode are used for training neural networks and for training the visual speech synthesizer.

Deep neural networks are widely used for solving computer vision problems such as object tracking [278]. But the concept is rarely used for tracking lips from frontal face talking images. In this work a CNN is designed to segment the inner and outer lip contour using land mark localization technique. The proposed CNN architecture with 6 convolution layers is finalized after experimenting with different network parameters. Rest of the chapter is organized as follows. After the introduction in section 6.1, section 6.2 discusses the mouth region pixel analysis and multiclass Bayesian classifier. Section 6.3 explains

ASM based lip land mark localization with experimental results. Section 6.4 discusses the theoretical aspects of CNN and its adaptation for lip tracking with experimental results. Section 6.5 concludes the work with future research directions.

6.2 Mouth region analysis in different Colour Spaces

Mouth area mainly consists of lip, skin neighborhood of lips, teeth, tongue and dark portions of mouth cavity. The range of human skin colour varies from white to black with discernable complexions including yellow, copper coloured and olive coloured [279]. The difference is primarily due to melanin content which varies greatly with ethnicities. The lip colour harmonizes with the background skin colour making lip segmentation using colour information quite difficult [280]. There are studies exploring the correlation of teeth colour with skin colour for various ethnicities [281]. So, while performing studies on colour based segmentation of skin or lip, researchers ensure the presence of images representing different ethnicities [282, 283, 284]. But the images selected are mostly of African, Mongolian and Anglo-Saxon ethnic groups. The wide ethnic and skin colour variability in Indian subcontinent is not addressed properly. This work performs the statistical and probabilistic analysis of mouth region pixel colors in Indian context for different color spaces. The results of analysis are incorporated to a multiclass Bayesian classifier for ranking the color components according to its effectiveness in segmenting different mouth regions. The conclusions derived from this work can be used for developing effective visual speech synthesis and recognition systems which incorporate the colour ethnic peculiarities and conditions of

Indian subcontinent. Even though lip segmentation and tracking have been studied deeply, only very few works have addressed the problem of segmenting different mouth regions including teeth and tongue. Tongue segmentation is addressed recently in some works, especially for applications related to traditional Chinese medicine [129-133]. Lip segmentation attempts can be broadly classified into two classes, namely Model based approach and Colour based approach [134]. In model based approach, mathematical models of lip contour are used as a set of model parameters for lip segmentation. Active shape and appearance model, snake model and deformable templates are widely used methods in this category [159,163]. In colour based approach, the colour triplet values of skin and lip pixels are used as the basic information for segmentation [291-293,144,294].

Lip, tongue, teeth and skin colour of human beings greatly vary with ethnicity. In Indian context the contrast between lip and skin colour is found to be low compared to other ethnicities. Statistical analysis of mouth region pixel colours in Indian context is an urgent requirement for developing native tools in the domain of visual speech recognition and synthesis. This work is an attempt towards this direction. Colour can be represented in different colour spaces such as RGB, normalized RGB, HSI, CIE Lab and YCbCr. The discriminating power among mouth regions is different for different colour spaces. Selection of optimum colour space which performs segmentation process effectively is still being actively debated among researchers. The optimum colour space is also varies depending on the colour properties of the population or ethnicity. The statistical analysis of pixels in skin and lip region of peoples in Indian subcontinent is performed in 5 different colour spaces. From the detailed review reported in chapter 2 on skin, lip and other mouth area segmentation, it is evident that there are 5 colour spaces which are most frequently used and reported to be most effective for the purpose. The following section describes the properties ofthese 5 colour spaces including RGB, Normalized RGB, HSV, CIE La*b*, and YCbCr which are used for this study.

6.2.1 Properties of Different Colour Spaces

A colour space is a mathematical representation of a set of colours. Most of the colour spaces are derived from the RGB colour space. Different colour spaces are suitable for different image processing applications. RGB, Normalized RGB, HSV, CIE La*b*, and YCbCr are the five colour spaces considered in this study.

a) RGB Colour Space

RGB colour space is the simplest and most widely used method in computer graphics. Luminance of a given RGB pixel is a linear combination of the R, G and B values. High correlation between channels, significant perceptual non-uniformity and mixing of chrominance and luminance data make RGB a bad choice for colour analysis and colour based recognition algorithms

b) *Normalised RGB Colour Space*

Normalised RGB is invariant to changes of surface orientation relative to the light source [295]. The normalised RGB values are represented as r, g, and b .These values are obtained by a normalisation procedure as shown bellow.

$$r = \frac{R}{R+G+B,}, g = \frac{G}{R+G+B}, b = \frac{B}{R+G+B} \qquad (6.1)$$

c) *HSV Colour Space*

Hue, Saturation and Value (HSV) colour space is a non-linear transformation of RGB colour space into a cylindrical coordinate representation. Hue is the wavelength at which the enegy is maximum or the colour of the pixel is most prominent. Saturation is the slope of the bandwidth curve around the central maximum[296]. This class of colour model is the most intuitive or artistic way of describing a colour. This model is closest to the way humans percvieve colour. Hence it is the most widely used model in computer vision applications. Jim *et al.* gives detailed algorithm for converting from RGB to HSV [297].

d) *CIE L*a*b* Colour Space*

In La*b* the ‘L’ channel represents the human perception of luminosity and the ‘a*’ and ‘b*’ components provide colour information. The range of L channel varies between black and white and the channel ‘a*’ varies from red to green similarly channel ‘b*’ varies from blue to yellow. It encodes the perceptual difference in colours when viewed by humans [298].

e) YCbCr Colour Space

In YCbCr, the gray scale information is carried by the 'Y' component and colour information is provided by two colour channels, *Cb*and *Cr*. Cb is obtained as the difference between blue component and a reference value and Cr is obtained as the difference between red component and a reference value. The Y value is obtained as a weighted sum of R, G, and B triplet [298].

6.2.2 Visual Speech Database Pre-Processing

The lip landmarked images taken from the MAVSC- IP and MAVSC-IW are used for analysis and classification experiments.The colour information from around one billion pixels belonging to face images of 10 speakers is used. The teeth and tongue regions inside the inner lip portion are segmented using a semi-automatic thresholding technique. The threshold for teeth and tongue are fixed separately for each speaker, through user intervention. A mask is created for each image in the data base with separate labels for skin, lip, tongue, teeth and inside mouth dark region (with corresponding labels 0,1,2,3and 4 respectively). Background pixels are eliminated by selecting a rectangular window based on the centroid of land mark points. The cropped images and the mask with labels are employed for performing statistical analysis of different regions. Figure 6.1 shows the output images and mask after preprocessing for two different images in the database. The images in the second column display the image mask with separate labels for skin, lip, tongue,teeth and inside mouth dark regions.

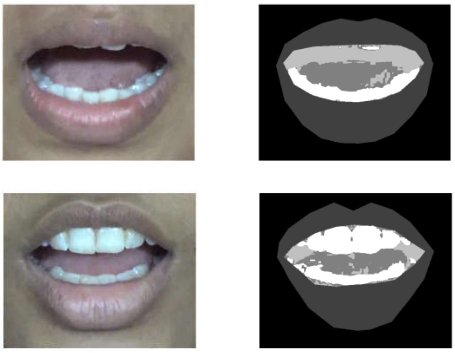


Figure 6.1: Sample images and corresponding mask image after pre-processing

6.2.3. Statistical Analysis of Colour Properties of Pixels of Mouth Regions

This section consolidates the statistical analysis performed on the images for understanding the colour properties of different mouth regions. Better understanding of the colour properties will help the development of mouth region segmentation tools adapted to the colour peculiarities of Indian subcontinent. The colour information can also be used for developing native speech animation applications. The properties of 15 colour components(from five colour spaces) in five different mouth regions is performed on the images from MAVSC- IP and MAVSC-IW datasets. Statistical properties including mean, standard deviation, class conditional probabilities and prior probabilities are estimated for pixels in each region. Other than skin, lip, tongue and teeth a separate class is employed to dark pixel region inside the mouth cavity. The statistics is obtained after analysing pixels in the mouth region images captured from 10 speakers. Table

6.1shows the mean and standard deviation of colour components belonging to 5 colour spaces obtained from 5 different mouth regions.

Table 6.1: Mean and standard deviation of mouth regions in 15 colour components from 5 different colour spaces

Colour Component	Mean					Standard Deviation				
	Skin	Lip	Teeth	Tongue	Dark	Skin	lip	Teeth	Tongue	Dark
	HSV									
H	0.116	0.682	0.564	0.770	0.865	0.370	0.353	0.045	0.079	0.095
S	0.131	0.134	0.137	0.125	0.238	0.024	0.041	0.043	0.031	0.088
V	0.521	0.513	0.737	0.522	0.289	0.054	0.088	0.136	0.106	0.110
La*b*										
L	53	49.48	71.01	50.52	25.86	5	7.68	12.74	10	11.63
a*	128.75	135.3	123.3	135.5	135.7	4.01	3.44	4.39	2.58	2.85
b*	134.29	127.8	120.6	121.6	125.5	3.88	2.47	2.94	2.32	2.47
RGB										
R	133	130.6	157.6	129.1	72.6	13.8	22.6	25.5	25.85	28.91
G	125.8	113.8	177.9	117.04	58.4	12.8	18.4	36.13	25.04	26.03
B	115.64	118.2	187.6	131.8	66.09	12.11	20.12	34.73	27.63	26.45
Normalised RGB										
R	0.3556	0.360	0.302	0.342	0.374	0.006	0.011	0.010	0.008	0.026
G	0.359	0.314	0.338	0.309	0.289	0.013	0.008	0.009	0.008	0.019
B	0.3088	0.325	0.359	0.348	0.336	0.010	0.008	0.009	0.008	0.019
YCbCr										
Y	108.93	102.5	148.5	105.0	54.59	10.77	16.80	28.32	21.86	22.96
Cb	122.45	127.3	135.1	132.4	129.1	3.01	1.98	2.81	2.11	1.55
Cr	131.85	134.9	118.4	132.0	133.5	2.26	3.19	5.05	2.17	2.67

RGB and YCbCr are defined in the 0- 255 range, while HSV and nRGB are defined in the range 0 - 1. In La*b* colour space L is defined over the range 0 - 100 and a* and b* is defined in the range 0 – 255. The dynamic range of mouth region pixels is below 50% of the possible range for all components in the RGB space. But the hue component of HSV space encompasses more than 75% of the possible range which is the maximum of all colour components. The class conditional probabilities and prior probabilities of different mouth region pixels are also estimated for 15 colour components. The estimated values are used in naive Bayesian classification discussed in the section 6.2.4. The class conditional probability distribution of different mouth regions for the Hue component in the HSV colour space and Cr component in the YCbCr is shown in figure 6.2 and figure 6.3 respectively.

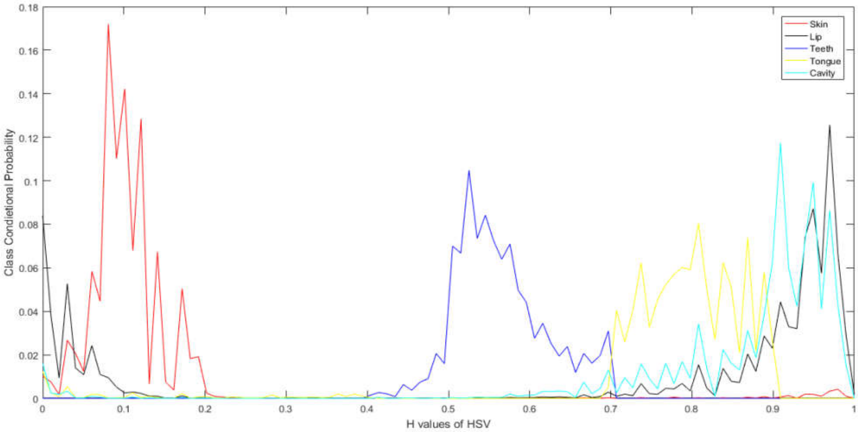


Figure 6.2: Class conditional probability of H Component of HSV

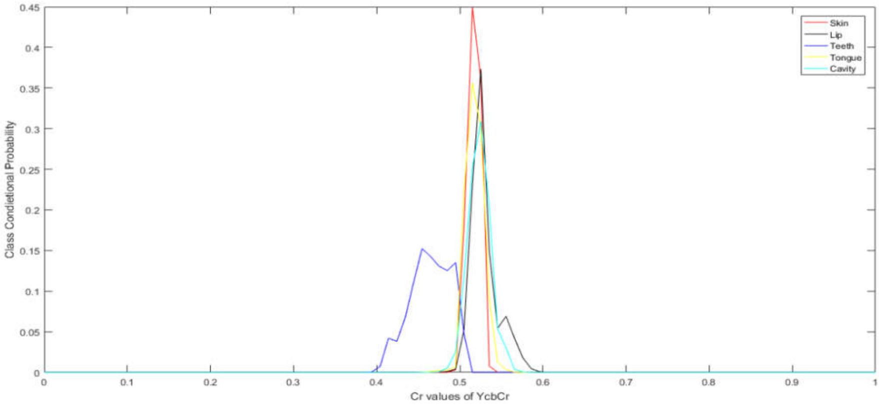


Figure 6.3: Class conditional probability of Cr Component of YCbCr(Normalised in the 0 – 1 range)

In the distribution of H component, teeth representing pixels have a distinct domain represented by blue colour in the figure 6.2, while lip pixels have colour presence in both low and high values of the spectrum. From the figure it is clear that, there is considerable overlap in the teeth, tongue and mouth cavity pixels in this colour space. The distinctiveness of teeth pixels is repeated in Cr values also, but the dynamic range of Cr values is comparatively small. In order to rank the effectiveness of colour components for practical purposes such as segmentation more objective evaluations are needed. The following section describes the performance evaluation of different colour components for mouth region classification implemented using naive Bayesian approach.

6.2.4 Mouth region Segmentation based on Bayesian Classifier

A 5 class Bayesian classifier is designed to test the distinguishing power of mouth region pixels in different colour spaces. Skin, lip, tongue, teeth and dark pixel region inside mouth cavity are the 5 visually different areas in the mouth region. The class conditional probabilities and prior probabilities computed from the MAVSC-IP and MAVSC-IW training dataset are used for the implementation of a multiclass Bayesian classifier for mouth region pixel classification. The detailed description of the proposed Bayesian classifier is given in following section.

Naive Bayesian Classifier

Bayesian classification has a training phase which computes the class conditional probability or likelyhood for each class and the prior probability for each class. Bayesian classification has been employed for solving skin and lip segmentation problems [115,299,300,301]. The idea is to divide the colour space in to histogram bins, where each bin stores the count corresponding to the occurrence of that colour in the training database. The counts are converted into class conditional probability measures. In this work, Bayesian classification is attempted to label each pixel either as skin, lip, teeth, tongue or dark region pixel inside the mouth cavity. Let w_i's be the set of classes (in this case 5 mouth regions) and x is the colour value of the current pixel, Bayes' theorem calculates the posterior probability$P\left({}^{w_i}/_{x}\right)$, the probability of observing the i^{th} mouth region, given a colour value x :

$$P\left({}^{w_i}/_{x}\right) = \frac{P\left({}^{x}/_{w_i}\right).P(w_i)}{P(x)} \qquad (6.2)$$

$P\left({}^{x}/_{w_i}\right)$ is the class conditional probability or likely hood,$P(w_i)$ is the prior probability for each class (both calculated from the training data) . As p(x) is the same for all classes, it is neglected and classification is based only on the value in the numerator of equation (6.2) [302]. A pixel with colour value x is assigned a label (representing the class membership) computed using equation (6.3)

$$label = Max_{i=0}^{4}\left(P\left({}^{x}/_{w_i}\right).P(w_i)\right) \qquad (6.3)$$

In the testing phase a mask is created corresponding to each image, where each pixel is assigned a label based on its identified class. Figure 6.4 shows the image and the mask obtained after Bayesian classification on an image which is not in the training set.

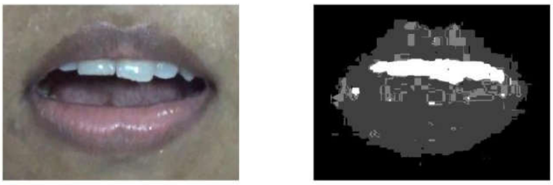


Figure 6.4: Original image and image mask created using Multiclass Beyesian Classifier in b* Colour Component of La*b* space

.

The mask thus obtained for 15 colour components can be compared with the ground truth mask prepared as part of pre-processing. Region wise and overall classification accuracies are computed by comparing the mask obtained from Bayesian classification and ground truth mask.

*Performance Evaluation*of Mouth Region Pixel Classification

The selection of colour space is crucial for discriminating different mouth regions. The objective of this work is to identify the optimal colour component for classifying mouth regions pixels in the context of skin tone of the Indian subcontinent. The performance of the multiclass segmentation in different colour spaces is compared. The segmented image mask obtained using Bayesian classification is used for comparison. The accuracy of classification is computed separately for each region by comparing with the ground truth mask. The 15 colour components are arranged according to their region wise performance in table 6.2. From the table it can be seen that, the overall performance and tongue segmentation performance is best for hue component in the HSV colour space. Lab –a for skin, nRGB – G for lip, YCbCr – Cr for teeth and RGB- R for dark regions inside mouth cavity are other toppers in the list. The classification 7 accuracy is found to be lowest for tongue segmentation. Figure 6.5 is the graphical representation of overall segmentation accuracy of different colour components.

Table 6.2:The 15 colour components are arranged according to their region wise classification performance

Over all	Skin	Lip	Teeth	Dark	Tongue
HSV – H	Lab – a	nRGB – G	HSV-H	RGB – R	HSV – H
Lab – b	RGB – G	nRGB – B	YCbCr - Cr	YCbCr – Y	Lab – b
Lab – a	YCbCr – Y	YCbCr – Cb	nRGB - R	HSV – S	Lab – a
YCbCr – Cb	Lab – L	Lab – b	nRGB - B	Lab – L	RGB – G
RGB – G	Lab – b	HSV – H	YCbCr - Cb	RGB – G	YCbCr – Cb
nRGB – G	RGB – R	Lab – a	Lab – b	RGB – B	nRGB – G
YCbCr – Y	HSV – H	RGB – G	RGB – B	nRGB – G	YCbCr – Y
Lab – L	HSV – S	YCbCr – Cr	Lab – a	nRGB – B	Lab – L
nRGB – B	YCbCr – Cr	nRGB – R	RGB – G	HSV – V	YCbCr – Cr
YCbCr – Cr	RGB – B	YCbCr – Y	Lab – L	nRGB – R	nRGB – B
RGB – R	nRGB - G	Lab – L	YCbCr - Y	HSV – H	RGB – R
RGB – B	YCbCr - Cb	RGB – R	RGB – R	YCbCr – Cb	RGB – B
HSV – S	nRGB - B	RGB – B	nRGB - G	YCbCr – Cr	nRGB – R
nRGB – R	nRGB - R	HSV – V	HSV – S	Lab – a	HSV – S
HSV – V	HSV – V	HSV – S	HSV – V	Lab – b	HSV – V

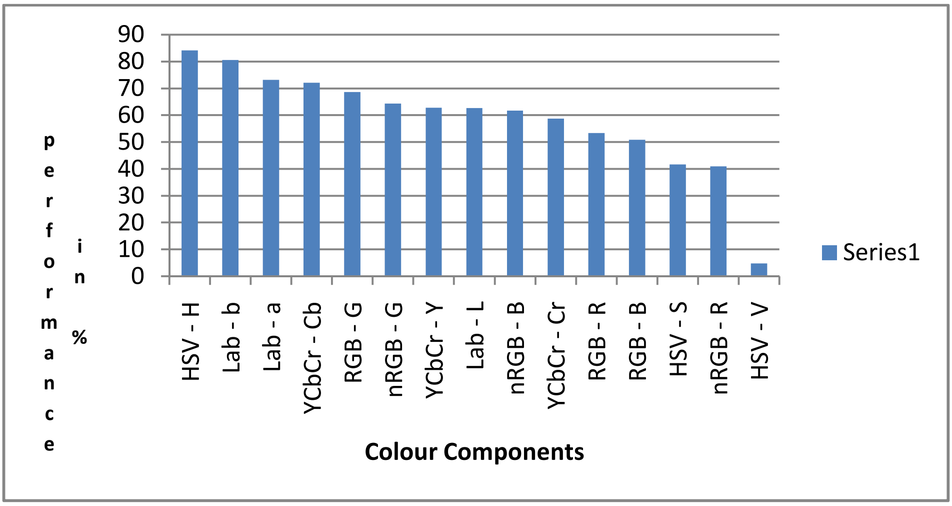


Figure 6.5: Overall segmentation accuracy of different colour components.

The percentage of overall accuracy ranges from 4.82 %(HSV - V) to 84.19 %(HSV-H). The skin accuracy is of the range 6.59 %(HSV - V) to 99.48 %(Lab-a), while the lip performance range is 2.02 %(HSV-S) to 92.73 %(nRGB-G). 99% of teeth pixels are correctly identified in both YCbCr – Cr and HSV-H colour components, while teeth pixel identification is very small for S and V components of HSV colour space. The range of accuracy for dark pixels varies from 0(Lab-a,b) to 96.8(RGB-R). The accuracy is lowest for tongue identification which is in the range 0(Lab-b) to 79.8(HSV-H).

6.3 Lip Segmentation using Active Shape Model (ASM)

Lip is the most important visible articulator in human speech production system. Lip motion accounts for more than 80% of visual cues during a conversation and plays a pivotal role in speech perception [117]. Hence the segmentation, tracking and synthesis of lips are one of the most attempted tasks. A detailed report of different algorithms and their effectiveness in lip segmentation and tracking is presented in chapter 2. The wide assortment of techniques used for lip tracking start with simple thresholding scheme to Bayesian classifiers and Gaussian Mixture Models(GMM) based Estimation Maximisation(EM) algorithms, and currently to the widely used machine learning algorithms including deep learning approaches.

This section elaborates on the use of Active Shape Model (ASM) , for lip segmentation and the experimental results obtained on the data set including MAVSC-IP and MAVSC-IW. ASM is a statistical model defined based on Principal Component Analysis. Many attempts are reported in literature which use ASM for lip segmentation and localisation. Fabian *et al.* used ASM for lip contour localisation [303]. The work analysis the tracking results by attempting improvements in initialisation methods, profile models and in search algorithms. *Luettin et al.* [164] and *Matthews et al.* [304] used ASM for lip tracking as part of feature extraction procedure. *Kyung et al.* uses an approach which blends ASM and snake model for lip contour extraction [289]. ASM has 3 components,

i. Training component from which a shape model is generated

ii. Profile model characterising the neighbourhood of landmark points

iii. A search procedure for locating shape in unknown images.

The following section describes the method used for lip shape modelling using ASM.

6.3.1 Lip Shape Modelling using ASM

Shape is the geometrical attribute of an object after removing translation, rotation and scale effects from its representation [305]. Shape being a salient visual feature in an image, its representation, searching and synthesis are decisive in many applications. Shape is represented by contour based and region based approaches. In contour based approaches only the contour of the interested shape is considered while the entire image region forming the shape is taken into account for region based approaches. Shape representation by a set of land mark points or shape feature points is the most prominent approach. A land mark is defined as a point of correspondence on each object that matches between and within populations [305]. This scheme is known as point distribution model for representing shape. For shape processing in 2 – D, the (x,y) coordinates of n land mark points representing the shape are concatenated to form a vector (x_1,y_1,x_2,y_2....x_n,y_n).

In this implementation outer and inner lip shape contours are represented by 36 land mark points or shape feature points. The outer lip contour is represented by 20 points, while the inner lip contour is represented by 16 points. A detailed description of the shape feature points and manual land marking performed on the visual speech corpora is described in chapter 5. Shape feature points are generated for all phonemes by manual land marking. The spatial arrangement of shape feature points in images representing ത/ t / and ഫ/ ph / are shown in figure 6.6.

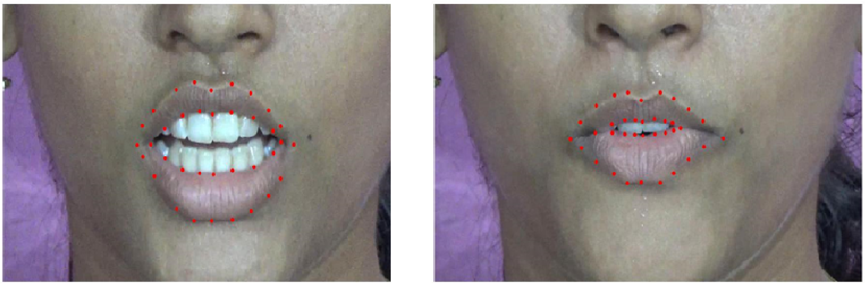


Figure 6.6: lip shape feature points for phonemesത / t / andഫ / p^h /

In ASM, the shape and its variability are represented in the PCA space. Before applying PCA, a pre-processing stage with Generalised Procrustes shape Alignment (GPA) procedure is applied on the shape coordinates of land mark points [309]. Distortions introduced by slight movements of the speaker are removed using Procrustres shape alignment procedure. Even though the video is captured in ideal conditions, variability introduced by changes in position and orientation of the talking faces exists for some images. Such changes that happened due to translation, rotation and scaling effects are corrected by applying Generalised Procrustes Analysis (GPA) to the shape vectors and Region of Interests (ROIs) of selected frames. GPA aligns the set of given shapes with respect to a mean shape which is computed from the same set of shapes. GPA is an iterative procedure which updates the orientation, scale and origin of the current mean on each pass. The iterative scheme used by Cootes *et al.* is used in this work [309]. Similarity transformations (Scaling, rotation and linear translations) are applied for aligning lip shapes.

GPA will make sure that PCA captures only local non rigid transformations [306] of lips by ensuring that GPA removes the effects due to translation, rotation and scaling. After this, PCA is applied on the coordinates of the shape feature points. PCA computes most appropriate and significant set of axes for representing the data using minimum dimensions, and hence it is one of the most widely used dimensionality reduction techniques [332]. In this work PCA is applied on the 36 shape feature points obtained from the manually land marked images. The shape feature points are vectorised, which forms a 72 dimensional vector corresponding to each image. PCA finds a more appropriate representation for these vectors using fewer number of dimensions and the steps for computing PCA is given bellow.

Step 1: The mean of the shape vectors can be calculated as

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N}(X_i)$$

Step 2: Calculate the Covariance matrix as

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(X_i - \overline{X})^T$$

Step 3: Compute the Eigen vectors φ_i and Eigen values λi of Σ arranged such that $\lambda i > \lambda_i + 1$

The principal axes in the covariance space are calculated as the Eigen vectors of the covariance matrix. The set of Eigen vectors S is arranged according to the decreasing value of λ_i, Eigen values. The set of Eigen vectors with the largest values of λ_i is only chosen for representation. The number of retaining Eigen vectors are a decision based on the trade-off between dimensionality reduction and accuracy [308]. The new set of coefficients in the PCA space is obtained by multiplying the matrix of Eigen vectors, with largest Eigen values with the original shape vectors.

Any shape in the training data set can now be approximated in the reduced dimensional PCA space. ASM training computes the mean shape from the set of training images. The variation from the mean is the PCA based model coefficient for representing a shape.

Let x be a member in the training set, x can be represented in the new space as

$$x = \bar{x} + sb \qquad (6.4)$$

Where p=(p_1,p_2,...,p_t) , the ser of Eigen vectors corresponding to the t largest Eigen values($\lambda_i's$) of the covariance matrix and b is a t – dimensional matrix given by

$$b = s^T(x - \bar{x}) \qquad (6.5)$$

b is the vector of model parameters representing the given shape x, known as shape parameters. We can vary the parameter b to obtain different shapes, by ensuring that b varies within the limit $3\sqrt{\lambda_i}$to-$3\sqrt{\lambda_i}$ [309] images. In this implementation 9 Eigen vectors with the largest Eigen values are retained for representing shape in the PCA space. Figure 6.7 shows the variation in the shape due to the first 3 modes corresponding to the 3 largest Eigen values.

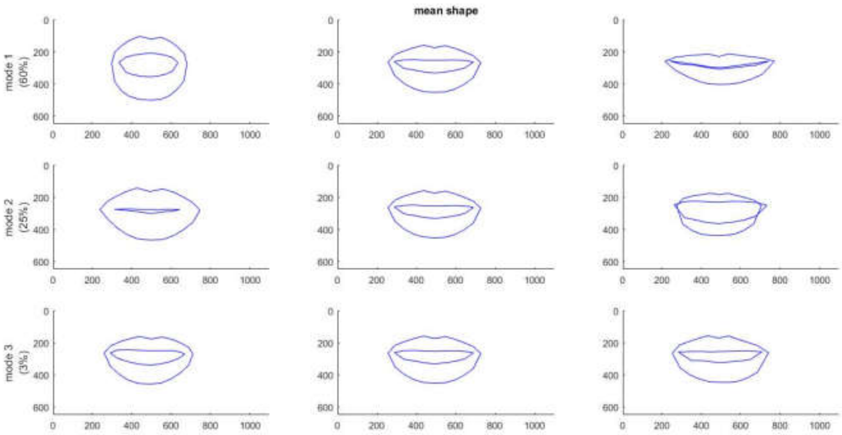


Figure 6.7: The variation in the shape due to the first 3 modes

6.3.2 ASM Profile Model

Profile model characterises the appearance context of each shape feature point in the shape model. The best fit among neighbourhood pixels of a candidate shape feature point is selected in ASM search by comparing the corresponding profile vectors. A profile model is either 1D or 2 D. In Coote's original model a 1-D whiskers, which runs through the land mark point and perpendicular to the shape boundary is used [309]. Figure 6.8 shows the placement of whiskers on lip land mark points.

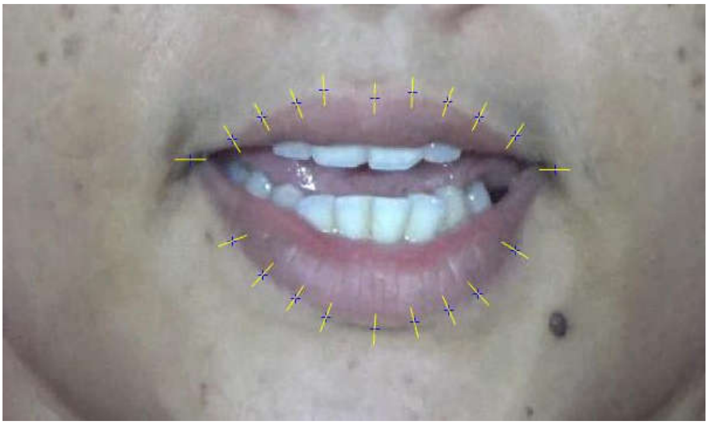


Figure 6.8: The 1 - D profile vector of outer land mark points

Let us discuss the making of the profile vector for a single land mark point. Let, P = [g_1,g_2,…g_N] is the 1-D whisker vector characterising the current land mark point. Instead of the original vector the gradient of the vector is used, which is represented as[g'_1,g'_2,…g'_N] and computed using equation (6.6)

$$g'_i = g_i - g_{i-1} \tag{6.6}$$

The gradient vector is normalised to remove the variation due to illumination changes using equation (6.7)

$$\hat{P}'_i = \frac{g'_i}{\sum_{i=1}^{N} |g'_i|} \tag{6.7}$$

The vector [$\hat{P}'_1, \hat{P}'_2 \ldots . \hat{P}'_N$] is the normalised profile vector of greyscale gradients. The data is assumed to be a multivariate Gaussian distribution and the profile model is represented as mean and covariance matrix computed separately for each land mark point. The mean profile $\overline{P_k}$and covariance matrix S_k are derived from normalised profile vectors of kth land mark points using all training images. The process is repeated to find the mean and covariance of all land mark points which forms the profile representation in ASM model.

During ASM search, a candidate profile vector is formed by applying equation (6.6) and (6.7) on texture values of the current whisker. The candidate profile vector, x is compared with the model profile of the land mark point using Mahalanobis distance. Mahalanobis distance is defined as

$$Distance = \sqrt{(x - \overline{P_k})^T {S_k}^{-1} (x - \overline{P_k})} 6.8$$

Where $\overline{P_k}$ is the mean and S_k is the covariance matrix [309,289]

A 2 – D profile, instead of the 1- D whiskers, is reported to be improving performance of ASM search [310]. The work by Fabian et al successfully implemented a 2- D profile vector for ASM based lip tracking [303]. The present work also employs the same technique. The 10x10 profile region around 4 land mark points is shown in figure 6.9. Unlike 1-D profile, which is orthogonal to shape edge the 2-D profile window is straight and aligned with the original image grid. Mean and covariance matrix of the profile vectors computed across training images is the profile representation in the ASM model. Each profile square region is made in to a long vector for computing mean and covariance. The following sequence steps is used for extracting 2-D profile vector

1. The intensity gradient is captured by applying a 3x3 mask to the pixels in the profile window around the neighbourhood of each landmark point. The mask used in the work is $\begin{bmatrix} 0 & 0 & 0 \\ 0 & -2 & 1 \\ 0 & 1 & 0 \end{bmatrix}$
2. Divide each element by the sum of absolute value of all elements (normalisation)
3. A sigmoid equaliser is applied to reduce the dynamic range and avoid outliers, which is defined as $x' = \frac{x}{abs(x)+c}$, where c decides the shape of the sigmoid function.
4. The square region is vectorised to calculate mean and covariance

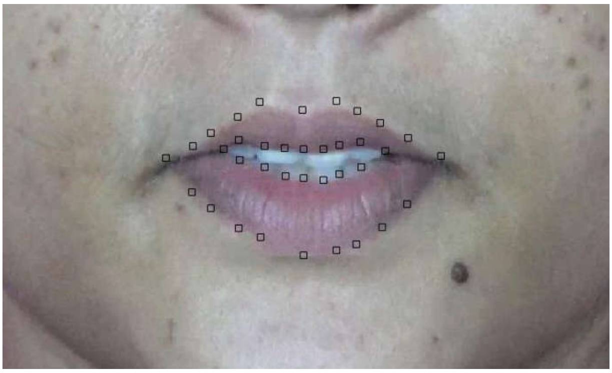


Figure 6.9:The 10x10 profile region around 4 land mark points

During ASM search x-displacements are orthogonal to the shape edges and y-displacements are tangents to the shape edges following the approach adopted by S. Milborrow [310].

6.3.3Active Shape Model Search

The ASM search is an iterative procedure with two phases. After initialisation the first phase relocate each shape feature point and second step corrects the location of points to maintain the relative positioning which is encoded in the shape model. The initial shape is either obtained from the previous frame or obtained by placing the mean shape in the image based on a computed centroid. The centroid is found either through user intervention or as the midpoint of mouth region image extracted using Viola John's algorithm [336].

The ASM search Algorithm is given bellow and the implementation is based on Cootes original work expect in profile selection. 2- D profile which is found to be better performing for lips than the 1-D whisker, is used for search.

```
ASM search Algorithm

Input: Image ROI, Model
Output: Land mark points representing lip shape in Y

1. Initialise shape Y
2. Repeat the following steps until convergence
       Find a new shape X from Y by recomputing the location of each land mark point
       Find Y by adjusting X to conform to the shape model
```

The steps for relocating the land mark points executes a profile vector matching across the neighbourhood pixels of the current land mark point. The profile vector is computed for each point in the neighbourhood. The computed values are compared with the model values based on the Mahalanobis distance given in equation (6.8) and the best fit neighbourhood pixel is the new land mark point.

Step 4 of ASM search algorithm performs the mapping of suggested shape from step 3 to the model space, to maintain the relative positioning of land mark points. The distance between X and model shape can be calculated as

$$distance(X, T(\bar{x} + Sb)) \tag{6.9}$$

Step 4 tries to find T and b that minimises the distance given in equation (6.9). T, the similarity transform on a pixel (x,y) can be represented as

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} tran\_x \\ tran\_y \end{bmatrix} \begin{bmatrix} S\ cos\emptyset & S\ sin\emptyset \\ -S\ sin\emptyset & Scos\emptyset \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{6.10}$$

Where tran_x and tran_y are translation coefficients and $\emptyset$ is the amount of rotation and S is the scaling parameter. The iterative procedure for finding optimum T and b is given in Cootes paper [309]. The proposed multi resolution based improvement of ASM search is used in this work. The shape feature points are extracted from the full resolution image. But the profile vector is calculated at each resolution levels, coarse to fine resolutions. Profile vector at separate resolutions expands the capture region of shape feature points. The ASM search starts with coarse resolution image and proceeds towards finer resolution image. An up sampled version of the coarse resolution search output is the input to the next high resolution search. Profile vectors at three resolution levels are used in this work, which is shown in figure 6.10.

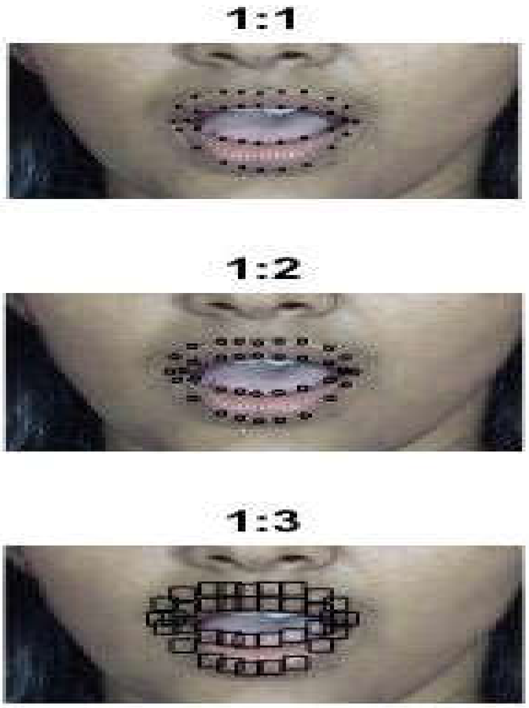


Figure 6.10: Profile vectors at 3 resolution levels

Figure 6.11 shows the results at different iterations. Red dots show the updated current points and blue dots show the values before current updating. Line is drawn connecting updated points.

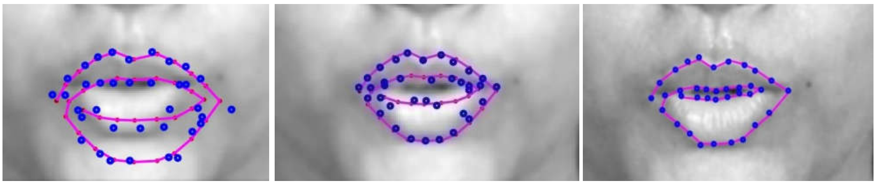


Figure 6.11: Intermediate results during ASM search, red dots show the updated current points and blue dots show the values before current updating

6.3.4 Experimental Results

The ASM training and searching are executed in different in different colour spaces.Land mark localisation using ASM is executed for 6 different top performing colour components (found using Bayesian classifier) using a single speaker dataset with 600 training images and 100 testing images. The Euclidian distance between the actual landmark points and the computed points for the test images is the metric used for comparisons. The sum of Euclidian distance for the 36 land mark points, between the computed and the actual is found for each image in the test data set, and the average of it for the 100 images is computed to be used as a comparison metric. Table 6.3 lists the average Euclidian distance in ASM search for different colour components for a single speaker training and testing framework. The lower Euclidian distance (highest performance) is obtained using H component of HSI colour space.

Table 6.3:The average Euclidian distance in ASM lip tracking for different colour components.

Sl. No	Colour Component	Average Euclidean Distance (in pixels)
1	HSV – H	55.43
2	Lab – b	67.6
3	Lab – a	79.3
4	YCbCr – Cb	81.2
5	RGB – G	92.67
6	nRGB – G	93.3

The average Euclidian distance for 1 – D whisker is 18.1 and for 2 – D profile vector is 15.43 in the hue space.Hence 2 – D profile vectors are used for tracking. The final tracked lip varies significantly based on the initialisation. Figure 6.12 shows different tracking results for the same image with 3 different centroid initialisations. The central image, which is the correctly tracked one, starts with centroid initialised approximately to the geometrical centre of the lip.

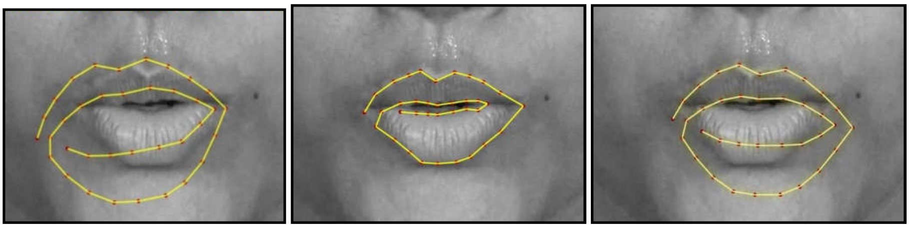


Figure 6.12: Different tracking results for the same image with 3 different centroid initialisations.

The tracking results on different images of a single speaker using H component of HSI space in the semi-automatic framework with centroid initialisation are given in figure 6.13.

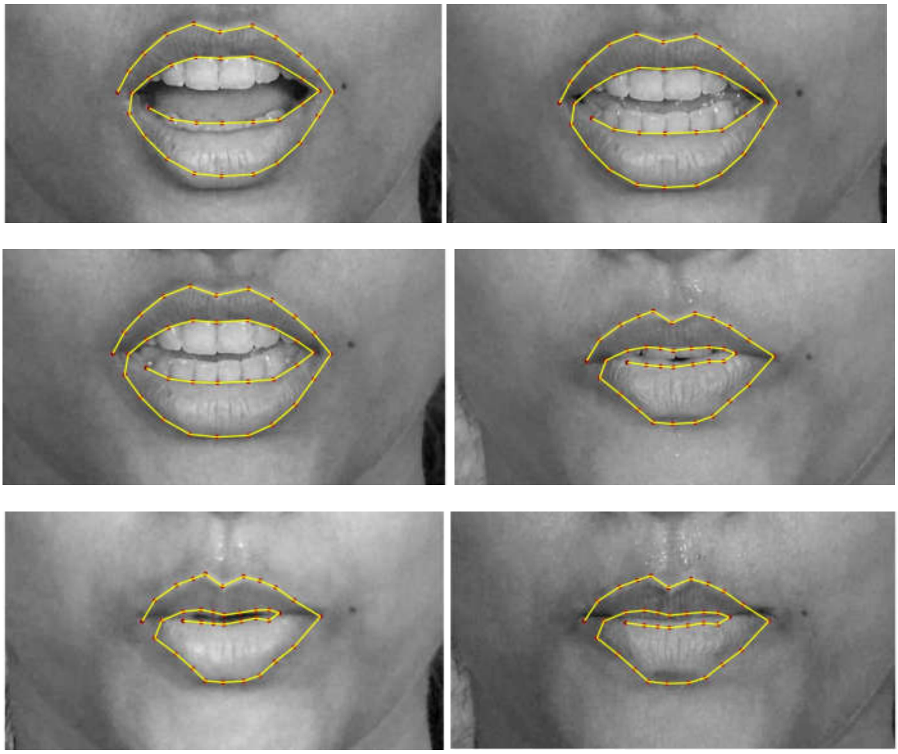


Figure 6.13: The tracking results on different images of a single speaker using H component of HSI space in the semi-automatic framework with centroid initialisation

A subset of the images in the MAVSC-IW and MAVSC-IW datasetsis manually landmarked (selected images of 3 speakers) for the extraction of lip geometric features, which is explained in the last chapter. ASM trained and searched on a single speaker dataset in hue component space is employed for land marking the remaining images in the dataset. Selected images for each speaker are land marked manually and the ASM is trained using these images. The minimum number of images used in the training set is 100, which is used to landmark a dataset with more than 1000 images. Lip shape is initialised either using manual centroid marking or initialisation from pervious frame methods. These images,landmarked using ASM search are used both for convolution neural network and for lip movement synthesis applications to be discussed in the next chapter.

6.4 Lip Segmentation using Convolution Neural Networks (CNNs)

Finding the appropriate set of features for representing data to solve a problem is crucial for both humans and machines. Conventional machine learning approaches use features selected by human beings and machine just computes these features. The task assigned to machine is the mapping from feature to the desired output. On the other hand deep learning discovers the appropriate feature for solving the problem by a process known as representation learning. The machine learned representation is found to outperform hand-designed representations in many domains. Auto encoders are a class of effective learning representation algorithms. Deep learning creates a hierarchy of representations in which higher level representations are explained in terms of lower level representations [337]. Deep learning is not a new technique. On the other hand its genesis dates back to 1940s. It appears to be new because of its huge popularity due to big successes in areas such as driver less cars and speech recognition. The first artificial neural networks, inspired by biological neurons have been created in the middle of last century. Deep learning systems are applied the multiple levels of composition of human brain perception framework. Another contributing factor to the success of the deep learning systems is the exponential increase in the training dataset. The generation of digital natives creates data from every aspect of their day to day life. Convolution Neural Networks (CNNs) are feed forward neural networks which use convolution layers for feature extraction. Conventional feed forward networks are the simplest ANNs first devised in 1960s. Section 6.4.1 explains the peculiarities and architecture of CNNs and section 6.4.2 describes the implementation details of CNN for lip segmentation and tracking with experimental results.

6.4.1 Architecture of Convolution Neural Networks (CNNs)

A neural network is a collection of artificial neurons. In each node(neuron) the input values are multiplied with weight to compute the net output of a neuron. Convolution neural networks are special feed forward neural networks designed for solving image processing tasks mostly in the domain of computer vision. CNNs consist of 3 type of layers, convolution layers, pooling layers and fully connected layers. Convolution layers, characterising the network are intended for extracting features from the images. Pooling layers are intended for down-sampling and the fully connected layers are exactly similar to conventional feed forward neural network layers. As CNNs deal mostly with images, the layers are 3-dimensionalwhich corresponds to the width, height and depth (number of colour channels) of images. The convolution layers are not fully connected and has an important property known as parameter sharing.

The number of parameters or weights in a fully connected network with image input is enormously large. Consider the first layer receiving a 512x512x3 images, in a fully connected network. Each neuron in the layer need to be connected to 7,86,432(512x512x3) nodes in the previous layer. Each connection is associated with a weight. With ‘m’ neurons in each layer and ‘n’ layers with 7,86,432 weights per neuron, the network has to operate in a huge parametric space. Naturally the whole process of training and applying will be very slow and may lead to over fitting. In fact the neurons in convolution layer operate on selected blocks to extract a single feature. The neurons in a single layer extract the same feature from different regions of an image which results in a uniform connection weight and the property is called shared weights. A different layer of neurons acts on the same region to extract a different feature. A hierarchy of features can be created by making the output of one convolution layer as the input of another convolution layer.

The convolution layer operates by mimicking the mathematical operation convolution. In image processing a convolution is obtained by sliding a convolution mask across the pixels in the image. The mask coefficients are multiplied with corresponding pixels and added, which replaces the original pixel. The convolution operation using a 3x3 mask is depicted in figure 6.14.

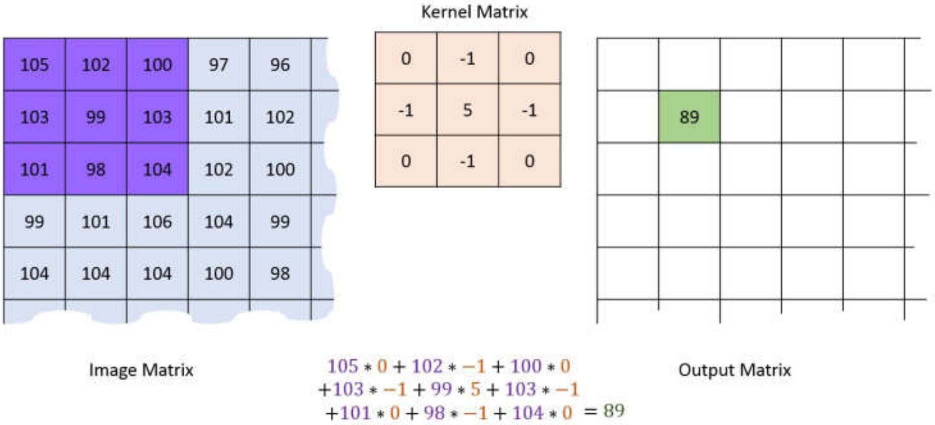


Figure 6.14: The convolution operation on two adjacent pixels with a 3x3 mask

The coefficients can be changed to extract different kinds of features. Every neuron performs a similar operation of element wise multiplication followed by addition. The connection weights are made to mask coefficients in order to perform convolution. If the size of the convolution mask is 5x5, the dimension of weight matrix for a neuron will be 5x5x3. The number of neurons in a layer depends on a factor known as stride, which is the distance of mask shift for each calculation. For example in a 512x512x3 image, with stride 1 the number of neurons will also be 512x512(padding is needed). But if the stride is 2 for the same image the number of neurons will be 256x256. Generally for an image of dimension WXH stride S, the number of neurons will be (W/S, H/S) with proper padding. In the first case (with stride 1) there will be 512x512x5x5x3 connection weighs. But all neurons share the same weight, bringing down the number of different weights to just 75(5x5x3). The above property is known as shared weights. The concept of drop out is used to overcome problems due to over fitting. The idea is to drop random nodes and its connections during training with a given probability. A network with n nodes can have 2n possible sub networks. During testing an approximate averaging method is used for finding the network parameters. The next section discusses the details of using CNNs for lip tracking using landmark localisation.

6.4.2 CNNs for Lip Segmentation by Landmark Localisation

The general approach used in this work for lip segmentation and tracking is to identify the landmark points representing inner and outer lip contours. The same approach is used for lip segmentation using CNNs. The network is trained with annoted images, which in turn learns the mapping from images to landmark points.The network is implemented using pytorch, which is one of the most popular research platforms [338]. The various phases of implementation is explained bellow.

1. Information about the images and keypoints in this dataset are summarized in CSV files
2. Image Pre-processing - A composition of the following operations is applied on the dataset.

I. The RGB images are converted to gray scale images

II. The image is rescaled to the size 224x224, after appropriate rescaling and croppingof the data.

III. Normalizing the images and key points; turning each RGB image into a grayscale image with a colour range of [0, 1] and transforming the given keypoints into a range of [-1, 1]

IV. Turning these images and keypoints into Tensors,which is a multidimensional array [338]

3. Define the network with dimension,stride and padding for each layer. The specification of the convolution layers of the CNN used in this work arrived after several experiments by varying network parameters and functions is given in table 6.4.

Table 6.4:The specification of the convolution layers of the CNN used in this work

Layer	Input Dimension	Filter Size	padding	Striding	Output dimension
Convolution layer 1	224x224x1	7x7x1	1	3x3	74x74x32
Convolution layer 2	74x74x32	5x5x1	0	3x3	24x24x64
Convolution layer 3	24x24x64	5x5	1	3x3	8x8x128
Convolution layer 4	8x8x128	3x3	0	1x1	6x6x256
Convolution layer 5	6x6x256	3x3	0	1x1	4x4x512
Convolution layer 6	4x4x512	1x1	0	1x1	4x4x512

The convolution layers are followed by three fully connected layers. The first layer consists of 1024 nodes and the number of nodes in the last output layer is fixed to be 72, which is the (x,y) coordinates of 36 landmark points. The dropout for fully connected layers is fixed to 0.3.

4. Training the network: The network is trained using 10,000 landmarked Images from MAVSC-IW and MAVSC-IW datasets. Images landmarked using ASM are also included in the training database after a manual inspection. Batch size used is 200, which is applied after shuffling. The Adaptive Moment Estimation or Adam optimization algorithm is used to train the network. The Adam optimization algorithm is a combination of gradient descent with momentum and RMSprop algorithms [339].
5. The test data set consists of 2000 images. Data is visualised by displaying original and tracked landmark points on test images. The output of 9 images in the testing dataset is given in figure 6.15. Original landmark points are shown in pink, while green circles indicates the land mark points found by the trained network.

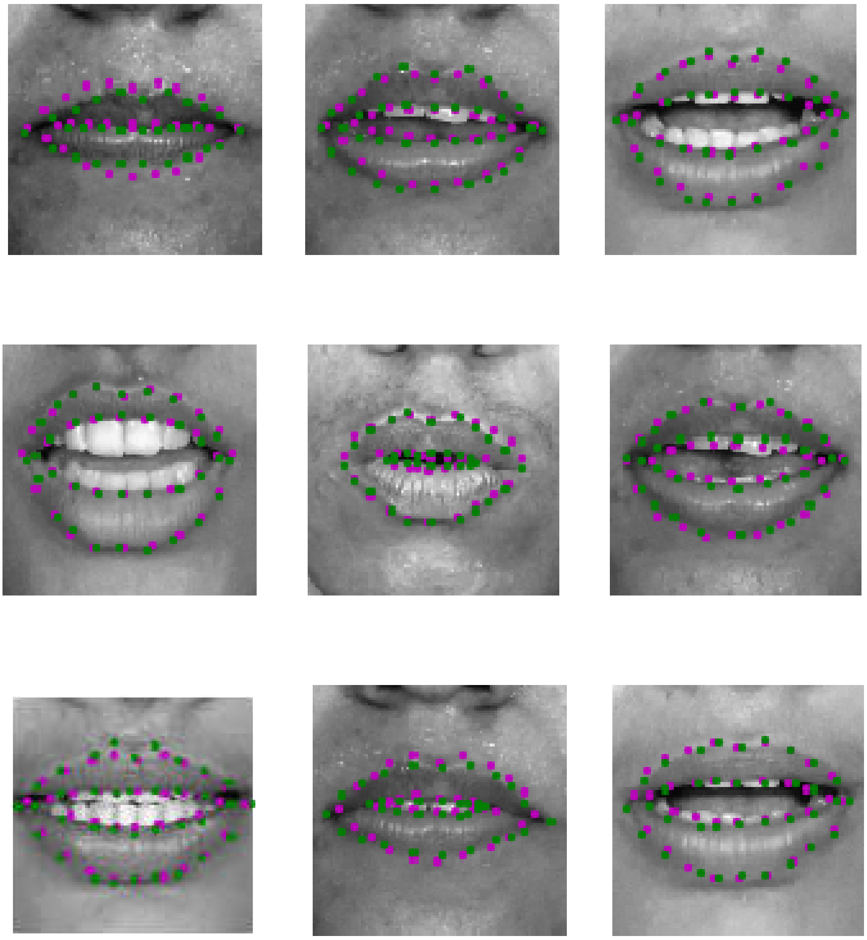


Figure 6.15: The output of 9 images in the testing dataset,original landmark points are shown in green, while pink circles indicates the land mark points found by the trained network.

Tracking experiments are conducted by varying network parameters and input data set. The Euclidian distance (the average of the training data set) between the actual land mark points and generated points is the metric used for comparison. The average Euclidian distance corresponding to different network configurations is shown in table 6.5. Finally the configuration with six convolution layers and Selu activation function, with lowest average Euclidian distance is used in this work.

Table 6.5: The average Euclidian distance for different network configurations with number of epochs=100and dropout=0.3

Layers	Activation function	Average Euclidean distance
Convolutional layer-6 fully conncted layer-3	Selu	17.5054
Convolutional layer-6 fully conncted layer-3	Relu	729.872
Convolutional layer-6 fully conncted layer-3	Elu	244.1181
Convolutional layer-5 fully conncted layer-3	Selu	17.6913
Convolutional layer-4 pooling layers-4 Fully connected layers-3	Selu	144.2143
Convolutional layer-4 pooling layers-4 Fully connected layers-3	Relu	188.5379
Convolutional layer-4 pooling layers-4 Fully connected layers-3	Elu	211.4157
Convolutional layer-5 fully conncted layer-3	Relu	117.2315
Convolutional layer-5 fully conncted layer-3	Elu	194.7534

Experiments are conducted by varying the number of epochs, from 100 to 700. But no improvements in the Euclidian distance are reported above 500 for any of the network configurations.

6.5 Conclusion

In this chapter statistical analysis is performed on the MAVSC-IW and MAVSC-IW images for understanding the colour properties of different mouth regions in Indian skin tone context. Hue of HSI colour space is identified to be the best performing colour component in mouth region pixel classification experiment conducted with Bayesian Classifier. ASM, trained and searched on a single speaker dataset in hue component space is employed for land marking images in the MAVSC-IW and MAVSC-IW dataset. These images, landmarked using ASM search is used both for experiments with CNN and for lip movement synthesis applications. Finally experiments are conducted with CNNs for lip segmentation using land mark localisation method.

Chapter 7

Comprehensive Malayalam visual speech synthesis framework using Independent Active Appearance Models

7.1 Introduction

Speech is bimodal in nature. The contribution of the visual part in the perception of speech, especially in the noisy environment is an established fact [311,242,312,276]. Image sequence of talking face with audio can create realistic conversation environment and will surely make machines more human-friendly. Audio visual speech synthesis has potential applications in facial animation automation, building user friendly interfaces, audio visual speech perception studies, developing support system for persons with hearing disabilities, virtual teleconferencing, e-learning *etc.*The first attempt for speech synthesis was purely mechanical, before the computing era where mechanical articulators supplemented an artificial vocal tract [313]. The attempts to automise facial movements started in 1970s, which was a basic polygonal mesh framework with closing and openings of mouth and eyes [314]. Then comes, limited set of hand drawn images representing posters such as mouth opening and mouth closed. where the realism is in the hands of an artist. Construction of hand drawings corresponding to expressive facial movements is time consuming and demands the involvement of expensive manpower. Demeny used a chrono photograph for displaying speech movements

[313].One of the initial works in the automisation of mouth synthesis was by Norman P. Erber and Carol Lee De Filippo in 1978. Lines representing mouth movements were analyzed and displayed using photo transistors and oscilloscopes [315]. An early attempt during digital era by Allen A. Montgomery in 1980, synthesised speech animation by displaying a sequence of limited set of primitive lip shapes [316]. The progress in 1980s started with facial muscle modeling, progressively leading towards the development of facial action system [317,188]. Tin Toy, a pixar production in 1998 is considered to be the first truly computer generated realistic character [199]. The last decades witnessed the emergence of a plethora of new techniques for facial movement automation, especially while talking. Different facial animation techniques are used in these days to automate the movements of visible articulators [315]. Hidden Markov Modeling based approaches to lip motion synthesis can successfully incorporate co-articulation effects to the system [318,171,319,320]. Time delay neural networks are also used for lip synthesise with temporal correlation and computational advantages [321]. Dimensionality reduction techniques are also used effectively to synthesis lip images by combining syllables [186]. Yu Ding *et al.* use Gaussian Mixture Model (GMM), called lip shape GMM for a lip animation synthesis framework [322]. Sign speech synthesis systems with special emphasis on lip movements are also emerging in different languages [323].

The lower jaw, tongue, teeth, and lips are the visible articulators of the human voice production system. The movement of visible articulators creates the varying image sequence corresponding to visual speech. Lip being the most dynamic visible articulator, lip motion synthesis is the most crucial component in making character animations believable. A realistic conversation experience can be imparted to the viewer by synthesizing realistic lip movements. Teeth visibility can also be explained in terms of lip shape.Speech animation is moving the facial features of a graphic model to synchronise the lip motion with the audio to give the impression of speech production [324]. Unrealistic speech animation distracts the viewers and prevents them from developing an attachment to the animated characters. The unpleasant experience created by the mismatch of lip movements and audio is a regular issue in cartoons in regional languages like Malayalam. Close up views of characters helps to develop attachment with characters. Hence production firms are keen on developing expressive speech animations. The conventional approach was to use performance capture systems together with a team of experienced animators. But this model is not suitable for low budget productions. Moreover multi lingual productions demand separate development for each language. The massive demand for low budget multilingual productions with lot of character animations is the main reason for the emergence of automated alternates to speech animation.

This study focuses on the development of visual speech synthesis framework in Malayalam, using Active Appearance Models (AAM). Active Appearance Model (AAM), a statistical data driven model, is used for modeling lip shapes and texture. Independent AAM, which models shape and textures using separate set of coefficients, is employed in this work. The framework is conceptualized as a text to visual speech synthesis system. The framework integrates all components explained in the previous chapters for the development of visual speech synthesizer in Malayalam. Even though the proposed framework is capable of synthesizing complete facial movements corresponding to a given sequence of phonemes or allophones, the work attempts to synthesis the lip movements only. The lip colour or texture used in character animations are not always their real life counterparts. So a separate synthesis framework is developed for creating the dynamics of lip shapes only. The complete shape and appearance synthesis framework uses morphing technique for generating intermediate frames. Perception experiments are conducted to compare the naturalness of phoneme, allophone and viseme based implementations of the synthesis framework.

The chapter is arranged as follows. After the introduction, section 7.2 introduces the theoretical foundations of lip modelling using independent AAMs. Section 7.3 describes the development of Malayalam visual speech synthesis framework using independent AAM and section 7.4 presents the experimental results of applying the framework for lip motion synthesis corresponding to Malayalam text. Section 7.5 concludes the work.

7.2 Talking Lip Modeling using Independent AAM

Active Appearance Models (AAMs) introduced by T.F. Cootes, G.J.Edvards and C.J.Taylor is considered to be the most successful model for interpreting deformable objects [324]. It is extensively used in medical image processing, computer vision and image synthesis applications to explain the shape and appearance of objects by model parameters. The complexity in modelling the variability of deformable objects persuades researchers to develop different avenues. Pieces and spring model based approach used by Martin *et al.* for representation and matching of images is perhaps the first deformable model used for image segmentation [325]. A viscous flow based model for brain is proposed by G.E. Christenson *et al.* [326]. Active contours, the idea popularly known as snakes proposed by Kass *et al.* are a landmark in the development of deformable models [288]. Eigen faces is one of the most successful methods in face recognition which uses PCA for representing face [327]. Active Shape Models (ASMs) and Appearance models belong to the category of interpretation by synthesis methods. This category of methods search objects by synthesising objects in their model parameter space. AAM has two components, a training component from which a shape and appearance model is generated and a search procedure using this model. This work employs modelling scheme used in AAM for representing the lip and its variability while talking for Malayalam.

AAMs are extensively used in visual speech synthesis applications. In the work by W Mattheyses *et al.*, independent AAMs with eight Eigen vectors for representing shape and 134 eigen vectors representing textures are used for photorealistic visual speech synthesis. The entire visual data base is converted into AAM model space and appropriate speech segments for the synthesis are selected in the model space [328]. In a similar work by W Mattheyses *et al.* segment selection is executed by using both target and join costs [92]. In their work Theobald *et al.* performed the subjective and objective comparison of acoustic driven and phonetic transcription based visual speech synthesisers. Independent AAM coefficients are used for representing the visual articulators in both schemes. Mel Frequency Cepstrum Coefficient (MFCC) is used as the acoustic feature and the mapping to AAM model space is performed by a feed forward Artificial Neural Network (ANN) [329]. A person specific facial appearance model using a variant of AAM with options for adding visual emphasis is a similar work [330]. The work by R. Anderson *et al.* tried to synthesis expressive speech. The inability of conventional AAMs to model local changes such as blinking of eyes is addressed in this work. An extension of HMM, known as cluster adaptive training, is used for synthesis in this work [331]. A multilevel training phase, employed in the work by M.M. Cohen *et al.* uses 184 prototype images corresponding to 45 phonemes. AAM search is used for land marking the remaining images in the corpora [214]. In the independent AAM approach, shape and appearance models are modelled separately, while a combined AAM models shape and appearance using a single set of parameters. In their work, Iain Matthews *et al.* discusses the advantages and disadvantages of both methods [306]. The following section describes theoretical foundations of independent AAM in detail.

7.2.1 Independent Active Appearance Model (AAM)

Active Appearance Model (AAM) is a generative statistical technique, which learns from the training examples. A set of images with landmark points charactering the object of interest is the training database for AAM modelling. The procedure adapted for fixing the landmark points in the inner and outer lip on images in MAVSC – IP and MAVSC – IW datasets are already explained in chapter 6.Representing shape using a set of landmark points is known as Point Distribution Model (PDP), in which land mark points are concatenated to a vector of vertex locations as $(x_0,y_0,x_1,y_1\ldots\ldots\ldots.x_n,y_n)^T$.The procrustres analysis based pre-processing employed on the land marked images in the training data set is also described in chapter 6. The next step is to apply Principal Component Analysis (PCA) to the set of concatenated vector of landmark points. PCA computes most appropriate and significant set of axes for representing the data using minimum dimensions [332]. The new set of axes in the PCA space is computed as the Eigen vectors of the covariance matrix. The most important Eigen vectors, based on Eigen value are only selected, resulting in dimensionality reduction. In the PCA space linear shape variations can be represented as deviations from the mean shape.

$$S = S_0 + \textstyle\sum_{i=1}^{l} S_i P_i \qquad (7.1)$$

where S_0 is the base mean shape, P_i are the set of shape model parameters and S_i are the eigen vectors corresponding to 'l' largest Eigen values.

For an independent AAM an appearance model is obtained by applying PCA on the appearance values of pixels inside the object boundaries. An image mask is applied to the training images to remove all pixels outside the object boundaries. The shape and hence the number of pixels inside the lip will be different for different frames. The training images are shape normalised by warping to the mean shape by using a piecewise affine warp. Warping involves coordinate transformations and image resampling. In piecewise affine warp employed in this implementation of AAM uses Delaunay triangulation of landmark points. A triangle correspondence is established between all S_0,S_i pairs. During a forward warp from S_0 to S_i the algorithm finds the enclosing triangle for a pixel in S0 and maps to S_i by finding the triangle correspondence [306,333]. The number of pixels inside the shape normalised patches for all images will be the same.

PCA applied on shape normalised textures creates a model space for appearance. In the AAM, space appearance can be modelled as

$$A = A_0 + \sum_{i=1}^{m} A_i \lambda_i \tag{7.2}$$

where A_0 is the base mean appearance, A_i are the Eigen images corresponding to the largest m eigen values and λ_i are the set of appearance model parameters.

In this work independent AAM is applied for synthesis of shape and appearance. In the independent AAM model the shape and appearance is a point in the $(p_1,p_2..p_l,\lambda_1\lambda_2....\lambda_m)$ space, i.e model coefficient space. So the object synthesis can be converted as a problem of finding the appropriate trajectory in the model space. Equation (7.1) and (7.2) for shape and appearance representation can be rewritten in matrix form as

$$s\_new = s_0 + sp \qquad (7.3)$$

$$a\_new = a_0 + a\lambda \qquad (7.4)$$

where s_0 is the base shape and a_0 is the base mean appearance, s is the matrix of shape Eigen vectors and a is the matrix of appearance Eigen vectors, p and λ are the shape and texture model parameters respectively, s_new is the synthesised shape and a_new is the synthesised texture.

The inverse mapping, to find the shape and appearance coefficients given an example image, can be obtained from (3) and (4) as

$$p = s^T(s\_new - s_0) \qquad (7.5)$$

$$\lambda = a^T(a\_new - a_0) \qquad (7.6)$$

Combined AAMs, which is not used in this work, are obtained by applying a further PCA on shape and appearance parameters. The next section explains the modelling and synthesis of talking lip movements using independent AAM.

7.2.2 Talking Lip Modelling

This section explains the use of independent AAM for modelling of lip area towards generating talking lips. A phoneme or allophone based training corpus used in this study is capable of accommodating all possible lip variability during talking. The procedure used to prepare the lip landmarked images in MAVSC – IP and MAVSC – IW datasets are already explained in chapter 6. Active Appearance Model (AAM) uses both hand annoted images and images accurately annoted by ASM and CNN for training and testing. Mouth area is the important dynamic portion for visual speech synthesis. The mouth area consists of lips, teeth, tongue and mouth cavity. Lips are always visible, but the visibility of other areas are highly dependent on the phoneme uttered.

The annoted lips can be used for the training AAM. In this work shape and texture are modelled separately. An image is considered as a point in the model space consisting of shape and texture coefficients. The synthesiser has to find the appropriate trajectory in the model space corresponding to a phoneme or allophone sequence. Inverse mapping is performed on the model coefficients in the trajectory to find a sequence of mouth region pixels. The outer lip contour is represented by 20 points, while the inner lip contour is represented by 12 points. Figure 7.1 displays the scatter plot of shape feature points (x_n,y_n), including outer lip and inner lip points in all frames of a single speaker.

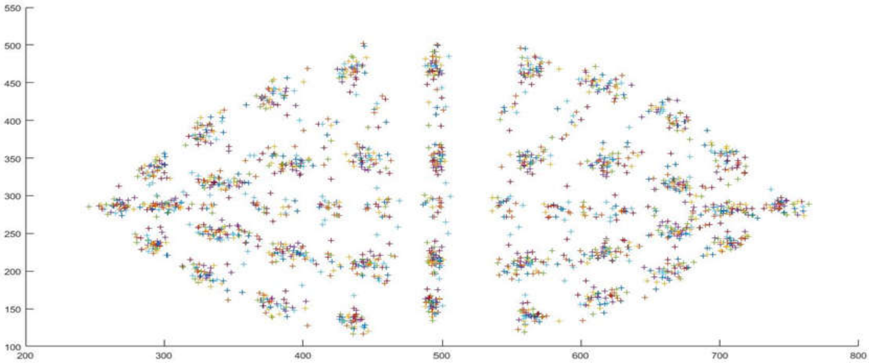
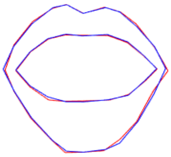
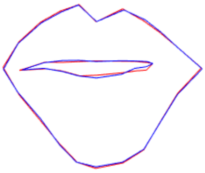
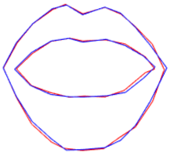
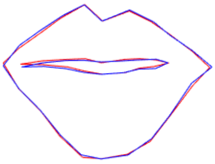
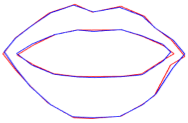
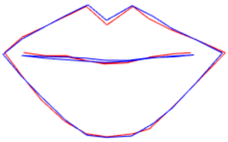


Figure 7.1 : Scatter Plot of shape feature points for 51 phonemes in Malayalam

In this work, 9 Eigen vectors with the largest Eigen values for representing shape in the PCA space are selected. After 9, the Eigen values is observed to be approaching zero. The model coefficient space is 9 dimensional instead of the original 72 dimensional shape vectors. Similarly this 50 Eigen vectors with the largest Eigen values for representing appearance in the PCA space are selected. After 50, the Eigen values is observed to be approaching zero. The model texture coefficient space is 50 dimensional instead of the original 72,000 (the number of pixels in the lip region) dimensional appearance vectors.

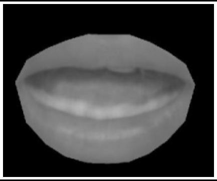
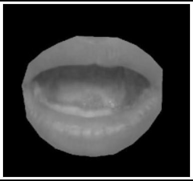
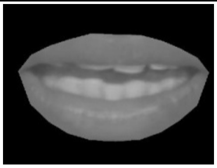
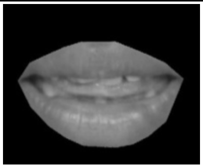
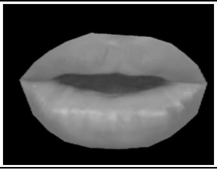
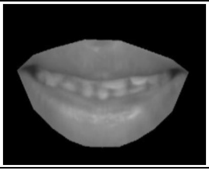
Phoneme based and allophone based AAM training is performed using MAVSC–IP and MAVSC–IW. The training and modelling is performed separately for single speaker and multi speaker mode. Frontal face images from 10 speakers, corresponding to Malayalam phonemes and allophones are used for training in the multi speaker mode. After the training phase, the shape and model coefficients corresponding to the key frames of phonemes or allophones are computed using inverse mappings. In the proposed implementation shape is represented by 9 shape coefficients. The shape generated from 9-dimensional ASM coefficients corresponding to five Malayalam short vowel phonemes and the consonant phoneme പ /P/ (with closed inner lip) are given in table 7.1. The original shapes from training images are also shown for comparison.

Table 7.1: Comparison between the original (red) and the model generated shape (blue) of selected Malayalam phonemes

Malayalam Phoneme	Shape Comparison	Malayalam Phoneme	Shape Comparison
അ/a/		ഉ/u/	
ഇ/i/		ഒ/o/	
എ/e/		പ/P/	

The mean normalised appearance image generated from 50-dimensional coefficient vector is wrapped to the generated shape of the corresponding phoneme. The synthesised appearances for selected phonemes is shown table 7.2.

Table 7.2: AAM synthesised lip appearances for selected Malayalam phonemes

Malayalam Phoneme	Synthesised Lip	Malayalam Phoneme	Synthesised Lip
അ/a/		ഔ/ou/	
ഇ/i/		ത/t/	
ഒ/o/		ച/c/	

A look up table of shape and appearance coefficients corresponding to phonemes, allophones and visemes generated after AAM training is stored for the purpose of visual speech synthesis. The AAM coefficient values averaged over all speakers is used for synthesising lip motions. The AAM coefficient values for each viseme is obtained by averaging the coefficient values corresponding to phonemes or allophones which are grouped to form the viseme.

The following section discusses the complete Malayalam visual speech synthesis framework developed using Independent AAMs. Even though the marking key points and subsequent modelling are conducted for mouth region alone, the framework is general and can be effectively applied for the complete talking face synthesis in Malayalam.

7.3 Malayalam Visual Speech Synthesis Framework using Independent AAM

A comprehensive visual speech synthesis framework for Malayalam using independent AAMs is proposed in this work . It is conceptualised as a text to visual speech synthesis system with options for synchronising with an input audio. The approach falls under the category of target based synthesis where trajectory is obtained by combining static frames, with certain intermediate approximations. Phoneme, allophone and viseme based look up tables for shape and appearance coefficients are the back bone of the proposed speech animation framework. The system is conceptualised as a word based framework consisting of 3 phases. In the first phase, the input text is converted to the sequence of Atomic Visual Units (AVU) of speech. The AVU can be phonemes, allophones or visemes (mapped from phonemes or mapped from allophones). In the second phase, the number of frames for each atomic unit is computed and the target image synthesis is performed in the final phase. Face image sequence corresponding to the given grapheme sequence, synchronised with the audio is generated in the final phase. Figure 7.2 depicts the 3 phase implementation of proposed visual speech synthesiser in Malayalam.

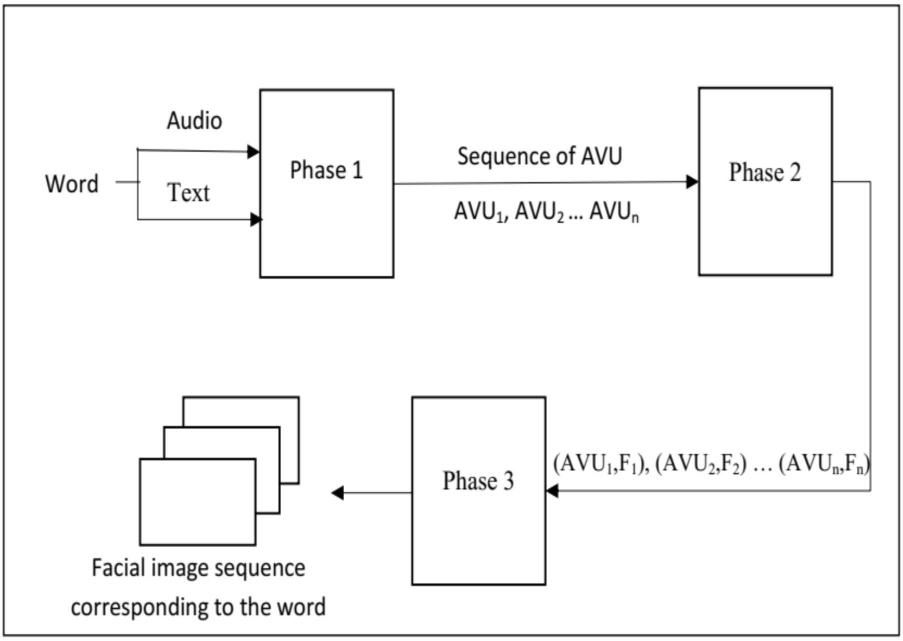

Audio
Word
Text
Phase 1
Sequence of AVU
$AVU_1, AVU_2 \ldots AVU_n$
Phase 2
$(AVU_1,F_1), (AVU_2,F_2) \ldots (AVU_n,F_n)$
Phase 3
Facial image sequence corresponding to the word

Figure 7.2: Process flow for the conversion of input text to the corresponding sequence of face images

In the first phase, the input text is converted in to sequence of AVUs , which is depicted in figure 7.3.

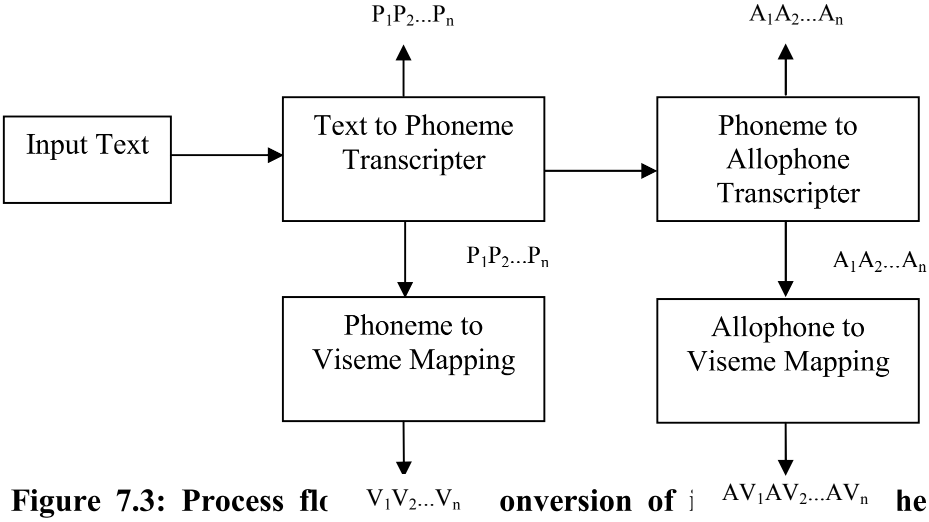

P1P2...Pn
A1A2...An
Input Text
Text to Phoneme Transcripter
Phoneme to Allophone Transcripter
P1P2...Pn
A1A2...An
Phoneme to Viseme Mapping
Allophone to Viseme Mapping
V1V2...Vn
AV1AV2...AVn

Figure 7.3: Process fl onversion of i he sequence of AVUs

(P_1,P_2..P_n), (A_1, A $_2$.. A $_n$), (V_1,V_2..V_n) and (AV_1,AV_2..AV_n) represent sequence of phonemes, allophones, visemes obtained from phonemes and visemes obtained from allophones respectively. The output from the first phase is a sequence of atomic visual speech units (AVU_1, AVU2...AVU_n) where *n* is the number of phonemes or allophones or visemes corresponding to a word.

In the second phase, the duration of each AVU in the output sequence is computed. Consider the use of framework in the application like character animation, the lip synchronisation with the input audio is very critical. The sequence of operations performed in phase 2 is given bellow

1. The audio of the spoken word (W_i) and the sequence of atomic units (AVU_1, AVU2...AVU_n) obtained from phase 1 are the input to this phase.

2. The actual duration of the input spoken word (W_i) is computed as D. The frame allocator computes the number of frames for each AVU. The procedure used for computing the number of frames for each AVU is explained in the following section. Duration table provides the average duration of Malayalam phonemes and allophones consolidated in chapter 2. The average duration of plosive consonants used is assigned as the sum of plosive duration and silence durations as explained in chapter 2.

3. Sequence of AVUs with its corresponding number of frames is the output from phase 2 operation.

The process flow of phase 2 operation is shown in figure 7.4.

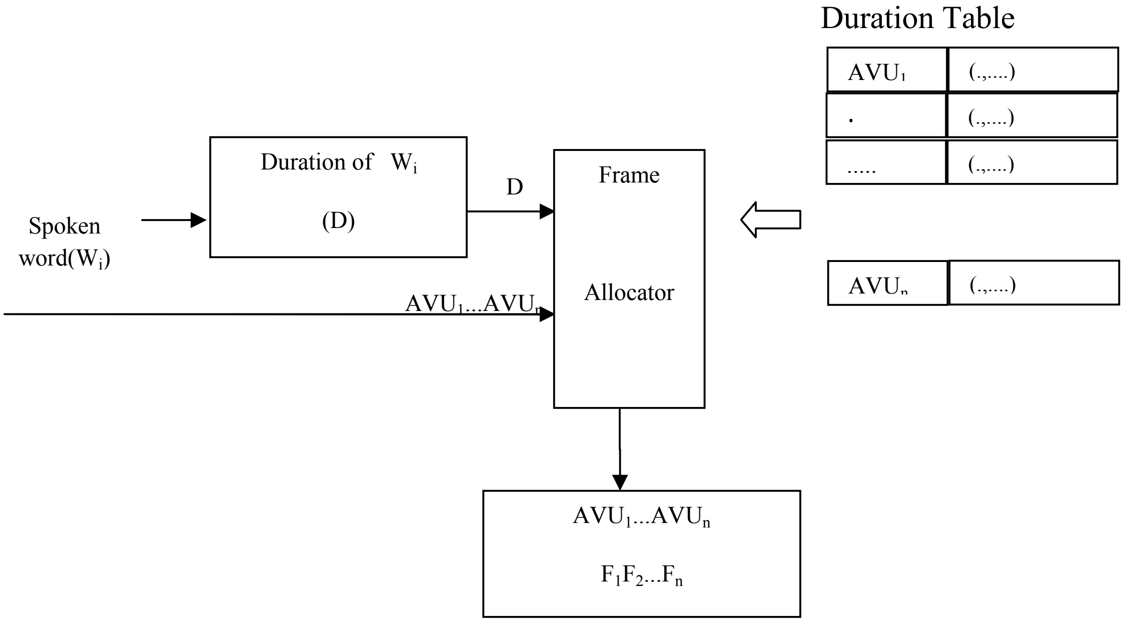


Figure 7.4: Process flow for computing the number of frames for AVUs

The frame allocator computes the number of frames for each AVU, using the following procedure. Let D is the actual duration of spoken word. The allocator first finds the sum of average durations of AVUs in the sequence, which is obtained from the duration table and let it be d. The duration of the word in actual utterance (D)is not the same as the sum of average phoneme durations (d).

Allocator finds the fractional unit duration of the i^{th} unit as

$$\text{fractional\_duration}_i = \text{average \_duration of AVU}_i/\text{d} \quad (7.7)$$

The target duration of the i^{th} unit for the synthesiser is computed as

$$\text{unit\_duration}_i = \text{fractional\_duration}_i \text{ x D} \qquad (7.8)$$

The number of frames corresponding to duration unit_duration_i is computed as

$$f_i = floor(\frac{unit\_duration_i}{frame\_duration}) \ (7.9)$$

where frame_duration is the duration of a frame obtained as 1/ frame rate. The elapsed time due to the round off to the nearest integer is assigned to the initial phonemes in the order. The remaining fractional time, which is less than a frame duration after the mentioned computations is summed for each word and will be added to the duration of the last phoneme in the sentence. In general application where the input audio is not available (where synchronisation need not be considered) the duration can be computed solely based on the average duration information.

In phase 3, the image synthesis module generates the visual speech image sequence using inputs obtained from phase 2.Three separate models are designed to perform this synthesis operations, which are described below. The first model synthesises the shape of target object , while the remaining 2 models can synthesise both shape and texture of frontal face talking images corresponding to a given text.

i. Shape Synthesis Framework (SSF)

SSF generates the sequence of shape of target object corresponding to a grapheme sequence. The SSF model depicted in figure 7.5 uses ASM coefficients to generate target object shapes corresponding to a phoneme, allophone or viseme. In this framework, the shape for an AVU is represented by a single set of coefficients corresponding to a single frame. The intermediate shapes are generated by linear interpolation. The model is particularly relevant for the synthesis of lip motion during talking. Lip shapes generated by the model can be used for lip motion synthesis in character animation. The lip colour or texture in character animations are not their real life counterparts. In such applications, lip colour is chosen based on the overall colour theme assigned to the character.

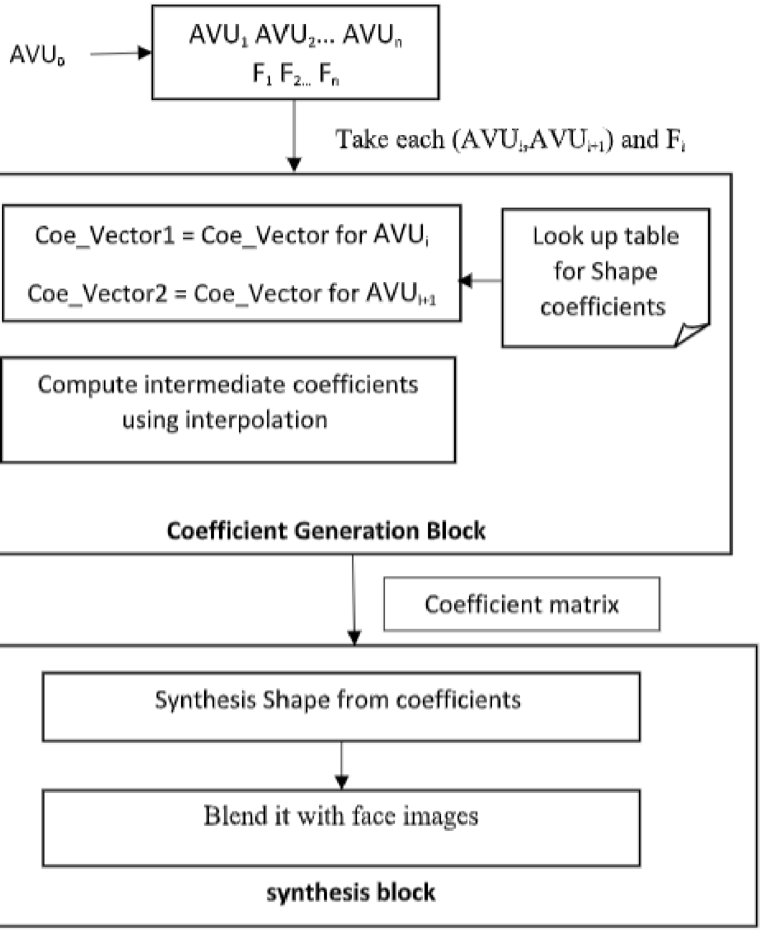

AVU0
AVU1 AVU2... AVUn
F1 F2... Fn
Take each (AVUi,AVUi+1) and Fi
Coe_Vector1 = Coe_Vector for AVUi
Coe_Vector2 = Coe_Vector for AVUi+1
Look up table for Shape coefficients
Compute intermediate coefficients using interpolation
Coefficient Generation Block
Coefficient matrix
Synthesis Shape from coefficients
Blend it with face images
synthesis block

Figure 7. 5: Block diagram of SSF framework

The sequence of operations performed in the proposed SSF model is explained bellow.

1. The phoneme sequence AVU_1 AVU_2... AVU_n and F_1, F_2...F_n , the number frames corresponding to each AVU (obtained from phase 2) are given as inputs to the SSF framework .

2. Coefficient generation block: The coefficient generation block works in an iterative fashion. The inputs (AVU_i, $AVUi_{+1}$) and F_i in a pass represent adjacent AVUs and the number frames for the corresponding transition respectively. The shape coefficients for AVU are taken from the look up table. Linear interpolation is used for generating the shape coefficients for intermediate frames. The increment in the model space is calculated as the difference between AVUs at the two ends, divided by the number of frames for the transition. AVU_0 is either the silence viseme or the last AVU of the preceding word.

3. Facial image synthesis block: The matrix of coefficients corresponding to the word is the input to this block. Shape is reconstructed from shape coefficients. The reconstructed shapes are combined with the background image after appropriate colouring and added to form the synthesised video corresponding to the given audio. Translation, rotation and scaling effects can be applied to the shape to suit different applications.

ii. Face Image Synthesis using Morphing In the Coefficient Space (SMCS)

Both shape and texture of frontal face images corresponding to the given can be synthesised using the SMCS framework depicted in figure 7.6. The basic flow is the same as that of the shape synthesiser.

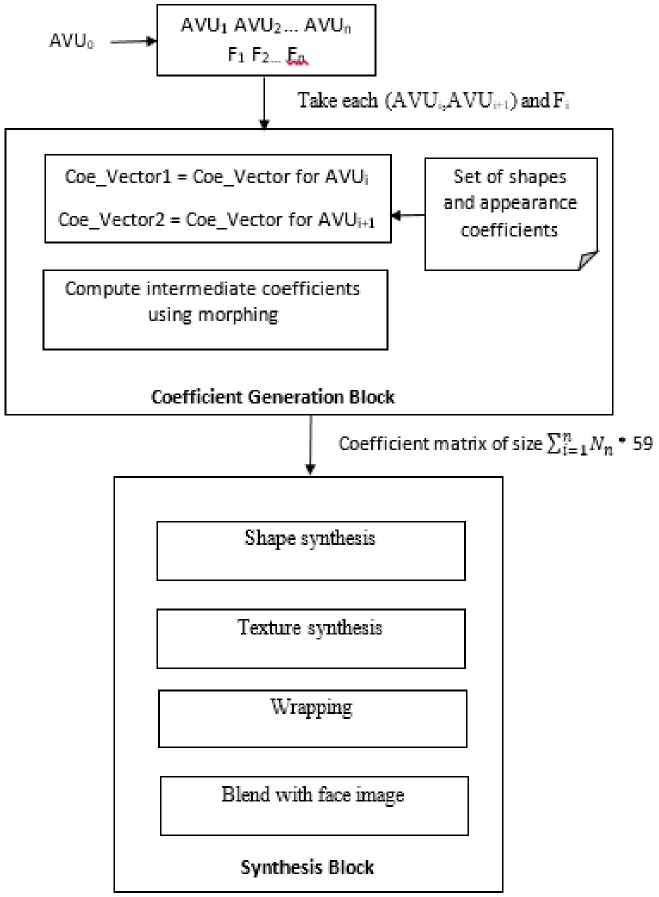

AVU₀
AVU₁ AVU₂... AVUₙ
F₁ F₂... Fₙ
Take each (AVUᵢ,AVUᵢ₊₁) and Fᵢ
Coe_Vector1 = Coe_Vector for AVUᵢ
Coe_Vector2 = Coe_Vector for AVUᵢ₊₁
Set of shapes and appearance coefficients
Compute intermediate coefficients using morphing
Coefficient Generation Block
Coefficient matrix of size $\sum_{i=1}^{n} N_n$ * 59
Shape synthesis
Texture synthesis
Wrapping
Blend with face image
Synthesis Block

Figure 7. 6: Block diagram of SMCS framework

The sequence of operations performed in the proposed SMCS model for facial image synthesis is listed below.

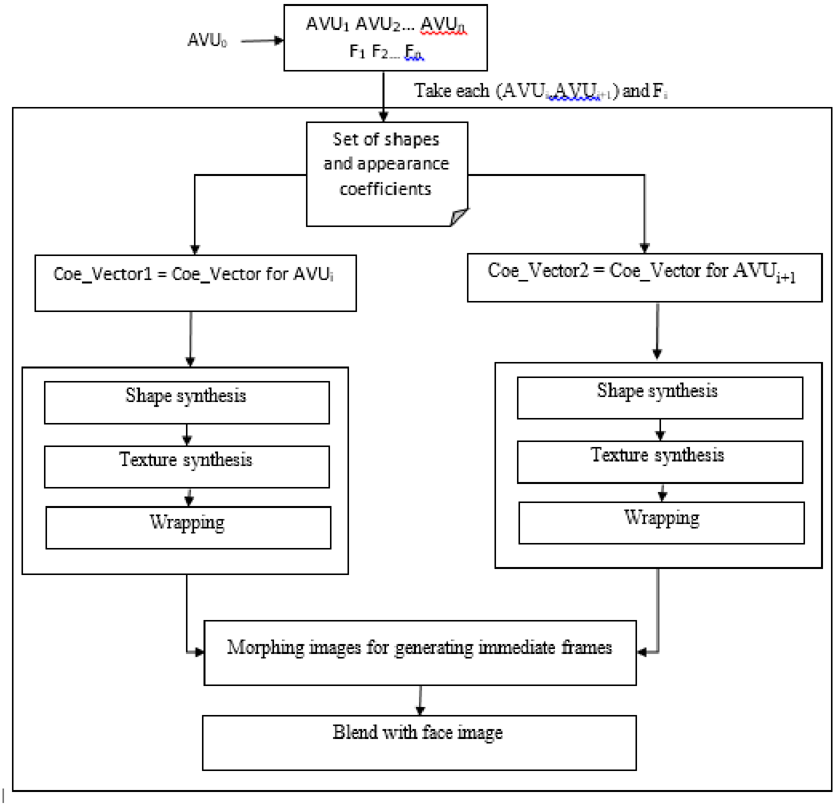
1. The phoneme sequence AVU_1 AVU_2... AVU_n and F_1, F2...F_n, the number frames corresponding to each phoneme are the inputs to the SMCS framework.

2. Coefficient generation block: The coefficient generation block works in an iterative fashion. The inputs (AVU_i,AVU_{i+1}) and F_i in a pass represents adjacent AVUs and the number frames for the corresponding transition respectively. The shape and appearance coefficients for AVU are taken from the look up table. The coefficients for intermediate frames are generated by morphing in the coefficient space. Linear, spline and cubic morphing methods are used and compared as part of the implementation of the framework. AVU_0 is either the silence viseme or the last AVU of the preceding word.

3. Facial image synthesis block: The matrix of shape and appearance coefficients corresponding to the input word is given as the input to this block. Shape is reconstructed from shape coefficients and texture for the mean shape is synthesised from the appearance coefficients. In the next step, the mean texture is wrapped to the target shape. The reconstructed images are combined with the background face image to generate the visual speech.

iii. Face Image Synthesis using Morphing in the Image Space (SMIS)

In this method morphing is performed in the image space, instead of the coefficient space as done in the SMCS method. The detailed flow of the SMCS framework is depicted in figure 7.7. The input and output for the SMIS module are the same as that of SMCS framework. For each iteration, the images corresponding to adjacent AVUs are synthesised from the shape and appearance coefficients. Intermediate frames are generated by morphing these images. Linear, spline and cubic morphing methods are used and compared as part of the implementation of the framework.



AVU0
AVU1 AVU2... AVUn
F1 F2... Fn
Take each (AVUi,AVUi+1) and Fi
Set of shapes and appearance coefficients
Coe_Vector1 = Coe_Vector for AVUi
Coe_Vector2 = Coe_Vector for AVUi+1
Shape synthesis
Texture synthesis
Wrapping
Shape synthesis
Texture synthesis
Wrapping
Morphing images for generating immediate frames
Blend with face image

Figure 7.7 : Block diagram for SMIS

The methods discussed so far can be used for the synthesis of frontal face talking images corresponding to a given grapheme sequence. In this work, the frameworks are applied for synthesising the lip movement sequence corresponding to a grapheme sequence. The details of experimental results using SSF,SMCS and SMIS frameworks for lip movement synthesis corresponding to Malayalam text are discussed in the next section in detail.

7.4 Experimental Results and Analysis on the implementation of SSF,SMCS and SMIS Frameworks for Lip Movement Synthesis

The framework is applied for generating lip movements corresponding to a given text. Speech animation is the task of moving the facial features of a graphics model to synchronize lip motion with the spoken audio and give the impression of speech production (301).Automatic Speech Animation has application for multimedia production, gaming, and low-bandwidth communication and in speech and language therapy. The synchronisation with incoming speech for lip movements is crucial as humans are sensitive to facial movements while talking. The frames in which a character speaks are generally taken as close up shots to develop an emotional attachment with the viewer. Hence a slightest glitch in the lip movements become noticeable and distracts the viewers from the movie. It is quite time consuming to create this demanding realism in speech animation. A plethora of techniques are used for speech animation in recent years.

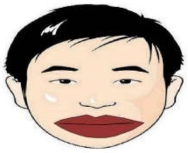
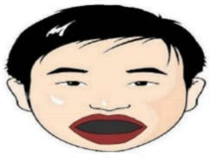
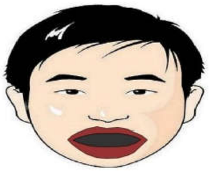
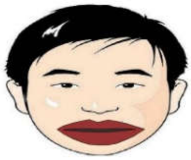
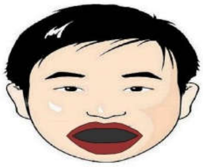
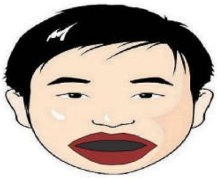
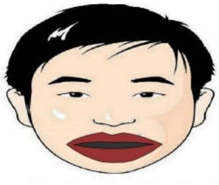
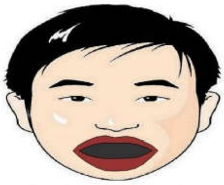
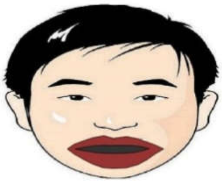
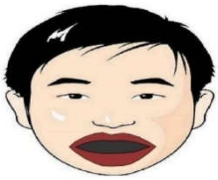
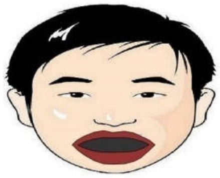
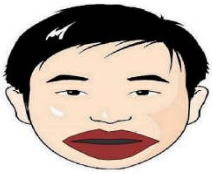
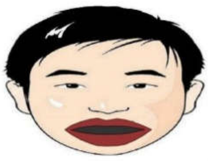
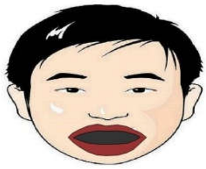
The simplest method, employed for low budget productions, is to keep a library of shapes corresponding to phonemes or visemes with an interpolation scheme for the generation of intermediate frames.

Multidimensional Morphable Model (MMM) is an example of this approach. Stitching real speech segments without interpolation is another popular low cost method. Performance driven animation with sophisticated post production work is used for high budget productions. Here the articulatory movements of a human actor are captured and transferred to a graphic model. Even though it lacks flexibility the model can incorporate the emotional nuances and subtleties of the human actor [37]. In this work SSF, SMCS and SMIS frameworks are used for lip movement synthesis. After performing independent AAM modelling each instance of the lip can be considered as a point in the 59 dimensional model space, comprising of 9 shape coefficients and 50 texture coefficients. A point in the model space corresponds to a particular lip shape and texture attained by the lip while talking. An utterance of a word can be generated as a trajectory in the model space using SSF, SMCS and SMIS frameworks. Synthesis experiments are conducted with all the 3 models. The results of experiments and its analysis are discussed in the following sections in detail.

7.4.1 Experimental Results for Lip Movement Synthesis using SSF Frameworks

The sequence of lip shapes corresponding to a text are synthesised using SSF framework. Table 7.3 shows the image sequence for the initial phonemes of the Malayalam word പകുതി/pakuthi/ (with sequence of phonemes as /പ്/+ /അ/+ /ക്/+ /ഉ/+/ത്/+/ഇ/) of duration 0.89 seconds with frames 25 per second.

Table 7.3: Synthesised images obtained for the initial 3 phonemes of Malayalam word /pakuthi/ using SSF based framework

Phonee	പ - അ Transition	അ- ക Transition	ക - ഉ Transition
Frame 1			
Frame 2			
Frame 3			
Frame 4			
Frame 5			

A simple colouring scheme, with separate colours for lip and inner mouth region are used in the implementation. Lip motion image sequences are generated using phoneme based, allophone based and viseme based models.

7.4.2 Experimental Results using SMCS Frameworks for Lip Movement Synthesis

This section discusses the the experimental results obtained on the generation of the lip images corresponding to most frequently occurring consonant-vowel combinations in Malayalam using SMCS framework. The sequence of lip images generated using allophone, viseme from allophone, phoneme and visme from phoneme as input to SMCS framework are shown in figure 7.8 to figure 7.11.

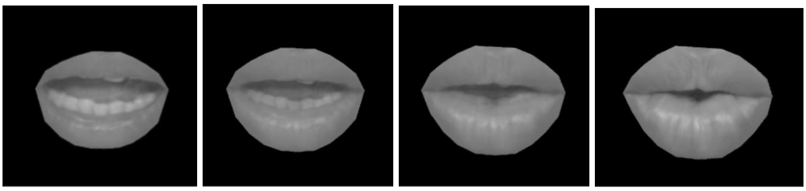


Figure 7.8 : Images corresponding to ക2 /ka/ to ഉ1/u/ transition generated with allophone based lip motion synthesis using SMCS

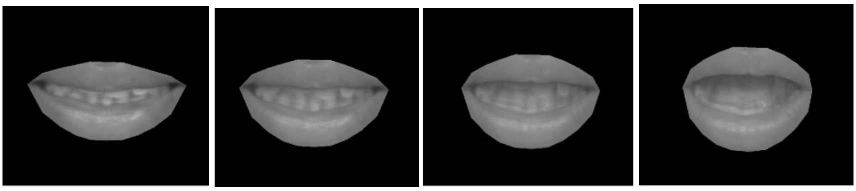


Figure 7.9 : Images corresponding ച /cha/ - ഔ /ou/ transition generated with viseme (from allophone) based lip motion synthesis using SMCS

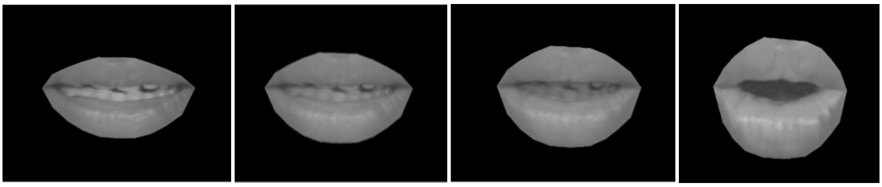


Figure 7.10 : Images corresponding to ട /ta/ - ഒ/o/ transition generated with phoneme based lip motion synthesis using SMCS

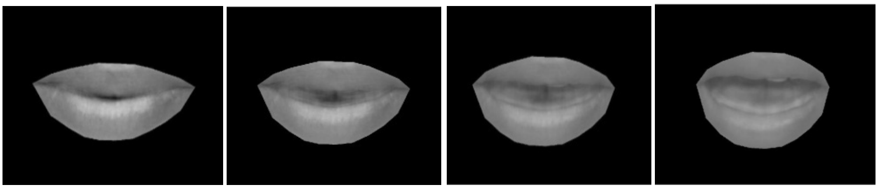


Figure 7.11: Images corresponding to പ /pa/ - അ /a/ transition generated with viseme from phoneme based lip motion synthesis using SMCS

Experiments are repeated by changing the method of morphing as linear, spline and cubic morphing. But these change in morphing methods is observed to be not making any significant difference in the synthesised image.

7.4.3 Experimental Results using SMIS Frameworks for Lip Movement Synthesis

This section discusses the the experimental results obtained on the generation of the lip images corresponding to most frequently occurring consonant-vowel combinations in Malayalam using SMIS framework. Figure 7.12 shows intermediate frames generated during the transition ത/tha/ to ഇ/e/.

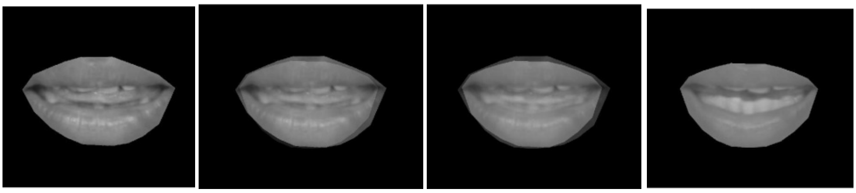


Figure 7.12 : Images corresponding to ത /tha/ - ഇ/e/ transition generated with viseme from phoneme based lip motion synthesis using SMIS

But morphing in the image space creates unrealsitic textures. Example intermediate frames generated during ക/ka/ to ഉ/u/, ച /cha/ - ഔ /ou/ consonant vowel transitions are depicted in figure 7.13.

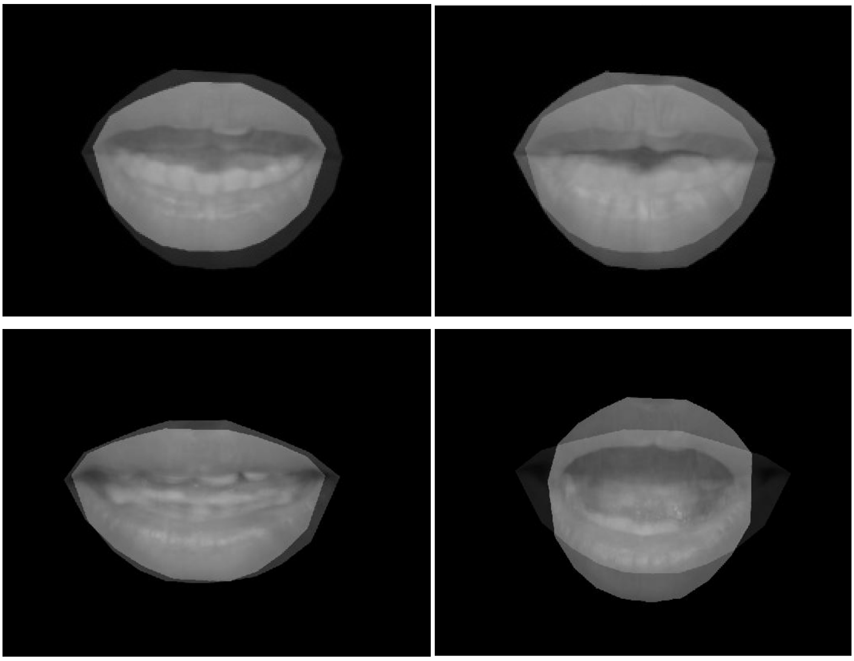


Figure 7. 13: Examples of intermediate unrealistic frames generated using SMIS lip motion synthesis framework

It can be observed that SMCS framework is generating realistic intermediate frames. Considering the advantage of SMCS framework in synthesising realistic textures, SMCS is used for the conduct of further experiments. Perception experiments with SMCS framework based lip image sequences are conducted for evaluating allophone, phoneme and viseme based AAM models, which is discussed in the next section.

7.4.4 Perception Experiments Based Evaluation of SMCS Framework for Lip Motion Synthesis

Set of 35 Malayalam words are selected for conducting experiments for the performance evaluation of different lip movement synthesis models. The set of words selected for the study, which are phoneme and allophone rich is listed in table 7.4.

Table 7.4: List of Malayalam words used for the conduct of perception experiments with corresponding allophonic transcription

Sl. No.	Word	Coresponding sequence of allophone
1	വഹ്നി	വ്1 അ2 ഹ്2 ന്11 ഇ3
2	ഖജനാവ്	ഖ്2 അ2 ജ്2 അ2 ന്2 ആ2 വ്2 ഉ4
3	ഈനാംപേച്ചി	ഈ1 ന്12 ആ2 മ്4 പ്3 ഏ3 ച്4 ഇ3
4	അനാച്ഛാദനം	അ1 ന്12 ആ2 ഛ്2 ആ2 ദ്2 അ2 ന്12 അ2 മ്4
5	കുഷ്ഠം	ക്1 ഉ7 ഷ്1 ഠ്2 അ2 മ്4
6	ഛത്രകം	ഛ1 അ2 ത്2 റ്1 അ2 ക്3 അ2 മ്4
7	മഠയൻ	മ്1 അ2 ഠ1 അ2 യ്1 അ2 ന്12
8	അഗ്നിഭൈരവൻ	അ1 ഗ്1 ന്11 ഇ1 ഭ1 ഐ1 ര്1 അ2 വ്1 അ2 ന്12
9	ഉന്ദുരുകർണിക	ഉ1 ന്12 ദ്2 ഉ3 ര്1 ഉ3 ക്3 അ2 ര്1 ണ്1 ഇ1 ക്3 അ2
10	കദംബപുഷ്പ	ക്1 അ2 ദ്2 അ2 മ്4 ബ്1 അ2 പ്2 ഉ3 ഷ്1 പ്4 അ2
11	കേരളത്തിലെ	ക്1 ഏ3 ര്1 അ2 ള്1 അ2 ത്2 ത്2 ഇ1 ല്1 എ2

12	ഒന്നരനൂറ്റാണ്ടുമുമ്പാണ്	ഒ1 ന്12 ന്12 അ2 ര്1 അ2 ന്12 ഊ1 റ്റ്2 ആ2 ണ്1 ട്1 ഉ3 മ്4 ഉ3 മ്4 പ്3 ആ2 ണ1 ഉ4
13	പൗരാവകാശചരിത്രം	പ്1 ഔ1 ര്1 ആ2 വ്1 അ2 ക്1 ആ1 ശ്1 അ2 ച്2 അ2 ര്1 ഇ1 ത്2 റ്1 അ2 മ്4
14	എഴുതപ്പെട്ടിരിക്കുന്ന	എ1 ഴ്1 ഉ3 ത്3 അ2 പ്4 പ്4 എ3 ട്4 ട്4 ഇ1 ര്1 ഇ1 ക്2 ക്2 ഉ3 ന്12 ന്12 അ2
15	ഇതുപോലെ	ഇ2 ത്2 ഉ3 പ്4 ഓ2 ല്1 എ2
16	ഉത്ഥാനവീരൻ	ഉ1 ത്2 ഥ്1 ആ2 ന്2 അ2 വ്1 ഈ2 ര്1 അ2 ന്12
17	തനിയേ	ത്1 അ2 ന്2 ഇ1 യ്1 ഏ2
18	ഏധിത	ഏ1 ധ്1 ഇ1 ത്3 അ2
19	ഓന്തുകൊത്തി	ഓ1 ന്12 ത്4 ഉ3 ക്3 ഒ2 ത്2 ത്2 ഇ3
20	ജലവൈദ്യുതി	ജ്1 അ2 ല്1 അ2 വ്1 ഐ1 ദ്1 യ്1 ഉ3 ത്3 ഇ3
21	ഡഫേദാർ	ഡ്1 അ2 ഫ1 ഏ3 ദ്2 ആ2 ര്1
22	ഝങ്കാരം	ഝ1 അ2 ങ്4 ക്4 ആ1 ര്1 അ2 മ്4
23	സുഖം	സ്1 ഉ7 ഖ്1 അ2 മ്4
24	തമാഖു	ത്1 അ2 മ്4 ആ2 ഖ്3 ഉ2
25	നക്ഷത്രചിഹ്നം	ന്2 അ2 ക്5 ഷ്1 അ2 ത്2 റ്1 അ2 ച്2 ഇ1 ഹ്2 ന്11 അ2 മ്4
26	നേർന്നുകെട്ടുക	ന്2 ഏ3 ര്1 ന്1 ന്1 ഉ3 ക1 എ3 ട്4 ട്4 ഉ3 ക്3 അ2
27	ഗൂഢാങ്ഘ്രലി	ഗ്2 ഊ1 ഢ്1 ആ2 മ്4 ങ്1 ഘ്1 അ2 ര്1 ഇ3
28	ചടകാമുഖം	ച്1 അ2 ട്2 അ2 ക്1 ആ1 മ്4 ഉ3 ഖ്1 അ2 മ്4
29	സംവരണം	സ്1 അ2 മ്3 വ്1 അ2 ര്1 അ2 ണ്1 അ2 മ്4
30	ചിരസഞ്ചിത	ച്1 ഇ1 ര്1 അ2 സ്1 അ2 ഞ്1 ച്3 ഇ1 ത്3 അ2

31	പടവലങ്ങ	പ്1 അ2 ട്2 അ2 വ്1 അ2 ല്1 അ2 ങ്1 ങ്1 അ2
32	ചെമ്പൊൽത്താർബാണൻ	ച്1 എ3 മ്4 പ്3 ഒ2 ല്1 ത്2 ത്2 ആ2 ര്1 ബ്1 ആ2 ണ്1 അ2 ന്12
33	ടോട	ട്1 ഓ2 ട്3 അ2
34	അവന്റെ	അ1 വ്1 അ2 ന്12 റ്റ്1 എ2
35	ചെയ്യപ്പെട്ട	ച്1 ഐ2 യ്1 അ2 പ്4 പ്4 എ3 ട്4 ട്4 അ2

Lip movement synthesis is performed from phoneme, allophone, viseme (mapped from phoneme) and viseme (mapped from allophone) based models using SMCS framework. Perception experiments are conducted to evaluate the synthesised images. For evaluation, the scheme proposed by Likert [334], in which the participants are asked to rate the naturalness in a 5 point scale is adopted. The scaling starts with 1(very unnatural) and ends with 5 (very natural) [335].

Three speakers with normal hearing and seeing conditions are selected as participants for the experiment. The synthesised image sequences corresponding to the 35 words using SCMS is displayed to the participant with audio from the original utterances. Four set of visuals is generated and displayed for each of the word with allophone, viseme(from allophone), phoneme and viseme(from phoneme) as basic AVUs used for AAM training. Each user marks one of the five options for rating the synthesised image corresponding to each word in each of the four modes. Options are provided for recording the rating.The sum of scaled responses for all words is used for preparing evaluation metric. A consolidated report of the performance evaluation experiments is given in table 7.5

Table 7.5: Consolidated report of responses evaluating SMCS synthesised lip images with different AVUs of 3 participants

Atomic Visual Speech Unit	**Participant 1**	**Participant 2**	**Participant 3**	**Average**
Synthesiser with allophone as basic unit	116	112	126	118
Synthesiser with Viseme obtained through mapping from Allophone as basic unit	110	108	120	112.67
Synthesiser with Phoneme as basic unit	100	96	100	98.67
Synthesiser with Viseme obtained through mapping from phonemes as basic unit	96	94	98	96

From the experimental results, it is evident that the allophone based approach for lip movement synthesis is found to be more promising. The difference between the allophone based and viseme (mapped from allophone) based approach is observed as very minimal. Similarly the difference between performance of the phoneme based and viseme (formed from phoneme) based approach is also found very minimal.

7.5 Conclusion

This chapter describes a novel comprehensive framework developed for visual speech synthesis for Malayalam. The components such as durational models, phoneme and allophone transcripters, viseme set generated using linguistic knowledge, perception experiments and data driven methods and lip movement synthesis algorithms are combined to form an efficient visual speech synthesiser. Independent AAM is used for the parametric modelling of visual speech. In this work the proposed framework is successfully applied for generating talking lip movements for Malayalam. Three separate models are implemented for lip movement synthesis. The SSF framework synthesises shape, whereas SMCS and SMIS frameworks synthesises both shape and texture. Perception experiments are conducted by taking allophone, phoneme, viseme (mapped from allophone) and viseme (mapped from phoneme) as inputs to the synthesiser. Perception based evaluation revealed the advantages of using allophone as the basic unit for lip synthesis applications. From the results of the perception experiments it is also evident that significant perception difference is not observed while using viseme instead of underlying phonemes or allophones as Atomic Visual Unit (AVUs).

Chapter 8
CONCLUSION

"Speech is rather a set of movements made audible than a set of sounds produced by movements"

8.1 Conclusion

Inputs from multiple channels, which are integrated in the brain makes human-human communication easy, accurate and robust. Incorporating this multimodal information integration capability of the brain in to machines is expected to revolutionise human computer interaction. The perception of speech is multimodal, even though speech perception is primarily based on the audio cues generated from the speech signal. But the importance of visual cues in the perception of speech and in the communication of non-linguistic factors such as emotional state is an established fact. The emphasis and punctuations added by facial expressions is one of the decisive factors in making face to face talking the most effective communication mechanism. The frequently quoted example is of a raised eyebrow, which acts as question mark. Blind people understand speech by touch. Taste and smell are not in anyway involved with speech. This thesis explored the visual aspects of speech for realising a visual speech synthesis framework in Malayalam.

Text to audio visual speech synthesis systems need to understand text, audio and visual domains of language representations.

Standardisations, mappings from one domain to another and statistical analysis based on atomic units in each domain need to be carried out for the target language. Graphemes are the basic unit in the textual representation of a language. Phonemes, defined as the smallest distinctive sound units in a language, are considered to be the basic unit for speech. But the properties of phonemes exhibit wide variations based on its position in the word and context. In Malayalam, phonemes are further categorised in to allophones based on the positional and contextual variability, *i.e.* the contextual and positional variability is encoded in the allophone characterisation of Malayalam language. This computational linguistic feature of Malayalam is exploited in designing many components of the proposed work. Malayalam phoneme set consists of 10 monophthongs, 2 diphthongs and 38 consonants. As part of the study, 107 allophones in Malayalam which include 76 consonant allophones, 28 vowel allophones and 3 allophones corresponding to diphthongs are identified and listed. Corpora in text, audio and visual modalities are created for the training and the conduct of various experiments related to tracking and synthesis.

Detailed investigation is performed on the duration of allophonic variations of Malayalam allophones, which is used as the durational model of visual speech synthesis framework. Further investigations are performed to understand the audio visual asynchrony in Malayalam. It is observed that the interphoneme silence in Malayalam is mainly attributed to the stop phase or obstruction phase of plosive formation. A rule based text to phoneme and allophone transcription system is developed as part of this study. The rule based approach gives satisfactory results in Malayalam transcription. Pre-processing brings different class of graphemes to a unified framework required for the phoneme and allophone transcription. A comprehensive statistical and probabilistic analysis based on the developed phoneme and allophone converter is performed on standard Malayalam word corpora.

Viseme set is the set of visually discernable mouth appearances, which is considered as the basic unit of visual speech. Separate viseme sets for Malayalam language are developed as part of this work based on phoneme to viseme mappings and allophone to viseme mappings. Phoneme to Viseme maps are developed from linguistic knowledge, perception experiments and data driven clustering methods. Geometric features of lips and Discrete Cosine Transform (DCT) are the visual features used for clustering. The final phoneme to viseme map, with 17 members is finalised by comparing viseme sets obtained using these three approaches.

Mouth area mainly consists of lip, skin neighbourhood of lips, teeth, tongue and dark portions of mouth cavity. The colour of these mouth regions shows significant variation with respect to ethnicity. The statistical and probabilistic analysis of colour of mouth region pixels in Indian context is performed as part of this study. The analysis is performed in different colour spaces. The colour spaces are ranked according to their performances against a multiclass Bayesian classifier. Lip tracking by land mark point localisation is another major problem Active Shape Modelling (ASM) and Convolution Neural Networks (CNN) are used for lip land mark point localisation. The ranking of colour spaces is exploited in designing the ASM based lip tracking system.

Finally a text to visual speech synthesis frame work using independent Active Appearance Models (AAM) is implemented. The lip movements corresponding to input text is synthesised using the framework. The frame work comprises of durational models, transcriptors, viseme set and lip tracking modules to develop the AAM based synthesiser.

8.2 Contributions of the thesis

The main contributions of the work are listed, which can be summarised in the following way.

i. The development of an isolated phoneme and isolated word audio visual speech corpora in Malayalam. Image sequences corresponding to speech utterances of 51 phonemes (from 10 speakers) and speech utterances of 214 words (from 10 speakers) are procured and arranged in a web framework for easy navigation and access. The corpora consist of around 13,000 frontal face images. The inner and outer lip contours are manually marked using 36 land mark points for a subset of images in the corpora. The corpora, the first of its kind in Malayalam, can be used for further research leading to the development of successful audio visual speech synthesis and recognition tools in Malayalam.

ii. An allophone based durational model, which accommodates the contextual variability of Malayalam phonemes in continuous speech is developed as part of this work. The preliminary studies performed revealed the nature of audio visual asynchrony in Malayalam.

iii. A grapheme to phoneme and allophone transcripter for Malayalam is developed as part of this work. The developed transcripter carries the potential for the development of language computing tools in Malayalam, which considers the inherent computational linguistic features of Malayalam language.

iv. The phoneme to viseme mappings generated in Malayalam using linguistic, perception and data driven approaches are the other important contribution of this work. The proposed viseme set will act as a foundation for future studies in the domain of Malayalam visual speech processing. The allophone to viseme maps generated is a catalyst towards allophone centric paradigm shift in Malayalam speech processing.

v. The ranking of colour components in different colour spaces for classification of mouth region pixels in to skin, lip, teeth and tongue is another contribution of the work. Considering the dearth of such studies in Indian skin tone context, the ranking can be used for developing native mouth motion analysis and tracking systems. A Hue (HSI colour model) based ASM lip segmentation framework is also proposed as part of this work. Works performing lip segmentation and tracking using deep neural network are rare in literature. A Convolution Neural Network based lip tracking system is also developed by landmark localisation and tested with MAVSC-IP and MAVSC-IW data sets.

vi. The thesis proposes a unified frame work for visual speech synthesis using independent AAM in Malayalam, using the components developed as part of the work. A lip motion animation synthesiser with grapheme sequence as input for Malayalam is developed and evaluated as part of this work. The reliable visual speech synthesisers in Malayalam for applications such as Human Computer Interaction (HCI), assistive technology for deaf community and talking face models may evolve in the public domain with the proposed framework as starting point.

8.3 Future Directions

More investigations are needed for exactly modelling audio visual asynchrony in Malayalam. A speech to phoneme and allophone transcriptor for Malayalam is required for developing speech driven facial motion synthesis systems. Extracting emotional cues from speech and creating expressive visual speech is also an important future direction. The statistical properties of phonemes and allophones in Malayalam derived using transcriptors can be used for language computing applications such as language identification.

Viseme is the set of visually separable units. The viseme set in Malayalam is developed using various methods in this work and the same is used as a component in the visual speech synthesis framework. The concept of visual severability is very important for developing visual speech recognisers in a language. The phoneme/allophone to viseme maps developed in this work, especially using data driven techniques can be incorporated in visual speech recognition systems in Malayalam. The visually separable phoneme classes can be validated using classifiers such as artificial neural networks.

A Performance driven facial animation systems has been used for many successful productions. Earlier approaches used markers and sensors for tracking target regions, which is retargeted to a synthesis framework. But the current practice is to use sophisticated algorithms for tracking regions and to use them for real time modelling and synthesis of facial movements. The detailed study performed on colour properties of mouth region pixels in Indian context can be used for designing real time tracking systems in an analysis synthesis framework.

A speech animation system in Malayalam can be developed for the character animation using the framework proposed in the work. In this work, the synthesiser framework is applied for creating lip movements corresponding to a text. A complete facial motion synthesis system can also be developed using the same framework. An image dataset with land marking on full face is the only additional requirement for developing facial motion synthesiser in Malayalam. Speech driven visual speech synthesisers can be developed using the AAM as part of future research. Deep Neural Network (DNN) based visual speech synthesisers trained with AAM coefficients is another possible direction for future implementations.

REFERENCES

[1] P. Kozierski, T. Sadalla, S. Drgas, and A. Dąbrowski, “Allophones in automatic whispery speech recognition,” in *2016 21st International Conference on Methods and Models in Automation and Robotics (MMAR)*, 2016, pp. 811–815.

[2] L. Nguyen, X. Guo, and J. Makhoul, “Modeling Frequent Allophones in Japanese Speech Recognition,” in *Seventh International Conference on Spoken Language Processing*, 2002.

[3] J. Xu, Y. Si, J. Pan, and Y. Yan, “Automatic allophone deriving for Korean speech recognition,” in *2013 Ninth International Conference on Computational Intelligence and Security*, 2013, pp. 776–779.

[4] F. Imedjdouben and A. Houacine, “Generation of allophones for speech synthesis dedicated to the Arabic language,” in *2015 First International Conference on New Technologies of Information and Communication (NTIC)*, 2015, pp. 1–4.

[5] P. Skrelin, “Allophone-and suballophone-based speech synthesis system for Russian,” in *International Workshop on Text, Speech and Dialogue*, 2000, pp. 271–276.

[6] W. Barkhoda, B. ZahirAzami, A. Bahrampour, and O.-K. Shahryari, “A comparison between allophone, syllable, and diphone based TTS systems for Kurdish language,” in *2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2009, pp. 557–562.

[7] G. C. Cawley and P. D. Noakes, “Allophone synthesis using a neural network,” in *Proceedings of the First World Congress on Neural Networks (WCNN-93)(2)*, 1993, pp. 122–125.

[8] M. Hamad and M. Hussain, “Arabic text-to-speech synthesizer,” in *2011 IEEE Student Conference on Research and Development*, 2011, pp. 409–414.

[9] A. Karpov, L. Tsirulnik, Z. Krňoul, A. Ronzhin, B. Lobanov, and M. Železný, “Audio-visual speech asynchrony modeling in a talking head,” in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[10] K.-H. Wong, W.-K. Lo, and H. Meng, “Allophonic variations in visual speech synthesis for corrective feedback in capt,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5708–5711.

[11] G. A. Kalberer, P. Müller, and L. J. Van Gool, “Speech Animation Using Viseme Space.,” in *VMV*, 2002, pp. 463–470.

[12] G. A. Kalberer, P. Müller, and L. J. Van Gool, “Generating Visemes for Realistic Animation.,” in *VMV*, 2002, pp. 233–240.

[13] O. Sayli, “Duration analysis and modeling for Turkish text-to-speech synthesis,” *Master’s thesis, Dep. Electr. Electron. Eng. Bogaziei Univ.*, 2002.

[14] R. Batusek, “A duration model for Czech text-to-speech synthesis,” in *Speech Prosody 2002, International Conference*, 2002.

[15] A. Lazaridis, P. Zervas, N. Fakotakis, and G. Kokkinakis, “A CART approach for duration modeling of Greek phonemes,” in *Proc. of SPECOM*, 2007, pp. 287–292.

[16] B. Chen, T. Bian, and K. Yu, “Discrete Duration Model for Speech Synthesis.,” in *INTERSPEECH*, 2017, pp. 789–793.

[17] Y. Hifny and M. Rashwan, “Duration modeling for arabic text to speech synthesis,” in *Seventh International Conference on Spoken Language Processing*, 2002.

[18] G. Norkevičius, G. Raškinis, and A. Kazlauskienė, “Knowledge-based grapheme-to-phoneme conversion of Lithuanian words,” in *SPECOM 2005, 10th International Conference Speech and Computer*, 2005, pp. 235–238.

[19] J. Pylkkonen and M. Kurimo, “Duration modeling techniques for continuous speech recognition,” in *Eighth International Conference on Spoken Language Processing*, 2004.

[20] Y. Demeke and S. Hailemariam, “Duration modeling of phonemes for Amharic text to speech system,” in *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, 2012, pp. 1–7.

[21] K. U. Ogbureke, J. P. Cabral, and J. Carson-Berndsen, “Explicit duration modelling in HMM-based speech synthesis using a hybrid hidden Markov model-Multilayer Perceptron,” in *SAPA-SCALE Conference*, 2012.

[22] S. Sovilj-Nikic, V. Delic, I. Sovilj-Nikic, and M. Markovic, “Tree-based phone duration modelling of the Serbian language,” *Elektron. ir Elektrotechnika*, vol. 20, no. 3, pp. 77–83, 2014.

[23] K. Samudravijaya, “Durational characteristics of Hindi phonemes in Continuous Speech,” *Tech. Report, TIFR*, 2003.

[24] K. Samudravijaya, “Durational characteristics of Hindi stop consonants,” in *Eighth European Conference on Speech Communication and Technology*, 2003.

[25] K. S. Rao and B. Yegnanarayana, “Modeling syllable duration in Indian languages using neural networks,” in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, vol. 5, p. V-313.

[26] K. S. Rao and B. Yegnanarayana, “Modeling syllable duration in indian languages using support vector machines,” in *Proceedings of 2005 International Conference on Intelligent Sensing and Information Processing, 2005.*, 2005, pp. 258–263.

[27] N. S. Krishna and H. A. Murthy, “Duration modeling of Indian languages Hindi and Telugu,” in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[28] L. Jahan, U. Kulsum, and A. Naser, “Bengali Diphone Duration Modeling for Bengali Text to Speech Synthesis System,” *Am. Acad. Sch. Res. J.*, vol. 7, no. 3, 2015.

[29] D. Govind, S. Mahanta, and S. R. M. Prasanna, “Significance of Duration in the Prosodic Analysis of Assamese,” in *Speech Prosody 2012*, 2012.

[30] S. Roy and N. Sinha, “Duration Modeling in Hindi,” *Int. J. Comput. Appl.*, vol. 97, no. 6, 2014.

[31] V. S. R. Bonda and P. N. Girija, “Duration Modeling For Telugu Language with Recurrent Neural Network.”

[32] B. L. Kanth, V. Keri, and K. S. Prahallad, “Durational characteristics of Indian phonemes for language discrimination,” in *International Conference on Information Systems for Indian Languages*, 2011, pp. 130–135.

[33] D. P. Gopinath, J. D. Sree, R. Mathew, S. J. Rekhila, and A. S. Nair, “Duration analysis for malayalam text-to-speech systems,” in *9th International Conference on Information Technology (ICIT’06)*, 2006, pp. 129–132.

[34] D. P. Gopinath, S. G. Veena, and A. S. Nair, “Modeling of Vowel Duration in Malayalam Speech using Probability Distribution,” *ISCA Arch. Speech Prosody, Campinas, Brazil*, pp. 6–9, 2008.

[35] J. R. Movellan, “Visual speech recognition with stochastic networks,” in *Advances in neural information processing systems*, 1995, pp. 851–858.

[36] C. C. Chibelushi, S. Gandon, J. S. D. Mason, F. Deravi, and R. D. Johnston, “Design issues for a digital audio-visual integrated database,” 1996.

[37] S. Pigeon and L. Vandendorpe, “The M2VTS multimodal face database (release 1.00),” in *International Conference on Audio- and Video-Based Biometric Person Authentication*, 1997, pp. 403–409.

[38] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, “XM2VTSDB: The extended M2VTS database,” in *Second international conference on audio and video-based biometric person authentication*, 1999, vol. 964, pp. 965–966.

[39] T. Chen, “Audiovisual speech processing,” *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 9–21, 2001.

[40] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, “Extraction of visual features for lipreading,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, 2002.

[41] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, “CUAVE: A new audio-visual database for multimodal human-computer interface research,” in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 2, p. II-2017.

[42] C. Sanderson and K. K. Paliwal, “The vidtimit database,” *Idiap Commun.*, pp. 2–6, 2002.

[43] E. Bailly-Bailliére *et al.*, “The BANCA database and evaluation protocol,” in *International conference on Audio-and video-based biometric person authentication*, 2003, pp. 625–638.

[44] B. Lee *et al.*, “AVICAR: Audio-visual speech corpus in a car environment,” in *Eighth International Conference on Spoken Language Processing*, 2004.

[45] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, “A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments,” in *Proceedings of the 6th international conference on Multimodal interfaces*, 2004, pp. 235–242.

[46] N. A. Fox, B. A. O’Mullane, and R. B. Reilly, “VALID: A new practical audio-visual database, and comparative results,” in *International Conference on Audio-and Video-Based Biometric Person Authentication*, 2005, pp. 777–786.

[47] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[48] S. J. Cox, R. W. Harvey, Y. Lan, J. L. Newman, and B.-J. Theobald, “The challenge of multispeaker lip-reading.,” in *AVSP*, 2008, pp. 179–184.

[49] J. Trojanová, M. Hrúz, P. Campr, and M. Železný, “Design and recording of czech audio-visual database with impaired conditions for continuous speech recognition,” in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, 2008.

[50] D. Petrovska-Delacrétaz *et al.*, “The IV 2 Multimodal Biometric Database (Including Iris, 2D, 3D, Stereoscopic, and Talking Face Data), and the IV 2-2007 Evaluation Campaign,” in *2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems*, 2008, pp. 1–7.

[51] D. Teferi and J. Bigun, “Evaluation protocol for the dxm2vts database and performance comparison of face detection and face tracking on video,” in *2008 19th International Conference on Pattern Recognition*, 2008, pp. 1–4.

[52] G. Zhao, M. Barnard, and M. Pietikainen, “Lipreading with local spatiotemporal descriptors,” *IEEE Trans. Multimed.*, vol. 11, no. 7, pp. 1254–1265, 2009.

[53] A. Vorwerk, X. Wang, D. Kolossa, S. Zeiler, and R. Orglmeister, “WAPUSK20-A Database for Robust Audiovisual Speech Recognition.,” in *LREC*, 2010.

[54] Y. A. El-Imam, “Phonetization of Arabic: rules and algorithms,” *Comput. Speech Lang.*, vol. 18, no. 4, pp. 339–373, 2004.

[55] A. B. Mosaddeque, N. UzZaman, and M. Khan, “Rule based automated pronunciation generator,” 2006.

[56] M. Divay and M. Guyomard, “Grapheme-to-phoneme transcription for French,” in *ICASSP’77. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1977, vol. 2, pp. 575–578.

[57] J. Halpern, “The Role of Phonetics and Phonetic Databases in Japanese Speech Technology.”

[58] J. Domokos, O. Buza, and G. Toderean, “Romanian phonetic transcription dictionary for speeding up language technology development,” *Lang. Resour. Eval.*, vol. 49, no. 2, pp. 311–325, 2015.

[59] A. Novák and B. Siklósi, "Grapheme-to-Phoneme Transcription in Hungarian.," *Int. J. Comput. Linguist. Appl.*, vol. 7, no. 1, pp. 161–173, 2016.

[60] Y. K. Thu, W. P. Pa, Y. Sagisaka, and N. Iwahashi, "Comparison of grapheme-to-phoneme conversion methods on a myanmar pronunciation dictionary," in *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, 2016, pp. 11–22.

[61] W. Aroonmanakun, N. Thapthong, P. Wattuya, B. Kasisopa, and S. Luksaneeyanawin, "Generating Thai Transcriptions for English Words," *SEALS XIV*, p. 13, 2004.

[62] A. Van Den Bosch and W. Daelemans, "Do not forget: Full memory in memory-based learning of word pronunciation," in *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, 1998, pp. 195–204.

[63] B. Kim, G. G. Lee, and J.-H. Lee, "Morpheme-based grapheme to phoneme conversion using phonetic patterns and morphophonemic connectivity information," *ACM Trans. Asian Lang. Inf. Process.*, vol. 1, no. 1, pp. 65–82, 2002.

[64] K. Nahar, H. Al-Muhtaseb, W. Al-Khatib, M. Elshafei, and M. Alghamdi, "Arabic Phonemes Transcription using Data Driven Approach.," *Int. Arab J. Inf. Technol.*, vol. 12, no. 3, 2015.

[65] M.-S. Liang, R.-Y. Lyu, and Y.-C. Chiang, "Data Driven Approaches to Phonetic Transcription with Integration of Automatic Speech Recognition and Grapheme-to-Phoneme for Spoken Buddhist Sutra," *Int. J. Comput. Linguist. Chinese Lang. Process. Vol. 13, Number 2, June 2008*, vol. 13, no. 2, pp. 233–254, 2008.

[66] C. Leitner, *Data-based automatic phonetic transcription*. na, 2008.

[67] J. Basu, T. Basu, M. Mitra, and S. K. Das Mandal, "Grapheme to Phoneme (G2P) conversion for Bangla," in *2009 Oriental COCOSDA International Conference on Speech Database and*

Assessments, 2009, pp. 66–71.

[68] S. A. Chowdhury, F. Alam, N. Khan, and S. R. H. Noori, “Bangla grapheme to phoneme conversion using conditional random fields,” in *2017 20th International Conference of Computer and Information Technology (ICCIT)*, 2017, pp. 1–6.

[69] R. R. Kiran, S. B. S. Kumar, K. E. Manjunath, B. Satapathy, A. Chaturvedi, and D. Pati, “Automatic phonetic and prosodic transcription for Indian languages: Bengali and Odia,” in *Tenth International Corference on Natural Language Processing*, 2013.

[70] A. V. Vidyapeetham and C. Ettimadai, “Grapheme to phone conversion for Hindi.”

[71] M. Choudhury, “Rule-based grapheme to phoneme mapping for hindi speech synthesis,” in *90th Indian Science Congress of the International Speech Communication Association (ISCA), Bangalore, India*, 2003.

[72] S. Chaware and S. Rao, “Rule-based phonetic matching approach for Hindi and Marathi,” *Comput. Sci. Eng.*, vol. 1, no. 3, 2011.

[73] S. N. Kayte, D. C. N. Kayte, and D. B. Gawali, “Grapheme-to-phoneme tools for the Marathi speech synthesis,” *Sangramsing Kayte al. Int. J. Eng. Res. Appl. ISSN*, pp. 2248–9622, 2015.

[74] A. Goel, D. Bansal, and K. Jindal, “Grapheme to Phoneme Conversion for Punjabi Language,” *SGI Reflections*, p. 3.

[75] H. Sarmad, “to-sound conversion for Urdu text-to-speech system,” in *Workshop on Computational Approaches to Arabic Script*, 2004.

[76] N. R. NANDARGE and G. P. REDDY, MALLAMMA V. SUMAN GOUDA, “KANNADA PHONETIC TRANSCRIPTION: NLP,” *Int. J. Adv. Electron. Comput. Sci.*, vol. 4, no. 6, pp. 51–53, 2017.

[77] A. G. Ramakrishnan and M. Laxmi Narayana, “Grapheme to phoneme conversion for Tamil speech synthesis,” in *Proc.*

Workshop in Image and Signal Processing (WISP-2007), IIT Guwahati, 2007, pp. 96–99.

[78] N. Udhyakumar, C. S. Kumar, R. Srinivasan, and R. Swaminathan, "Decision tree learning for automatic grapheme-to-phoneme conversion for Tamil," in *9th Conference Speech and Computer*, 2004.

[79] S. S. Nair, C. R. Rechitha, and C. S. Kumar, "Rule-based grapheme to phoneme converter for Malayalam," *Int. J. Comput. Linguist. Nat. Lang. Process.*, vol. 2, no. 7, pp. 417–420, 2013.

[80] C. Kurian and K. Balakrishnan, "Automated Transcription System for Malayalam Language," *Int. J. Comput. Appl.*, vol. 19, no. 5, pp. 5–10, 2011.

[81] E. Bozkurt, C. E. Erdem, E. Erzin, T. Erdem, and M. Ozkan, "Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation," in *2007 3DTV Conference*, 2007, pp. 1–4.

[82] J. Lander, "Read my lips: Facial animation techniques," *Game Dev. Mag. C. Media Gr.*, pp. 17–21, 1999.

[84] A. A. Montgomery and P. L. Jackson, "Physical characteristics of the lips underlying vowel lipreading performance," *J. Acoust. Soc. Am.*, vol. 73, no. 6, pp. 2134–2144, 1983.

[85] C. Neti *et al.*, "Audio-visual speech recognition," in *Final workshop*, 2000, pp. 40–41.

[86] C. A. Binnie, P. L. Jackson, and A. A. Montgomery, "Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation," *J. Speech Hear. Disord.*, vol. 41, no. 4, pp. 530–539, 1976.

[87] C. G. Fisher, "Confusions among visually perceived consonants," *J. Speech Hear. Res.*, vol. 11, no. 4, pp. 796–804, 1968.

[88] H. L. Bear, R. W. Harvey, and Y. Lan, "Finding phonemes: improving machine lip-reading," *arXiv Prepr. arXiv1710.01142*,

2017.

[89] S. Taylor, B.-J. Theobald, and I. Matthews, “A mouth full of words: Visually consistent acoustic redubbing,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4904–4908.

[90] J. Jeffers and M. Barley, *Speechreading (lipreading)*. Charles C. Thomas Publisher, 1980.

[91] E. Setyati, S. Sumpeno, M. H. Purnomo, K. Mikami, M. Kakimoto, and K. Kondo, “Phoneme-Viseme Mapping for Indonesian Language Based on Blend Shape Animation.,” *IAENG Int. J. Comput. Sci.*, vol. 42, no. 3, 2015.

[92] W. Mattheyses, L. Latacz, and W. Verhelst, “Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis,” *Speech Commun.*, vol. 55, no. 7–8, pp. 857–876, 2013.

[93] T. Seko, N. Ukai, S. Tamura, and S. Hayamizu, “Improvement of lipreading performance using discriminative feature and speaker adaptation,” in *Auditory-Visual Speech Processing (AVSP) 2013*, 2013.

[94] D. Yu, O. Ghita, A. Sutherland, and P. F. Whelan, “A novel visual speech representation and HMM classification for visual speech recognition,” in *Pacific-Rim Symposium on Image and Video Technology*, 2009, pp. 398–409.

[95] A. G. Chitu and L. J. M. Rothkrantz, “Visual speech recognition-automatic system for lip reading of dutch,” *J. Inf. Technol. Control. vol. year vii*, vol. 3, pp. 2–9, 2009.

[96] P. Damien, N. Wakim, and M. Egéa, “Phoneme-viseme mapping for Modern, Classical Arabic language,” in *2009 International Conference on Advances in Computational Tools for Engineering Applications*, 2009, pp. 547–552.

[97] J. Melenchón, J. Simó, G. Cobo, and E. Martínez, “Objective viseme extraction and audiovisual uncertainty: estimation limits between auditory and visual modes.,” in *AVSP*, 2007, p. 13.

[98] B. Aschenberner and C. Weiss, "Phoneme-viseme mapping for german video-realistic audio-visual-speech-synthesis," 2005.

[99] N. Akhter, "A viseme recognition system using lip curvature and neural networks to detect Bangla vowels." BRAC Univeristy, 2016.

[100] P. Varshney, O. Farooq, and P. Upadhyaya, "Hindi viseme recognition using subspace DCT features," *Int. J. Appl. Pattern Recognit.*, vol. 1, no. 3, pp. 257–272, 2014.

[101] S. S. Kumar, "Encoding Malayalam visemes for facial image synthesis," in *2008 IET International Conference on Wireless, Mobile and Multimedia Networks*, 2008, pp. 271–274.

[102] Y. Dai and Y. Nakano, "Face-texture model based on SGLD and its application in face detection in a color scene," *Pattern Recognit.*, vol. 29, no. 6, pp. 1007–1017, 1996.

[103] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 696–706, 2002.

[104] L. M. Bergasa, M. Mazo, A. Gardel, M. A. Sotelo, and L. Boquete, "Unsupervised and adaptive Gaussian skin-color model," *Image Vis. Comput.*, vol. 18, no. 12, pp. 987–1003, 2000.

[105] D. A. Brown, I. Craw, and J. Lewthwaite, "A SOM Based Approach to Skin Detection with Application in Real Time Systems.," in *BMVC*, 2001, vol. 1, no. 2001, pp. 491–500.

[106] T. S. Caetano, S. D. Olabarriaga, and D. A. C. Barone, "Performance evaluation of single and multiple-gaussian models for skin color modeling," in *Proceedings. XV Brazilian Symposium on Computer Graphics and Image Processing*, 2002, pp. 275–282.

[107] N. Oliver, A. P. Pentland, and F. Berard, "Lafter: Lips and face real time tracker," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997, pp. 123–129.

[108] K. Schwerdt and J. L. Crowley, "Robust face tracking using color," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000, pp. 90–95.

[109] N. Sebe, I. Cohen, T. S. Huang, and T. Gevers, "Skin detection: A bayesian network approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, 2004, vol. 2, pp. 903–906.

[110] M. Störring, T. Kočka, H. J. Andersen, and E. Granum, "Tracking regions of human skin through illumination changes," *Pattern Recognit. Lett.*, vol. 24, no. 11, pp. 1715–1723, 2003.

[111] M.-H. Yang and N. Ahuja, "Gaussian mixture model for human skin color and its applications in image and video databases," in *Storage and retrieval for image and video databases VII*, 1998, vol. 3656, pp. 458–467.

[112] C. Chen and S.-P. Chiang, "Detection of human faces in colour images," *IEE Proceedings-Vision, Image Signal Process.*, vol. 144, no. 6, pp. 384–388, 1997.

[113] S. J. McKenna, S. Gong, and Y. Raja, "Modelling facial colour and identity with gaussian mixtures," *Pattern Recognit.*, vol. 31, no. 12, pp. 1883–1892, 1998.

[114] Y. Wang and B. Yuan, "A novel approach for human face detection from color images under complex background," *Pattern Recognit.*, vol. 34, no. 10, pp. 1983–1992, 2001.

[115] D. Chai and A. Bouzerdoum, "A Bayesian approach to skin color classification in YCbCr color space," in *2000 TENCON Proceedings. Intelligent Systems and Technologies for the New Millennium (Cat. No. 00CH37119)*, 2000, vol. 2, pp. 421–424.

[116] P. Kakumanu, S. Makrogiannis, R. Bryll, S. Panchanathan, and N. Bourbakis, "Image chromatic adaptation using ANNs for skin color adaptation," in *16th IEEE International Conference on Tools with Artificial Intelligence*, 2004, pp. 478–485.

[117] M. J. Jones and J. M. Rehg, "Statistical color models with

application to skin detection," *Int. J. Comput. Vis.*, vol. 46, no. 1, pp. 81–96, 2002.

[118] B. Jedynak, H. Zheng, and M. Daoudi, "Statistical models for skin detection," in *2003 Conference on Computer Vision and Pattern Recognition Workshop*, 2003, vol. 8, p. 92.

[119] J. Y. Lee and S. I. Yoo, "An elliptical boundary model for skin color detection," in *Proc. of the 2002 International Conference on Imaging Science, Systems, and Technology*, 2002.

[120] Z. Fu, J. Yang, W. Hu, and T. Tan, "Mixture clustering using multidimensional histograms for skin detection," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, 2004, vol. 4, pp. 549–552.

[121] I. Anagnostopoulos, C. Anagnostopoulos, V. Loumos, and E. Kayafas, "A probabilistic neural network for human face identification based on fuzzy logic chromatic rules," *IEEE MED03*, 2003.

[122] M.-J. Seow, D. Valaparla, and V. K. Asari, "Neural network based skin color model for face detection," in *32nd Applied Imagery Pattern Recognition Workshop, 2003. Proceedings.*, 2003, pp. 141–145.

[123] H. Greenspan, J. Goldberger, and I. Eshet, "Mixture model for face-color modeling and segmentation," *Pattern Recognit. Lett.*, vol. 22, no. 14, pp. 1525–1536, 2001.

[124] Q. Huynh-Thu, M. Meguro, and M. Kaneko, "Skin-color extraction in images with complex background and varying illumination," in *Sixth IEEE Workshop on Applications of Computer Vision, 2002.(WACV 2002). Proceedings.*, 2002, pp. 280–285.

[125] S. Marcel and S. Bengio, "Improving face verification using skin color information," in *Object recognition supported by user interaction for service robots*, 2002, vol. 2, pp. 378–381.

[126] K.-M. Cho, J.-H. Jang, and K.-S. Hong, "Adaptive skin-color filter," *Pattern Recognit.*, vol. 34, no. 5, pp. 1067–1073, 2001.

[127] M. Soriano, B. Martinkauppi, S. Huovinen, and M. Laaksonen, "Adaptive skin color modeling using the skin locus for selecting training pixels," *Pattern Recognit.*, vol. 36, no. 3, pp. 681–690, 2003.

[128] L. Sigal, S. Sclaroff, and V. Athitsos, "Skin color-based video segmentation under time-varying illumination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 7, pp. 862–877, 2004.

[129] X. Zhai, H. Lu, and L. Zhang, "Application of image segmentation technique in tongue diagnosis," in *2009 International Forum on Information Technology and Applications*, 2009, vol. 2, pp. 768–771.

[130] M.-J. Shi, G.-Z. Li, F.-F. Li, and C. Xu, "Computerized tongue image segmentation via the double geo-vector flow," *Chin. Med.*, vol. 9, no. 1, p. 7, 2014.

[131] K. Wu and D. Zhang, "Robust tongue segmentation by fusing region-based and edge-based approaches," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 8027–8038, 2015.

[132] Z. Cui, W. Zuo, H. Zhang, and D. Zhang, "Automated tongue segmentation based on 2D Gabor filters and fast marching," in *International Conference on Intelligent Science and Big Data Engineering*, 2013, pp. 328–335.

[133] J. Ning, D. Zhang, C. Wu, and F. Yue, "Automatic tongue image segmentation based on gradient vector flow and region merging," *Neural Comput. Appl.*, vol. 21, no. 8, pp. 1819–1826, 2012.

[134] E. Skodras and N. Fakotakis, "An unconstrained method for lip detection in color images," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 1013–1016.

[135] J. Du, Y. Lu, M. Zhu, K. Zhang, and C. Ding, "A novel algorithm of color tongue image segmentation based on HSI," in *2008 International Conference on BioMedical Engineering and Informatics*, 2008, vol. 1, pp. 733–737.

[136] N. Eveno, A. Caplier, and P.-Y. Coulon, “New color transformation for lips segmentation,” in *2001 IEEE Fourth Workshop on Multimedia Signal Processing (Cat. No. 01TH8564)*, 2001, pp. 3–8.

[137] A. Panning, R. Niese, A. Al-Hamadi, and B. Michaelis, “A new adaptive approach for histogram based mouth segmentation,” *Proc. World Acad. Sci. Eng. Technol.*, vol. 56, pp. 779–784, 2009.

[138] A. D. Gritzman, V. Aharonson, D. M. Rubin, and A. Pantanowitz, “Automatic computation of histogram threshold for lip segmentation using feedback of shape information,” *Signal, Image Video Process.*, vol. 10, no. 5, pp. 869–876, 2016.

[139] R. Stiefelhagen, U. Meier, and J. Yang, “Real-time lip-tracking for lipreading,” in *Fifth European Conference on Speech Communication and Technology*, 1997.

[140] R. Stiefelhagen, J. Yang, and A. Waibel, “A model-based gaze tracking system,” *Int. J. Artif. Intell. Tools*, vol. 6, no. 02, pp. 193–209, 1997.

[141] S.-H. Leung, S.-L. Wang, and W.-H. Lau, “Lip image segmentation using fuzzy clustering incorporating an elliptic shape function,” *IEEE Trans. image Process.*, vol. 13, no. 1, pp. 51–62, 2004.

[142] S. Lucey, S. Sridharan, and W. Chandran, “Chromatic lip tracking using a connectivity based fuzzy thresholding technique,” in *ISSPA’99. Proceedings of the Fifth International Symposium on Signal Processing and its Applications (IEEE Cat. No. 99EX359)*, 1999, vol. 2, pp. 669–672.

[143] A.-C. Liew, S. H. Leung, and W. H. Lau, “Segmentation of color lip images by spatial fuzzy clustering,” *IEEE Trans. Fuzzy Syst.*, vol. 11, no. 4, pp. 542–549, 2003.

[144] A. D. Gritzman, D. M. Rubin, and A. Pantanowitz, “Comparison of colour transforms used in lip segmentation algorithms,” *Signal, Image Video Process.*, vol. 9, no. 4, pp. 947–957, 2015.

[145] M. Li and Y. Cheung, “Automatic segmentation of color lip images based on morphological filter,” in *International Conference on Artificial Neural Networks*, 2010, pp. 384–387.

[146] M. Li, “Automatic Segmentation of Lip Images Based on Markov Random Field.”

[147] V. Vezhnevets, “Face and facial feature tracking for natural human-computer interface,” in *International Conference on Computer Graphics between Europe and Asia (GraphiCon-2002)*, 2002, pp. 86–90.

[148] M. U. R. Sánchez, J. Matas, and J. Kittler, “Statistical chromaticity-based lip tracking with B-splines,” in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997, vol. 4, pp. 2973–2976.

[149] Y.-P. Guan, “Automatic extraction of lip based on wavelet edge detection,” in *2006 Eighth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, 2006, pp. 125–132.

[150] N. Eveno, A. Caplier, and P.-Y. Coulon, “Accurate and quasi-automatic lip tracking,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 706–715, 2004.

[151] X. Liu and Y. Cheung, “A robust lip tracking algorithm using localized color active contours and deformable models,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 1197–1200.

[152] G. I. Prajapati and N. M. Patel, “DToLIP: Detection and tracking of lip contours from human facial images using Snake’s method,” in *2011 International Conference on Image Information Processing*, 2011, pp. 1–6.

[153] S. S. Morade and B. S. Patnaik, “Automatic lip tracking and extraction of lip geometric features for lip reading,” *Int. J. Mach. Learn. Comput.*, vol. 3, no. 2, p. 168, 2013.

[154] S. G. Yuanyao Lu, “A Key Point Extraction Method of Lip Contours Based on Jumping Snake,” in *6th International*

Conference on Manufacturing Science and Engineering, 2015.

[155] P. Kuo, P. Hillman, and J. Hannah, “Improved lip fitting and tracking for model-based multimedia and coding,” 2005.

[156] H. Seyedarabi, W. Lee, and A. Aghagolzadeh, “Automatic lip tracking and action units classification using two-step active contours and probabilistic neural networks,” in *2006 Canadian Conference on Electrical and Computer Engineering*, 2006, pp. 2021–2024.

[157] P. Dalka, P. Bratoszewski, and A. Czyzewski, “Visual lip contour detection for the purpose of speech recognition,” in *2014 International Conference on Signals and Electronic Systems (ICSES)*, 2014, pp. 1–4.

[158] J. Luettin, N. A. Thacker, and S. W. Beet, “Visual speech recognition using active shape models and hidden Markov models,” in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 1996, vol. 2, pp. 817–820.

[159] J. Luettin, N. A. Thacker, and S. W. Beet, “Active shape models for visual speech feature extraction,” in *Speechreading by humans and machines*, Springer, 1996, pp. 383–390.

[160] T. A. Faruquie, A. Majumdar, N. Rajput, and L. V. Subramaniam, “Large vocabulary audio-visual speech recognition using active shape models,” in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, 2000, vol. 3, pp. 106–109.

[161] S. Lucey, S. Sridharan, and V. Chandran, “Initialised eigenlip estimator for fast lip tracking using linear regression,” in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, 2000, vol. 3, pp. 178–181.

[162] L. Xie, X.-L. Cai, Z.-H. Fu, R.-C. Zhao, and D.-M. Jiang, “A robust hierarchical lip tracking approach for lipreading and audio visual speech recognition,” in *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 04EX826)*, 2004, vol. 6, pp. 3620–3624.

[163] K. L. Sum, W. H. Lau, S. H. Leung, A. W.-C. Liew, and K. W. Tse, “A new optimization procedure for extracting the point-based lip contour using active shape model,” in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, 2001, vol. 3, pp. 1485–1488.

[164] J. Luettin, N. A. Thacker, and S. W. Beet, “Statistical lip modelling for visual speech recognition,” in *1996 8th European Signal Processing Conference (EUSIPCO 1996)*, 1996, pp. 1–4.

[165] Q. D. Nguyen and M. Milgram, “Multi features models for robust lip tracking,” in *2008 10th International Conference on Control, Automation, Robotics and Vision*, 2008, pp. 1333–1337.

[166] B. Beaumesnil and F. Luthon, “Real time tracking for 3D realistic lip animation,” in *18th International Conference on Pattern Recognition (ICPR’06)*, 2006, vol. 1, pp. 219–222.

[167] F. Luthon and B. Beaumesnil, “Real-time liptracking for synthetic face animation with feedback loop,” in *International Conference on Computer Vision Theory and Applications (VISAPP’06)*, 2006, vol. 2, pp. 402–407.

[168] S.-L. Wang, W.-H. Lau, A. W.-C. Liew, and S.-H. Leung, “Robust lip region segmentation for lip images with complex background,” *Pattern Recognit.*, vol. 40, no. 12, pp. 3481–3491, 2007.

[169] E. Cosatto, G. Potamianos, and H. P. Graf, “Audio-visual unit selection for the synthesis of photo-realistic talking-heads,” in *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, 2000, vol. 2, pp. 619–622.

[170] G. Englebienne, T. Cootes, and M. Rattray, “A probabilistic model for generating realistic lip movements from speech,” in *Advances in neural information processing systems*, 2008, pp. 401–408.

[171] O. Govokhina, G. Bailly, G. Breton, and P. Bagshaw, “A new trainable trajectory formation system for facial animation,” in *Workshop on Experimental Linguistics*, 2006, pp. 25–32.

[172] M. S. Gordon and M. Hibberts, “Audiovisual speech from emotionally expressive and lateralized faces,” *Q. J. Exp. Psychol.*, vol. 64, no. 4, pp. 730–750, 2011.

[173] J. Yang, J. Xiao, and M. Ritter, “Automatic selection of visemes for image-based visual speech synthesis,” in *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, 2000, vol. 2, pp. 1081–1084.

[174] D. Cosker, S. Paddock, D. Marshall, P. L. Rosin, and S. Rushton, “Toward perceptually realistic talking heads: Models, methods, and mcgurk,” *ACM Trans. Appl. Percept.*, vol. 2, no. 3, pp. 270–285, 2005.

[175] F. Elisei, M. Odisio, G. Bailly, and P. Badin, “Creating and controlling video-realistic talking heads.,” in *AVSP*, 2001, pp. 90–97.

[176] J. Beskow and M. Nordenberg, “Data-driven synthesis of expressive visual speech using an MPEG-4 talking head,” in *Ninth European Conference on Speech Communication and Technology*, 2005.

[177] O. Engwall, “Evaluation of a system for concatenative articulatory visual speech synthesis,” in *Seventh International Conference on Spoken Language Processing*, 2002.

[178] P. Dey, S. C. Maddock, and R. Nicolson, “Evaluation of A Viseme-Driven Talking Head.,” *TPCG*, vol. 10, pp. 139–142, 2010.

[179] Z. Deng and U. Neumann, “eFASE: expressive facial animation synthesis and editing with phoneme-isomap controls,” in *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, 2006, pp. 251–260.

[180] J. M. De Martino, L. P. Magalhães, and F. Violaro, “Facial

animation based on context-dependent visemes," *Comput. Graph.*, vol. 30, no. 6, pp. 971–980, 2006.

[181] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: a back-projected human-like robot head for multiparty human-machine interaction," in *Cognitive behavioural systems*, Springer, 2012, pp. 114–130.

[182] E. Cosatto, J. Ostermann, H. P. Graf, and J. Schroeter, "Lifelike talking faces for interactive services," *Proc. IEEE*, vol. 91, no. 9, pp. 1406–1429, 2003.

[183] I. Steiner and S. Ouni, "Progress in animation of an EMA-controlled tongue model for acoustic-visual speech synthesis," *arXiv Prepr. arXiv1201.4080*, 2012.

[184] I. Albrecht, J. Haber, K. Kahler, M. Schroder, and H.-P. Seidel, "' May I talk to you?:-)'-facial animation from text," in *10th Pacific Conference on Computer Graphics and Applications, 2002. Proceedings.*, 2002, pp. 77–86.

[185] E. Cosatto and H. P. Graf, "Photo-realistic talking-heads from image samples," *IEEE Trans. Multimed.*, vol. 2, no. 3, pp. 152–163, 2000.

[186] M. Aharon and R. Kimmel, "Representation analysis and synthesis of lip images using dimensionality reduction," *Int. J. Comput. Vis.*, vol. 67, no. 3, pp. 297–312, 2006.

[187] R. Gutierrez-Osuna *et al.*, "Speech-driven facial animation with realistic dynamics," *IEEE Trans. Multimed.*, vol. 7, no. 1, pp. 33–42, 2005.

[188] T. Ezzat, G. Geiger, and T. Poggio, *Trainable videorealistic speech animation*, vol. 21, no. 3. ACM, 2002.

[189] F. J. Huang, E. Cosatto, and H. P. Graf, "Triphone based unit selection for concatenative visual speech synthesis," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 2, p. II-2037.

[190] E. Agelfors, J. Beskow, I. Karlsson, J. Kewley, G. Salvi, and N. Thomas, "User evaluation of the SYNFACE talking head

telephone," in *International Conference on Computers for Handicapped Persons*, 2006, pp. 579–586.

[191] A. Verma, N. Rajput, and L. V. Subramaniam, "Using viseme based acoustic models for speech driven lip synthesis," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, 2003, vol. 5, p. V-720.

[192] S. Deena, S. Hou, and A. Galata, "Visual speech synthesis by modelling coarticulation dynamics using a non-parametric switching state-space model," in *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, 2010, p. 29.

[193] A. V. Tanveer A. Faruuie, Chalapathy Neti, Nitendra Rajput L., Venkata Subraniam, "Animating expressive faces to speak in Indian languages."

[194] A. R. A. and N. Ahmad, "Urdu Viseme Identification," pp. 68–71.

[195] A. N. Jothilakshmi, "Modeling of 3D human face for Tamil visual speech synthesis," *IJSSST*, pp. 68–71.

[196] C. for L. Minorities and G. of I. Ministry of Minority Affairs, "Report of the Commissioner for linguistic minorities: 50th report (July 2012 to June 2013)."

[197] N. M. Brooke and S. D. Scott, "Two-and three-dimensional audio-visual speech synthesis," in *AVSP'98 International Conference on Auditory-Visual Speech Processing*, 1998.

[198] B.-J. Theobald, J. A. Bangham, I. A. Matthews, J. R. W. Glauert, and G. C. Cawley, "2.5 D Visual Speech Synthesis Using Appearance Models.," in *BMVC*, 2003, pp. 1–10.

[199] A. Turkmani, "Visual analysis of viseme dynamics." University of Surrey, 2008.

[200] M. Chu and Y. Feng, "Study on factors influencing durations of syllables in Mandarin," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[201] V. R. Prabodhachandran Nair, *Svanavijnjaanam (Phonetics)*. SIL, 1980.

[202] D. H. Klatt, “Synthesis by rule of segmental durations in English sentences,” *Front. Speech Comm. Res.*, pp. 287–299, 1979.

[203] K. Bartkova and C. Sorin, “A model of segmental duration for speech synthesis in French,” *Speech Commun.*, vol. 6, no. 3, pp. 245–260, 1987.

[204] A. R. M. Simões, “Predicting sound segment duration in connected speech: An acoustical study of Brazilian Portuguese,” in *The ESCA Workshop on Speech Synthesis*, 1991.

[205] A. Bellur, K. B. Narayan, K. R. Krishnan, and H. A. Murthy, “Prosody modeling for syllable-based concatenative speech synthesis of Hindi and Tamil,” in *2011 National Conference on Communications (NCC)*, 2011, pp. 1–5.

[206] “Malayalam Phonetic Archive,” 2016. [Online]. Available: http://www.cmltemu.in.

[207] T. Balasubramanian, *A textbook of English phonetics for Indian students*. Macmillan, 1981.

[208] R. Lawrence, *Fundamentals of speech recognition*. Pearson Education India, 2008.

[209] R. E. Asher, *Malayalam*. Routledge, 2013.

[210] K. Sekiyama, “Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility,” *J. Acoust. Soc. Japan*, vol. 15, no. 3, pp. 143–158, 1994.

[211] O. Govokhina, G. Bailly, and G. Breton, “Learning optimal audiovisual phasing for a HMM-based control model for facial animation,” in *6th ISCA Workshop on Speech Synthesis (SSW6)*, 2007, pp. 1–4.

[212] R. De Mori, *Computer models of speech using fuzzy algorithms*. Springer Science & Business Media, 2013.

[213] B. H. Repp and H. Lin, “Acoustic properties and perception of stop consonant release transients,” *J. Acoust. Soc. Am.*, vol. 85, no. 1, pp. 379–396, 1989.

[214] M. M. Cohen and D. W. Massaro, “Modeling coarticulation in synthetic visual speech,” in *Models and techniques in computer animation*, Springer, 1993, pp. 139–156.

[215] B. Kim, G. Lee, and J.-H. Lee, “Hybrid grapheme to phoneme conversion for unlimited vocabulary,” *Nat. Lang. Eng.*, 1998.

[216] M. Razavi, R. Rasipuram, and M. M.- Doss, “Acoustic data-driven grapheme-to-phoneme conversion in the probabilistic lexical modeling framework,” *Speech Commun.*, vol. 80, pp. 1–21, 2016.

[217] U. Reichel, H. R. Pfitzinger, and H.-U. Hain, “English grapheme-to-phoneme conversion and evaluation,” *Speech Lang. Technol.*, vol. 11, pp. 159–166, 2008.

[218] J. Gros, F. Mihelič, S. Dobrišek, T. Erjavec, and M. Žganec, “Rules for automatic grapheme-to-allophone transcription in slovene,” in *International Workshop on Text, Speech and Dialogue*, 2000, pp. 171–176.

[219] F. Sindran, F. Mualla, T. Haderlein, K. Daqrouq, and E. Nöth, “Rule-based standard arabic phonetization at phoneme, allophone, and syllable level,” *Int. J. Comput. Linguist.*, vol. 7, pp. 23–37, 2016.

[220] M. Wypych, E. Baranowska, and G. Demenko, “A grapheme-to-phoneme transcription algorithm based on the sampa alphabet extension for the Polish language,” *Mach. Learn.*, vol. 21, no. 22, 1973.

[221] P. Bonaventura, F. Giuliani, J. M. Garrido, and I. Ortin, “Grapheme-to-phoneme transcription rules for Spanish, with application to automatic speech recognition and synthesis,” in *Proceedings of the Workshop on Partially Automated Techniques for Transcribing Naturally Occurring Continuous Speech*, 1998, pp. 33–39.

[222] V. P., "HIDDEN MARKOV MODEL BASED KEYWORD SPOTTING FOR MALAYALAM SPEECH ANALYTICS," UNIVERSITY OF CALICUT, 2017.

[223] E. Agirre, O. Ansa, E. Hovy, and D. Martinez, "Enriching WordNet concepts with topic signatures," *arXiv Prepr. cs/0109031*, 2001.

[224] J. R. Bellegarda, "Large vocabulary speech recognition with multispan statistical language models," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 1, pp. 76–84, 2000.

[225] P. B. Denes, "Statistics of spoken English," *J. Acoust. Soc. Am.*, vol. 34, no. 12, pp. 1978–1979, 1962.

[226] E. J. Yannakoudakis and P. J. Hutton, "An assessment of n-phoneme statistics in phoneme guessing algorithms which aim to incorporate phonotactic constraints," *Speech Commun.*, vol. 11, no. 6, pp. 581–602, 1992.

[227] C. Basztura, "Rozmawiac z komputerem (Eng. To speak with computers)," *Wrocław: Format*, 1992.

[228] M. Larson, "Sub-word-based language models for speech recognition: implications for spoken document retrieval," *Whorkshop Lang. Model. Inf. Retr.*, 2001.

[229] B. Ziółko, J. Gałka, S. Manandhar, R. C. Wilson, and M. Ziółko, "The use of statistics of Polish phonemes in speech recognition."

[230] M. P. S. Ms. Sunder, "RESEARCH ON PHONEME SEQUENCES FOR LANGUAGE IDENTIFICATION AND CONCURRENT VOICE TRANSMISSION," *IJCSMC*, vol. 3, no. 6, pp. 116–120, 2014.

[231] P. Dalsgaard, O. Andersen, H. Hesselager, and B. Petek, "Language-identification using language-dependent phonemes and language-independent speech units," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, 1996, vol. 3, pp. 1808–1811.

[232] P. Matějka, I. Szöke, P. Schwarz, and J. Černocký, "Automatic

language identification using phoneme and automatically derived unit strings," in *International Conference on Text, Speech and Dialogue*, 2004, pp. 147–153.

[233] S. Verberne, "Context-sensitive spell checking based on word trigram probabilities," *Unpubl. master's thesis, Univ. Nijmegen*, 2002.

[234] B. Bansal, M. Choudhury, P. R. Ray, S. Sarkar, and A. Basu, "Isolated-word Error Correction for Partially Phonemic Languages using Phonetic Cues," in *International Conference on Knowledge based Computer Systems (KBCS 2004)*, 2004, pp. 509–519.

[235] K. Tamaoka and S. Makioka, "Frequency of occurrence for units of phonemes, morae, and syllables appearing in a lexical corpus of a Japanese newspaper," *Behav. Res. Methods, Instruments, Comput.*, vol. 36, no. 3, pp. 531–547, 2004.

[236] N. Smirnova and P. Chistikov, "Statistics of Russian monophones and diphones," *Proc. Specom-2011. Kazan, Russ.*, pp. 218–223, 2011.

[237] I. Esquerra, A. Febrer, and C. Nadeu, "Frequency analysis of phonetic units for concatenative synthesis in Catalan," in *Fifth International Conference on Spoken Language Processing*, 1998.

[238] P. Kłosowski, "Statistical analysis of orthographic and phonemic language corpus for word-based and phoneme-based Polish language modelling," *EURASIP J. Audio, Speech, Music Process.*, vol. 2017, no. 1, p. 5, 2017.

[239] N. Sreedevi and M. Irfana, "Frequency of occurrence of phonemes in Calicut and Eranakulam dialects of Malayalam.," *J. All India Inst. Speech Hear.*, vol. 32, 2013.

[240] D. Poeppel and P. J. Monahan, "Speech perception: Cognitive foundations and cortical implementation," *Curr. Dir. Psychol. Sci.*, vol. 17, no. 2, pp. 80–85, 2008.

[241] M. F. Woodward and C. G. Barber, "Phoneme perception in

lipreading," *J. Speech Hear. Res.*, vol. 3, no. 3, pp. 212–222, 1960.

[242] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, p. 746, 1976.

[243] E. Owens and B. Blazek, "Visemes observed by hearing-impaired and normal-hearing adult viewers," *J. Speech, Lang. Hear. Res.*, vol. 28, no. 3, pp. 381–393, 1985.

[244] E. T. Auer Jr and L. E. Bernstein, "Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness," *J. Acoust. Soc. Am.*, vol. 102, no. 6, pp. 3704–3710, 1997.

[245] E. Cosatto and H. P. Graf, "Sample-based synthesis of photo-realistic talking heads," in *Proceedings Computer Animation'98 (Cat. No. 98EX169)*, 1998, pp. 103–110.

[246] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces.," in *Siggraph*, 1999, vol. 99, no. 1999, pp. 187–194.

[247] M. Dimitrijevic, S. Ilic, and P. Fua, "Accurate face models from uncalibrated and ill-lit video sequences," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 2004, vol. 2, pp. II–II.

[248] S. L. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews, "Dynamic units of visual speech," in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2012, pp. 275–284.

[249] I. S. Pandzic and R. Forchheimer, *MPEG-4 facial animation: the standard, implementation and applications*. John Wiley & Sons, 2003.

[250] W. Mattheyses, L. Latacz, and W. Verhelst, "Automatic viseme clustering for audiovisual speech synthesis," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[251] L. Cappelletta and N. Harte, "Phoneme-to-viseme mapping for

visual speech recognition.," in *ICPRAM (2)*, 2012, pp. 322–329.

[252] P. Lucey, T. Martin, and S. Sridharan, "Confusability of phonemes grouped according to their viseme classes in noisy environments," in *Australian International Conference on Speech Science & Tech*, 2004, pp. 265–270.

[253] B. B. Owens E, "Visemes observed by hearing-impaired and normal-hearing adult viewers.," *J Speech Hear Res.*, vol. 28(3), 1985.

[254] H. L. Bear, R. W. Harvey, B.-J. Theobald, and Y. Lan, "Which phoneme-to-viseme maps best improve visual-only computer lip-reading?," in *International Symposium on Visual Computing*, 2014, pp. 230–239.

[255] N. Ahmad, S. Datta, D. Mulvaney, and O. Farooq, "A comparison of visual features for audiovisual automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 123, no. 5, p. 3939, 2008.

[256] L. Cappelletta and N. Harte, "Viseme definitions comparison for visual-only speech recognition," in *2011 19th European Signal Processing Conference*, 2011, pp. 2109–2113.

[257] B. Lidestam and J. Beskow, "Visual phonemic ambiguity and speechreading," *J. Speech, Lang. Hear. Res.*, 2006.

[258] N. Alothmany, R. Boston, C. Li, S. Shaiman, and J. Durrant, "Classification of visemes using visual cues," in *Proceedings ELMAR-2010*, 2010, pp. 345–349.

[259] M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, "DCT-based video features for audio-visual speech recognition," in *Seventh International Conference on Spoken Language Processing*, 2002.

[260] M. J. Zaki, W. Meira Jr, and W. Meira, *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[261] "Hierarchical Clustering." [Online]. Available: https://in.mathworks.com/help/stats/hierarchical-clustering.html.

[262] V. A. Kozhevnikov, “Rech: Artikulatsiya i Vospriatatie (Moscow-Leningrad),” *Trans. Speech Articul. Perception. Washington, DC Jt. Publ. Res. Serv.*, vol. 30, p. 543, 1965.

[263] F. Bell-Berti and K. S. Harris, “Anticipatory coarticulation: Some implications from a study of lip rounding,” *J. Acoust. Soc. Am.*, vol. 65, no. 5, pp. 1268–1270, 1979.

[264] A.-P. Benguerel and M. K. Pichora-Fuller, “Coarticulation effects in lipreading,” *J. Speech, Lang. Hear. Res.*, vol. 25, no. 4, pp. 600–607, 1982.

[265] A. P. Breen, E. Bowers, and W. Welsh, “An investigation into the generation of mouth shapes for a talking head,” in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP’96*, 1996, vol. 4, pp. 2159–2162.

[266] S. Hilder, B.-J. Theobald, and R. Harvey, “In pursuit of visemes,” in *Auditory-Visual Speech Processing 2010*, 2010.

[267] A. Weissenfeld, K. Liu, and J. Ostermann, “Video-realistic image-based eye animation via statistically driven state machines,” *Vis. Comput.*, vol. 26, no. 9, pp. 1201–1216, 2010.

[268] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, “Rigid head motion in expressive speech animation: Analysis and synthesis,” *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 15, no. 3, pp. 1075–1086, 2007.

[269] H. Seyedarabi, A. Aghagolzadeh, and S. Khanmohammadi, “Facial expressions animation and lip tracking using facial characteristic points and deformable model,” *Int. J. Inf. Technol.*, vol. 1, no. 4, pp. 165–168, 2004.

[270] H. Li, J. Yu, Y. Ye, and C. Bregler, “Realtime facial animation with on-the-fly correctives.,” *ACM Trans. Graph.*, vol. 32, no. 4, pp. 41–42, 2013.

[271] S. Bouaziz, Y. Wang, and M. Pauly, “Online modeling for realtime facial animation,” *ACM Trans. Graph.*, vol. 32, no. 4, p. 40, 2013.

[272] B. Beaumesnil, F. Luthon, and M. Chaumont, “Liptracking and MPEG4 animation with feedback control,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006, vol. 2, pp. II–II.

[273] Z. Wu, P. S. Aleksic, and A. K. Katsaggelos, “Lip tracking for MPEG-4 facial animation,” in *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, 2002, pp. 293–298.

[274] C. Benoit, T. Guiard-Marigny, B. Le Goff, and A. Adjoudani, “Which components of the face do humans and machines best speechread?,” in *Speechreading by humans and machines*, Springer, 1996, pp. 315–328.

[275] D. Burnham, R. Campbell, G. Away, and B. J. Dodd, *Hearing eye II: the psychology of speechreading and auditory-visual speech*. Routledge, 2013.

[276] W. H. Sumby and I. Pollack, “Visual contribution to speech intelligibility in noise,” *J. Acoust. Soc. Am.*, vol. 26, no. 2, pp. 212–215, 1954.

[277] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, “Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins Summer 2000 Workshop,” in *2001 IEEE Fourth Workshop on Multimedia Signal Processing (Cat. No. 01TH8564)*, 2001, pp. 619–624.

[278] C. LI, Y. Bo, and C. LI, “Deep Learning Based Visual Tracking: A Review,” *DEStech Trans. Comput. Sci. Eng.*, no. smce, 2017.

[279] S. D. Cotton and E. Claridge, “Do all human skin colours lie on a defined surface within LMS space?,” *Sch. Comput. Sci. Res. REPORTS-UNIVERSITY BIRMINGHAM CSR*, 1996.

[280] M. Sadeghi, J. Kittler, and K. Messer, “Modelling and segmentation of lip area in face images,” *IEE Proceedings-Vision, Image Signal Process.*, vol. 149, no. 3, pp. 179–184, 2002.

[281] S. B. Haralur, A. M. Dibas, N. A. Almelhi, and D. A. Al-

Qahtani, “The tooth and skin colour interrelationship across the different ethnic groups,” *Int. J. Dent.*, vol. 2014, 2014.

[282] R. Hassanpour, A. Shahbahrami, and S. Wong, “Adaptive Gaussian mixture model for skin color segmentation,” *World Acad. Sci. Eng. Technol.*, vol. 41, pp. 1–6, 2008.

[283] V. A. Oliveira and A. Conci, “Skin Detection using HSV color space,” in *H. Pedrini, & J. Marques de Carvalho, Workshops of Sibgrapi*, 2009, pp. 1–2.

[284] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, “A survey of skin-color modeling and detection methods,” *Pattern Recognit.*, vol. 40, no. 3, pp. 1106–1122, 2007.

[285] K. K. Bhoyar and O. G. Kakde, “Skin color detection model using neural networks and its performance evaluation,” in *Journal of computer science*, 2010.

[286] M. R. Tabassum *et al.*, “Comparative study of statistical skin detection algorithms for sub-continental human images,” *arXiv Prepr. arXiv1008.4206*, 2010.

[287] Q. D. Nguyen and M. Milgram, “Semi adaptive appearance models for lip tracking,” in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 2437–2440.

[288] M. Kass, A. Witkin, and D. Terzopoulos, “Snakes: Active contour models,” *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, 1988.

[289] K. S. Jang, “Lip contour extraction based on active shape model and snakes,” *Int. J. Comput. Sci. Netw. Secur.*, vol. 7, no. 10, pp. 148–153, 2007.

[290] R. Kushwahaa, N. Naina, and P. Jangraa, “Extraction of Lip Contour from Face,” 2012.

[291] M. Celenk, “A color clustering technique for image segmentation,” *Comput. Vision, Graph. image Process.*, vol. 52, no. 2, pp. 145–170, 1990.

[292] K.-S. Fu and J. K. Mui, “A survey on image segmentation,”

Pattern Recognit., vol. 13, no. 1, pp. 3–16, 1981.

[293] Y. Cheung and M. Li, "MAP-MRF based LIP segmentation without true segment number," in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 769–772.

[294] A. B. A. Hassanat and S. Jassim, "Color-based lip localization method," in *Mobile Multimedia/Image Processing, Security, and Applications 2010*, 2010, vol. 7708, p. 77080Y.

[295] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques," in *Proc. Graphicon*, 2003, vol. 3, pp. 85–92.

[296] J. Bigun, *Vision with direction*. Springer, 2006.

[297] J. Schwiegerling, "Field guide to visual and ophthalmic optics," 2004.

[298] N. A. Ibraheem, M. M. Hasan, R. Z. Khan, and P. K. Mishra, "Understanding color models: a review," *ARPN J. Sci. Technol.*, vol. 2, no. 3, pp. 265–275, 2012.

[299] J. Yang, W. Lu, and A. Waibel, "Skin-color modeling and adaptation," in *Asian Conference on Computer Vision*, 1998, pp. 687–694.

[300] B. D. Zarit, B. J. Super, and F. K. H. Quek, "Comparison of five color models in skin pixel classification," in *Proceedings International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. In Conjunction with ICCV'99 (Cat. No. PR00378)*, 1999, pp. 58–63.

[301] T.-W. Yoo and I.-S. Oh, "A fast algorithm for tracking human faces based on chromatic histograms," *Pattern Recognit. Lett.*, vol. 20, no. 10, pp. 967–978, 1999.

[302] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, vol. 3, no. 22, pp. 41–46.

[303] F. Schneiter, "Lip Contour Localization using Statistical Shape Models." Master Thesis Supervised by Gabriele Fanelli

Computer Vision Institute …, 2009.

[304] and J. A. B. Iain Matthews, Tim Cootes_, Stephen Cox, Richard Harvey, "Lip-reading using shape and scale," in *Int. Conf on Auditory-Visual Speech Processing (AVSP'98)*, 1998.

[305] I. L. Dryden, "Shape analysis," *Wiley StatsRef Stat. Ref. Online*, 2014.

[306] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 135–164, 2004.

[307] N. A. Campbell and W. R. Atchley, "The geometry of canonical variate analysis," *Syst. Biol.*, vol. 30, no. 3, pp. 268–280, 1981.

[308] M. B. Stegmann, R. Fisker, B. K. Ersbøll, H. H. Thodberg, and L. Hyldstrup, "Active appearance models: Theory and cases," in *in Proc. 9th Danish Conf. Pattern Recognition and Image Analysis*, 2000.

[309] T. Cootes, E. R. Baldock, and J. Graham, "An introduction to active shape models," *Image Process. Anal.*, pp. 223–248, 2000.

[310] S. Milborrow, "Locating facial features with active shape models." University of Cape Town, 2007.

[311] A. MacLeod and Q. Summerfield, "Quantifying the contribution of vision to speech perception in noise," *Br. J. Audiol.*, vol. 21, no. 2, pp. 131–141, 1987.

[312] D. G. Stork and M. E. Hennecke, *Speechreading by humans and machines: models, systems, and applications*, vol. 150. Springer Science & Business Media, 2013.

[313] D. Lindsay, "Talking head," *Am. Herit. Inven. Technol.*, vol. 13, pp. 56–63, 1997.

[314] F. I. Parke and K. Waters, *Computer facial animation*. AK Peters/CRC Press, 2008.

[315] W. Mattheyses and W. Verhelst, "Audiovisual speech synthesis: An overview of the state-of-the-art," *Speech Commun.*, vol. 66,

pp. 182–217, 2015.

[316] N. P. Erber and C. L. De Filippo, “Voice/mouth synthesis and tactual/visual perception of/pa, ba, ma,” *J. Acoust. Soc. Am.*, vol. 64, no. 4, pp. 1015–1019, 1978.

[317] P. Ekman and W. V Friesen, “Facial coding action system (FACS): A technique for the measurement of facial actions.” Palo Alto, CA: Consulting Psychologists Press, 1978.

[318] E. Yamamoto, S. Nakamura, and K. Shikano, “Lip movement synthesis from speech based on hidden Markov models,” *Speech Commun.*, vol. 26, no. 1–2, pp. 105–115, 1998.

[319] G. Hofer, J. Yamagishi, and H. Shimodaira, “Speech-driven lip motion generation with a trajectory HMM,” 2008.

[320] L. Wang, X. Qian, W. Han, and F. K. Soong, “Photo-real lips synthesis with trajectory-guided sample selection,” in *Seventh ISCA Workshop on Speech Synthesis*, 2010.

[321] S. Curinga, F. Lavagetto, and F. Vignoli, “Lip movements synthesis using time delay neural networks,” in *1996 8th European Signal Processing Conference (EUSIPCO 1996)*, 1996, pp. 1–4.

[322] Y. Ding and C. Pelachaud, “Lip animation synthesis: a unified framework for speaking and laughing virtual agent.,” in *AVSP*, 2015, pp. 78–83.

[323] Z. Krňoul, “Refinement of lip shape in sign speech synthesis,” 2009.

[324] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 6, pp. 681–685, 2001.

[325] M. A. Fischler and R. A. Elschlager, “The representation and matching of pictorial structures,” *IEEE Trans. Comput.*, no. 1, pp. 67–92, 1973.

[326] M. M. GE Christensen, RD Rabbitt, “Topological properties of smooth anatomic maps,” 1995.

[327] M. Turk and A. Pentland, “Eigenfaces for recognition,” *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[328] W. Mattheyses, L. Latacz, and W. Verhelst, “Active appearance models for photorealistic visual speech synthesis,” in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[329] B.-J. Theobald and I. Matthews, “Relating objective and subjective performance measures for aam-based visual speech synthesis,” *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 20, no. 8, pp. 2378–2387, 2012.

[330] J. Melenchón, E. Martínez, F. De La Torre, and J. A. Montero, “Emphatic visual speech synthesis,” *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 17, no. 3, pp. 459–468, 2009.

[331] R. Anderson, B. Stenger, V. Wan, and R. Cipolla, “Expressive visual text-to-speech using active appearance models,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3382–3389.

[332] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, 2010.

[333] B.-J. Theobald, “Visual speech synthesis using shape and appearance models.” University of East Anglia, 2003.

[334] R. Likert, “A technique for the measurement of attitudes.,” *Arch. Psychol.*, 1932.

[335] B.-J. Theobald, S. Fagel, G. Bailly, and F. Elisei, “LIPS2008: Visual speech synthesis challenge,” in *9th Annual Conference of the International Speech Communication Association (Interspeech 2008)*, 2008, pp. 2310–2313.

[336] Viola, Paul, and Michael Jones. "Robust real-time object detection." *International journal of computer vision* 4.34-47 (2001): 4.

[337] http://www.deeplearning.org.

[338] http://www.pytorch.org.

[339] Ruder, Sebastian. "An overview of gradient descent optimization algorithms." *arXiv preprint arXiv:1609.04747*(2016).

LIST OF PUBLICATIONS OF THE AUTHOR

[1] **Sandesh E. PA.**, Lajish V.L, Bibish Kumar K T, R.K.Sunil Kumar"A Comparative Study of Colour Spaces for Mouth Region Segmentation in Indian Context" *International Journal of Research in advent Technology*, Volume 6, Issue 8, August 2018.

[2] Vivek P, **Sandesh E PA**, Lajish V L "Durational Characteristics of Allophonic Variations in Malayalam Vowel Phonemes" *International Journal of Reserachin in Electronics and Computer Engineering (IJRECE)*, Vol. 6. Issue 3 July-September, 2018.

[3] **Sandesh E.PA,** Lajish V.L. "Lip Motion Synthesis for Speech Animation using Active Shape Model" *IEEE International Conference on Intelligent Computing and Control Systems (ICICCS - 2018).*

[4] Bibish Kumar K T, R K Sunil Kumar, **Sandesh E PA**, Lajish V L., "Colour Thresholding based Approaches to Lip Segmentation for Visual Speech Recognition."*L7th National Conference on Indian Language Computing (NCILC-2017),* 17-18 February 2017, Department of Computer Applications, CUSAT, Cochin.