

Viseme Identification and Analysis for Recognition of Malayalam Speech in Intense Background Noise

In Partial Fulfilment of the Requirements for the Degree of
Doctor of Philosophy
in
Physics

A Thesis Submitted by
BIBISH KUMAR K T

Under the Guidance of
Dr. R. K. Sunil Kumar
Assistant Professor
Department of Information Technology
School of Information Science and Technology
Kannur University, Kerala, India-670567



DEPARTMENT OF PHYSICS
GOVERNMENT COLLEGE MADAPPALLY
VADAKARA, CALICUT, KERALA,
INDIA – 673102
(Affiliated to University of Calicut)

NOVEMBER 2021

KANNUR UNIVERSITY
DEPARTMENT OF INFORMATION TECHNOLOGY
(School of Information Science and Technology)
KANNUR, KERALA 670567

Date...21/11/2022.

CERTIFICATE

This is to certify that the thesis entitled "**Viseme Identification and Analysis for Recognition of Malayalam Speech in Intense Background Noise**" is a report of original work carried out by **Mr. Bibish Kumar K. T.** under my supervision and guidance in the Department of Physics, Govt. College, Madappally, Vadakara, Calicut, Kerala and that no part thereof has been presented for the award of any other degree.



Dr. R. K. Sunil Kumar
Research Supervisor
Department of Information Technology
Kannur University

DECLARATION

I hereby declare that the work presented in this thesis entitled "**Viseme Identification and Analysis for Recognition of Malayalam Speech in Intense Background Noise**", is the original work done by me under the guidance of **Dr. R. K. Sunil Kumar**, Department of Information Technology, Kannur University, Kerala and no part thereof has been presented for the award of any other degree.

Madappally
November 2021



Bibish Kumar K T
Research Scholar
Govt. College, Madappally, Vadakara
Calicut, Kerala

ACKNOWLEDGEMENTS

This thesis is the confluence of many experiences I had at Govt. College, Madappally, from dozens of exceptional individuals whom I wish to thank. First and foremost, I wish to thank my guide Dr R. K. Sunil Kumar, Department of Information Technology, Kannur University, Kerala. It is an honour to be his first PhD student. I wish to express my deep gratitude to him for his insightful guidance and invaluable help to steer this work. He gave me the moral support and freedom to progress during the most difficult times.

I extend my sincere gratitude to Dr Udayakumar O. K., Principal, Govt. College, Madappally, Calicut, Kerala for providing the right resources and facilities in the department to accomplish my research work. I am extremely thankful to Prof. K. Suresh Babu, Former Head of my department and Dr P. Ramakrishnan, Former Principal of the same institution, for extending their valuable suggestions and support during the work tenure. I am deeply indebted to my department teachers, especially Dr Suneera T. P. (Head of the Department), Dr Harikrishnan G. and Dr Nithyaja B, for their inestimable support. I am extremely grateful to Dr Rajendran Edathumkara, Department of Malayalam, Govt. College, Madappally, Calicut for the generous donation of his time and valuable input for developing a linguistic background in my work and his open-mindedness by extending the discussion to other linguistic personalities. At the same time, I thank the linguistic people behind creating the Malayalam Phonetic archive owned by Thunchath Ezhuthachan Malayalam University, Kerala, India. I also express my gratitude to all teaching, non-teaching and research scholars in various departments in my college for their support and smiling faces.

I am extremely thankful to Dr Lajish V. L., Assistant Professor, Department of Computer Science, University of Calicut, for his guidance and providing the facilities of his research lab. I extend my gratitude towards his PhD students, especially Dr Vivek P., Assistant Professor of the same department and Dr Sandesh. E. PA for their keen support and collaboration during my work.

My special thanks to the M.Sc. project students, especially Sourabh S. and Devadathan S. R. for their indebted support during the recording and preparation of the database mentioned in this work. I extend my thanks to all my friends, especially Dr Subramanian, Department of Sanskrit, Zamorin's Guruvayurappan College, Calicut, Kerala and Dr Rakesh, Freelance Journalist.

Words are beyond to express my gratitude towards my colleagues, Mr Muraleedharan K. M., Mr Sunil John and Ms Aljinu Khadar K.V. for their consistent whole-hearted co-operation and encouragement during this work.

I owe a huge debt of gratitude to my parents and family members, whose unwavering encouragement and support served as a constant source of motivation for this work. Finally, I applaud myself for not losing my mind and completing the course despite receiving no stipend or fellowship.

Bibish Kumar K. T.

To my family & my research family....

ABSTRACT

In Malayalam, speech processing is still in its infancy stage, with only a few works focusing only on audio speech. Thus, there is a significant gap between the computational and linguistic aspects of the Malayalam language, making it an under-resourced language. It is well established that combining the audio and visual speech information improves the system's overall performance, especially in acoustically noisy conditions. However, few efforts, particularly in resourced languages, have attempted to recognise speech in noisy environments and struggled to meet satisfactory performance. This study aims to develop a Malayalam audio-visual speech recognition system that utilises visual speech information in noisy environments to improve the overall performance by proposing new algorithms. This is the first audio-visual speech recognition system in Malayalam. In this work, the visual speech unit (viseme) is utilised initially to recognise the Malayalam speech (phoneme) using the SVM classifier in noisy acoustical conditions having noise levels up to -20 dB.

A Malayalam audio-visual speech database named "MOZHI" is proposed to carry out this work, recorded in various environments for various research purposes, making it the first of its kind in the language. In this study the first category of the database is used which includes 30 speakers uttering 50 phonemes and 207 connected words that include all allophonic variations. Phonemes are the relatively distinct and fundamental utterances of a language. An allophone is a version of a phoneme that is phonetically distinct. The place and phonetic surroundings in the word generally characterises the allophones of the same phoneme. Viseme is a visual language unit that describes distinct speech movements of the visual speech articulators. Three noises, white Gaussian noise, pink noise, and red noise, were added to the segmented and labelled clean speech with a noise level ranging from 20 dB to -20 dB to

perform the proposed work in noisy situations. This study used statistical analysis of the duration of all phonemes and allophones to estimate audio-visual asynchrony for better understanding of coarticulation of Malayalam.

Since the proposed work relies mostly on visual speech, a comprehensive analysis of viseme is carried out to interpret the visual information from the lips. Linguistic expertise is used to choose the relevant frame for the underlined phoneme's visual appearance. A viseme is represented by two preceding and following frames from the chosen frame. Using the lip corner coordinate points and hue channel information from the HSV colour model, the lip region is extracted and made rotational and translational invariant. Linguistically 50 phonemes were grouped into 14 viseme classes. A data-driven approach is carried out by clustering the DCT features using K-means, and an optimum 16 viseme class were selected using the gap statistic method. Allophone-to-viseme mapping is performed by the data-driven approach, which emphasises the need for linguistic classification of allophones.

Noise-robust audio speech features are extracted for the proposed task as it is carried out in noisy conditions. This work proposed noise-robust acoustic features using the autocorrelation function (ACR). It includes ACR for fundamental frequency estimation, ACR Cepstrum for formant frequencies and ACR MFCC. The proposed audio speech features are compared with the most prominent methods and have outperformed even in intense background noise.

The SVM classifier is used in this study to identify the underlying phoneme by efficiently combining the extracted audio and visual speech features. To determine an optimum value for the SVM classifier hyperparameters and evaluate model performance, a nested stratified 5-fold cross-validation strategy is utilised, which minimises overfitting concerns and imbalance in the target class distribution. Since the proposed work relies on visual speech, a minor change in the visual appearance could cause the

underlying phoneme to be misjudged. A modified hierarchical approach addresses this problem by proposing a modified phoneme-to-viseme mapping or broad viseme class.

Precision, recall, F1-score, and accuracy are used to evaluate the presented systems' performance. In all metrics, the visual-only and audio-only speech recognition systems perform at 90% and 96%, respectively (in clean speech). Based on the broad viseme approach, the performance of the second phase of the AVSR (Audio-Video Speech Recognition) system has improved the average accuracy by 9%, 8% and 3% for audio-only speech recognition in white Gaussian noise, pink noise and red noise even at -20 dB noise level. It can be concluded that the improved performance of the proposed system, at intense background noise, is due to the process involved in the audio and visual feature extraction and the implementation aspects of the SVM classifier.

CONTENTS

	<i>Page No.</i>
List of Tables	ix
List of Figures	xi
List of Abbreviations	xvii
1. Introduction	1
1.1 Background	1
1.2 Motivation	3
1.3 Thesis outline	5
2. Literature Review	11
2.1 Introduction	11
2.2 Review on known Audio-Visual Speech Database	12
2.3 Review on Phoneme and Allophone Durational Modelling and Audio-Visual Speech Asynchrony	32
2.4 Review on known Viseme set	35
2.5 Review on Lip Segmentation	43
2.6 Review on AVSR systems in Intense background Noise	44
2.7 Conclusions	51
3. An Audio-Visual Speech Database in Malayalam – “MOZHI”	53
3.1 Introduction	53
3.2 Database Design	55
3.2.1 Language Material	56
3.2.1.1 Speech Production	57
3.2.1.2 Vowel Phonemes	58
3.2.1.3 Diphthong Phonemes	60
3.2.1.4 Consonant Phonemes	61
3.2.1.5 Allophones	69
3.2.2 Database Acquisition Hardware	73
3.2.3 Recording Setup	73

3.3 Audio and Video Segmentation and Labelling	78
3.4 Simulated Noisy Signal	84
3.4.1 White Gaussian Noise	86
3.4.2 Pink Noise	88
3.4.3 Red Noise	90
3.5 Audio-Visual Asynchrony	91
3.6 Conclusions	106
4. Malayalam Viseme set Identification	109
4.1 Introduction	109
4.2 Challenges in Viseme set Identification	111
4.3 Viseme set Formation Approaches	115
4.4 Selection of Relevant Frames	116
4.5 Colour-based Approach to Lip Segmentation	117
4.5.1 Image Thresholding	119
4.6 Lip Region Extraction	121
4.7 Malayalam Phoneme-to-Viseme / Many-to-One Mapping	126
4.7.1 Linguistic Approach	127
4.7.2 Data-driven Approach	130
4.7.2.1 Geometric Visual Features	131
4.7.2.2 Discrete Cosine Transform (DCT) Visual Features	132
4.7.2.3 Viseme set Formation by Clustering in the Parametric Space	133
4.7.3 Phoneme-to-Viseme Mapping – An Overview	145
4.8 Allophone-to-Viseme Mapping / Many-to-Many Mapping	148
4.9 Conclusions	153
5. Effect of Intense Background Noise on Acoustical Speech Parameters	155
5.1 Introduction	155
5.2 Pre-processing	157
5.2.1 Amplitude Normalization	157
5.2.2 Pre-emphasis Filtering	158
5.2.3 Framing	159
5.2.4 Windowing	161
5.2.5 Autocorrelation	163

5.3 Acoustical Speech Parameter – Fundamental Frequency (F0)	173
5.4 Acoustical Speech Parameter – Formant Frequencies	182
5.5 Acoustical Speech Parameter - ACR MFCC	190
5.6 Conclusions	194
6. Audio-Visual Speech Recognition using Support Vector Machine Classifier	195
6.1 Introduction	195
6.2 Support Vector Machine (SVM) Classifier	198
6.3 Problems and Strategies: Before and Audio Audio-Visual Integration	202
6.3.1 Multi-class Problem	202
6.3.2 Audio-Visual Integration Approach	204
6.3.3 Dataset Preparation and Hyper Parameter Estimation	210
6.4 Proposed Audio-Visual Speech Recognition System	222
6.5 Experimental Results	222
6.5.1 Visual-only Speech Recognition	223
6.5.2 Audio-only Speech Recognition	226
6.5.3 Audio-Visual Speech Recognition	229
6.6 Conclusions	237
7. Thesis Conclusions	239
7.1 Conclusions	239
7.2 Future Research Directions	242
Reference	245
List of Publications	269

LIST OF TABLES

<i>Table No.</i>	<i>Title</i>	<i>Page No.</i>
2.1	Summary of known Audio-Visual Speech Database	14
2.2	Summary of known Viseme sets	37
2.3	Summary of AVSR systems employed in negative SNR Conditions	46
3.1	Linguistic Classification of Vowel Phonemes	59
3.2	Linguistic Classification of Malayalam Consonant Phonemes	68
3.3	Malayalam Vowel and Diphthong Phonemes with its Allophonic variations	70
3.4	Malayalam Consonant Phonemes with its Allophonic variations	71
3.5	MOZHI Database Profile	77
3.6	Naming rule of MOZHI database files	83
3.7	SNR and Corresponding Relation between Signal Power and Noise Power	85
4.1	Performance HSV and CIE L*a*b* Colour Models in Lip Segmentation Problem	121
4.2	Linguistic Classification of Vowel and Diphthong Phonemes Visually	129
4.3	Linguistic Classification of Consonant Phonemes Visually	130
4.4.	Phoneme-to-viseme mapping using one frame based on the Geometric features	139
4.5.	Phoneme-to-viseme mapping using one frame based on the DCT features	140
4.6.	Phoneme-to-viseme mapping using three frames based on the Geometric features	141
4.7.	Phoneme-to-viseme mapping using three frames based on the DCT features	142

4.8.	Phoneme-to-viseme mapping using 5 frames based on the Geometric features	143
4.9.	Phoneme-to-viseme mapping using 5 frames based on the DCT features	143
4.10	Allophone-to-viseme mapping based on the DCT feature vector	150
4.11	Vowel Allophone-to-viseme mapping based on the DCT feature vector	151
4.12	Consonant Allophone-to-viseme mapping based on DCT feature vector	152
5.1	Variation of F0 estimation accuracy with SNR of differed noise types based on Praat, YIN, PEFAC, RAPT and ACR	179
5.2	Variation of F1 & F2 estimation accuracy with SNR of differed noise types based on Praat and ACR Cepstrum	189
5.3	Performance of ACR MFCC and MFCC with SNR of differed noise type for Short Vowel Phoneme	192
6.1	Performance of SVM Classifier for Different Kernel Functions	202
6.2	Modified Phoneme-to-Viseme Mapping	208
6.3	Grid search of 14 Viseme	214
6.4	Grid search of 50 Viseme	215
6.5	Grid search of 50 Phonemes	216
6.6	Confusion Matrix	223
6.7	Comparison of Performance of Visual-only, Audio-only and Proposed AVSR in different Acoustical Noisy Conditions	236

LIST OF FIGURES

<i>Figure No</i>	<i>Title</i>	<i>Page No.</i>
3.1	Schematic Diagram of Speech Production System	58
3.2	Time Domain Representation of Short Vowel Phonemes: (a) അ /a/, (b) ഇ /i/, (c) എ /e/, (d) ഒ /o/ and (e) ഉ /u	60
3.3	Time Domain Representation of Diphthong Phonemes: (a) ഐ /ai/ and (b) ഔ /au/	61
3.4	Place of Articulation	62
3.5	Time Domain Representation of Bilabial Consonant Phonemes: (a) പ /P/, (b) ഫ /p ^h /, (c) ബ /b/, (d) ഭ /b ^h / and (e) മ /m/	62
3.6	Time Domain Representation of Labiodentals Consonant Phoneme: വ /v/	63
3.7	Time Domain Representation of Dental Consonant Phoneme: (a) ത /t/, (b) ഥ /t ^h /, (c) ദ /d/, (d) ധ /d ^h / and (e) ന /n/	63
3.8	Time Domain Representation of Alveolar Consonant Phonemes: (a) ട /t̪/, (b) ന് /n/, (c) സ് /s/, (d) ര /r/, (e) റ /r̪/ and (f) ല /l/	64
3.9	Time Domain Representation of Retroflex Consonant Phonemes: (a) ട /t̪/, (b) ഠ /t̪ ^h /, (c) ഡ /d̪/, (d) ഡ /d̪ ^h /, (e) ണ /ɳ/, (f) ഷ /ʃ/, (g) ള /ʎ/ and (h) ഴ /z/	65
3.10	Time Domain Representation of Palatal Consonant Phonemes: (a) ച /c/, (b) ഛ /c ^h /, (c) ജ /j/, (d) ഝ /j ^h /, (e) ഞ /ɲ/, (f) ശ /ʃ/ and (g) യ /y/	65
3.11	Time Domain Representation of Velar Consonant Phonemes: (a) ക /k/, (b) ഖ /k ^h /, (c) ഗ /g/, (d) ഘ /gh/ and (e) ങ /ŋ/	66
3.12	Time Domain Representation of Glottal Consonant Phoneme: ഹ /h/	66
3.13	MOZHI recording setup	75
3.14	Block Diagram of Spectral Subtraction method	80

3.15	Speech Segmentation Process	81
3.16	Properties of White Gaussian Noise	87
3.17	Effect of White Gaussian Noise on Speech signal in Time and Frequency domain	88
3.18	Properties of Pink Noise	89
3.19	Illustration of effect of Pink Noise on Speech signal in Time and Frequency domain	89
3.20	Properties of Red Noise	90
3.21	Illustration of effect of Red Noise on Speech signal in Time and Frequency domain.....67	91
3.22	Time alignment of vowel phoneme അ /a/	94
3.23	Time alignment of consonant phoneme ത് /t/	95
3.24	Time alignment of vowel allophone എ [ye] in the word എവിടെ [evite]	96
3.25	Time alignment of vowel allophone എ [ey] in the word പിന്നെ [pinne]	97
3.26	Time alignment of vowel allophone എ [E] in the word വെളുത്ത [ve[utta]	97
3.27	Graphical representation of Durational Statistics of Malayalam phonemes from acoustic speech	99
3.28	Graphical representation of Durational Statistics of Malayalam phonemes from visual speech	100
3.29	Graphical representation of Durational Statistics of Malayalam vowel allophones from acoustic speech	102
3.30	Graphical representation of Durational Statistics of Malayalam vowel allophones from visual speech	102
3.31	Graphical representation of Durational Statistics of Malayalam consonant allophones from acoustic speech	103
3.32	Graphical representation of Durational Statistics of Malayalam consonant allophones from visual speech	104
3.33	Histogram of Asynchrony distribution in Malayalam Phonemes	106
3.34	Histogram of Asynchrony distribution in Malayalam Allophones	106
4.1	Viseme representation using a single frame (upper), three frames (middle) and five frames (lower)	114
4.2	Sequential frames for phoneme - ച് /c/	117
4.3	Manual labeling of landmark points in ROI	122

4.4	Estimation of Angle of Tilt	123
4.5	Rotated Image	123
4.6	Rotated Image along with landmark Points	124
4.7	Selected Frame of the Phoneme e /e/	124
4.8	Selected Frame of the Phoneme u /u/	125
4.9	Extracted Lip Region	125
4.10	Extraction of Seven Geometric Features from a Frame	132
4.11	DCT Coefficient Feature Extraction	134
4.12	A two cluster example: (a) data; (b) within sum of square function W_k ; (c) functions $\log(W_k)$ (O) and $E \cdot n \{\log(W_k)\}$ (E); (d) gap curve	137
4.13	Gap curve	138
4.14	Thresholded Lip Region	147
4.15	Centroid of Connected White Pixel Regions	147
4.16	Extracted Lip Region from HSV Colour Model	148
5.1	Normalized of Speech Signal	158
5.2	Pre-emphasized Speech Signal	159
5.3	Framing of Speech Signal	161
5.4	Hamming Window	162
5.5	Hamming Windowed Speech Signal	162
5.6	Power Spectrum and Autocorrelation function of Hamming windowed speech signal	163
5.7	Comparison of different Noises at 10 dB level in Time Domain and Autocorrelation Domain	167
5.8	Power Spectrum of Speech Signal and corresponding ACR	168
5.9	DDR Hamming Window	169
5.10	Illustrations of DDR Hamming window on One-sided ACR. (a) One-sided ACR. (b) DDR Hamming window. (c) Windowed ACR	170
5.11	Comparison of Power Spectrum. (a) Hamming windowed Speech signal. (c) DDR Hamming windowed ACR. Smooth spectral envelope of speech signal (b) and DDR Hamming windowed ACR (d)	171
5.12	Comparison of Speech Power Spectrum and ACR Power Spectrum in different Noisy Condition	172
5.13	Pre-processing Steps	173
5.14	The top figure shows the recorded microphone waveform over time, whereas the bottom waveform represents the	176

	laryngograph signal of the word “encyclopedias”. Speech includes voiced and unvoiced parts	
5.15	Block Diagram of proposed ACR F0 Estimation Algorithm	178
5.16	Performance of Different Pitch Estimation Methods in White Gaussian Noise added Speech signal	181
5.17	Performance of Different Pitch Estimation Methods in Pink Noise added Speech signal	181
5.18	Performance of Different Pitch Estimation Methods in Red Noise added Speech signal	182
5.19	Smooth Spectral envelope of vowel phoneme	183
5.20	Spectrogram of vowel Phoneme a/	183
5.21	F1 vs. F2 plot of short Malayalam vowel phonemes	184
5.22	Block Diagram of proposed ACR Cepstrum Algorithm	186
5.23	Cepstrum of Speech Segment	188
5.24	Mel-scale Filter Bank	190
5.25	Block Diagram of ACR-MFCC Feature Extraction Algorithm	192
5.26	Unified Frame work of Acoustical Feature Extraction	193
6.1	Hyperplanes for Classifying the Non-separable Data points	199
6.2	Linear Separating Hyperplane for the Non-separable Data points	200
6.3	Transformation of Non-Separable Data points in Feature Space to Separable Data points in Kernel Space	201
6.4	One-vs-All SVM Classifier	203
6.5	Hierarchical Approach for Audio Visual Integration	207
6.6	Modified Hierarchical Approach for Audio Visual Integration	209
6.7	Distribution of Observations in 14 Viseme Classes	211
6.8	Distribution of Observations in 50 Phonemes and 50 Visual representation of Phonemes	211
6.9	Identification of best fold ‘k’ for Nested Stratified Cross Validation	218
6.10	Visualization of Nested 5-fold Stratified Cross Validation Method for Visual-only and Audio-only Speech Recognition	219
6.11	Visualization of Nested 5-fold Stratified Cross Validation Method for the First Phase of Audio-Visual Speech Recognition	220

6.12	Visualization of Nested 5-fold Stratified Cross Validation Method for the Second Phase of Audio-Visual Speech Recognition	221
6.13	Schematic diagram of Proposed Audio-Visual Speech recognition system	222
6.14	Confusion Matrix of 14 Viseme Classes	224
6.15	Performance of First Phase of AV Speech Recognition (Viseme Recognition)	224
6.16	Performance of Visual-only Speech Recognition of Visual representation of 50 Phonemes	225
6.17	Performance of Audio-only Speech Recognition	226
6.18	Performance of Audio-only Speech Recognition in White Gaussian Noise	227
6.19	Performance of Audio-only Speech Recognition in Pink Noise	228
6.20	Performance of Audio-only Speech Recognition in Red Noise	229
6.21	Performance of Second Phase of AV Speech Recognition of Visual representation of 50 Phonemes	231
6.22	Performance of Second Phase of AV Speech Recognition of 50 Phonemes from Audio	232
6.23	Performance of AV Speech Recognition of 50 Phonemes in White Gaussian Noise	233
6.24	Performance of AV Speech Recognition of 50 Phonemes in Pink Noise	234
6.25	Performance of AV Speech Recognition of 50 Phonemes in Red Noise	235
6.26	Graphical Representation of Performance of Proposed AVSR system	237

ABBREVIATIONS

AAM	-	Active Appearance Model
ACR	-	Autocorrelation Function
ANN	-	Artificial Neural Network
ASM	-	Active Shape Model
ASR	-	Automatic Speech Recognition
AVSR	-	Audio-Visual Speech Recognition
CV	-	Consonant-Vowel
dB	-	Decibel
DCT	-	Discrete Cosine Transform
DDR	-	Double Dynamic Range
DWT	-	Discrete Wavelet Transform
ERM	-	Empirical Risk Minimisation
F0	-	Fundamental Frequency
FFT	-	Fast Fourier Transform
fps	-	frames per second
fs	-	Sampling frequency
HMM	-	Hidden Markov Model
GMM	-	Gaussian Mixture Model
IDFT	-	Inverse Discrete Fourier Transform
IFFT	-	Inverse Fast Fourier Transform
K-NN	-	K-Nearest Neighbors
LDA	-	Linear Discriminant Analysis
LPC	-	Linear Predictive Coding
MFCC	-	Mel Frequency Cepstral Coefficients
PCA	-	Principal Component Analysis
PSD	-	Power spectrum density

RAPT	- Robust Algorithm for Pitch Tracking
ROI	- Region of Interest
SD	- Standard Deviation
SNR	- Signal-to-Noise Ratio
SRM	- Structural Risk Minimisation
SVM	- Support Vector Machine
SWIPE	- Saw tooth Waveform Inspired Pitch Estimator
TEMU	- Thunchath Ezhuthachan Malayalam University
YAAPT	- Yet Another Algorithm for Pitch Tracking

CHAPTER 1

INTRODUCTION

1.1 Background

“Speech is the most sophisticated behaviour of the most complex organism in the known universe” – Moore, R. K. (2007) [1]. *Speech processing* is a distinct discipline that covers a broad area by incorporating various technologies and applications that enable humans to interact seamlessly with intelligent systems. One of the most challenging features especially for researchers in the speech processing realm is its multidisciplinary nature, which necessitates knowledge and expertise from various disciplines. Research and development of ASR (automatic speech recognition) systems began in the 1950s in Bell Labs, with simple digit recognition systems [2]. Since then, the recognition tasks have become more complex, from speaker-dependent isolated word recognition then speaker-independent continuous speech recognition with an extensive vocabulary up to spontaneous speech recognition in a noisy environment. Automatic speech recognition is now used in various areas of our lives, including speech-enabled electronic devices, navigation systems, and inquiry systems, etc.

The existing audio-only speech-based application system is only successful in relatively controlled surroundings or under calm conditions. Any intense background noise has a negative impact on the performance of such a system. Humans, on the other hand, can typically compensate for such ambiguity by incorporating additional speech information sources, such as the speaker's facial appearance. To recognise speech, humans integrate both visual and audio cues wisely. This effect is known as the “*McGurk effect*” [3]. Gesture or deformation of the mouth, especially lips, plays a vital role in overcoming

audio-only speech perception limitation. This idea of using visual cues inspired many researchers to devote their efforts to develop intelligent systems using audio and visual speech information, namely Audio-Visual Speech Recognition (AVSR).

Following Petajan's first attempt at a visual speech recognition system in 1984 [4], a wide range of AVSR systems has been produced, with results confirming that AVSR systems have a better recognition rate than audio-only speech recognition systems, especially in noisy environments. Because visual speech mainly conveys information about the point of articulation, which is easily confused when the audio modality is noisy. Even though these systems analyse and combine audio and visual information differently, they generally have a similar system architecture. Following is the general overview of how AVSR systems work. First, the speech signal is collected and recorded in both visual and audio formats. The signals were then processed for features that represent each. Finally, the two modalities were integrated to recognise the speech. Combining audio and visual information aims to provide the best feasible recognition outcomes in noisy environments.

Audio-visual speech-based applications were devised successfully for resourced languages like English and are growing rapidly. However, this technology for under-resourced languages does not attain an adequate pace in its research and development. There have been several technological improvements in speech processing for a range of languages in recent years; however, most researchers in this field are focused on developing speech-based applications for European languages, particularly English. The fascination with speech processing aroused speech researcher's focus on developing speech processing systems in their native language. Speech-based application systems have been developed in India for various languages, including Hindi, Tamil, Kannada, Oriya, and Punjabi. However, speech processing in Malayalam is still

in its early stages, with only a few works focusing only on audio speech. As a result, there is a significant gap between the computational and linguistic aspects of the Malayalam language, particularly in intense background noise. A significant amount of effort must be made in this regard to develop a noise-robust audio-visual speech recognition system in Malayalam.

1.2 Motivation

The speech recognition area has grown significantly in various contexts over the past 70 years of research and development by considering speech signals' inherent variability and diversity. Motivated by the ease of data transfer through spoken words rather than through typing using keyboards has propelled the human-machine interaction to the world of reality. Despite these technical achievements, ASR falls short of human performance in various tasks and situations, especially when there is much background noise. Furthermore, most ASR systems claim that users must have a basic understanding of English to communicate with them naturally and effectively. Thus, it has recognised the need for an effective human-machine interaction system based on regional languages. As a result, researchers are working hard to increase the accuracy of speech processing systems by incorporating visual speech information in these languages so that a user-friendly interactive system can be developed. Even though speech processing has been a focus of research in India for several years, Malayalam speech processing research is still in its infancy. As a result, having an effective Malayalam speech recognition system has great relevance in the present scenario.

Malayalam is a Dravidian language spoken in India that serves as the official language of Kerala. It has 50 phonemes and 106 allophones. Malayalam is syllabic with exact correspondence between spoken and written syllables. The lack of standard speech database and supporting computational linguistic resources is the major roadblock to Malayalam speech processing. A

phonetically rich audio-visual speech database in Malayalam suitable for various research in audio and visual speech processing is presented. The information like the duration of elementary speech units (phonemes and visemes), asynchrony between audio and visual speech signals etc., are essential for their combined use in speech processing systems. In addition, linguistic information of articulation of different phonemes and visemes and their coarticulation variations is also vital for developing such systems. These informations are language-specific and are unexplored in many regional languages, including Malayalam. This research delves into the durational analysis of Malayalam audio-visual speech signals to better understand Malayalam's coarticulation nature.

This will be the first study in Malayalam that addresses visual speech processing. It is necessary to establish the relationship between phoneme and its visual equivalent by establishing the viseme set in Malayalam to analyse visual speech and make decisions based on audio-visual speech data. Language-specific exploration is necessary to analyse visemes, the visual equivalent of phonemes. Since the lip region of a talking face contains most visual speech information, extracting the lip region is critical for boosting the system's accuracy. Furthermore, there is no unique agreement on the lip extraction algorithm, especially in the Indian context, where low lip-skin colour contrast and facial hairs are major issues. While standard parameters for characterising an audio speech signal are well understood, the visual speech parameters that are most beneficial for automatic speech recognition are still being disputed. This work investigates the highlighted problems associated with the visual speech signal to capture relevant articulatory information from the talking face.

Extraction of noise-robust acoustic features from speech signals is still an active research area. Researchers have introduced new noise-robust feature

extraction algorithms and modified existing traditional methods. Such features must have acceptable accuracy even with ambient noise and use computationally simple algorithms in real-time applications. However, few works have even addressed speech signal processing at negative dB (decibel) level noise and failed to achieve satisfactory results. Fundamental frequency, formant frequencies, and MFCCs (Mel Frequency Cepstral Coefficients) are the most often utilised acoustical parameters in literature. This research focuses on extracting a noise-resistant version of these features using relatively simple noise-robust algorithms and evaluating their performance in various noisy speech signals, even up to -20 dB. The experimental result of this type has significant importance in real-time speech-based applications.

The effective merging of information extracted from audio and visual speech signals remains an open problem. The system can effectively adapt to any noisy environment by optimising the weights associated with each modality. The major goal is to develop a Malayalam speech recognition system that utilises visual speech information in noisy environments to improve overall performance by proposing new algorithms and models. The motivation for this research stems from the fact that little attempt was reported to build a multimodal Malayalam speech recognition system. This research proposes solutions for the problems mentioned above to support the development of Malayalam AVSR systems.

1.3 Thesis Outline

This research aims to investigate the performance of a Malayalam speech recognition system in acoustically noisy conditions by utilising visual speech information. Experimental findings are used to validate the proposed methods and system components, which are thoroughly discussed. The rest of this thesis is organised as follows.

The purpose of Chapter 2 is to lay the groundwork for the subsequent chapters. A detailed review of the known audio-visual speech database, including features like durational analysis and audio-visual asynchrony is discussed here. Phoneme-to-viseme mapping, allophone-to-viseme mapping and visual feature extraction is discussed in detail in the following sections. In the next section, studies on various lip region extraction techniques are reviewed. The last section reviewed the audio-visual speech recognition systems in intense background noise.

The third chapter proposes a Malayalam audio-visual speech database, “MOZHI”. In Malayalam, there are linguistically classified 50 phonemes based on articulatory points and positions and 106 allophones. Phonemes are the relatively distinct and fundamental utterances of a language. An allophone is a version of a phoneme that is phonetically distinct. The purpose of a speech database depends intensely on the language material, speaker population, and the quality of the recording, which should be optimally adapted to attain the prime goal. This database consists of 3 categories of recording, which can be used for various research works. There is one category of audio-only speech database and two categories of audio-visual speech database. This database is designed to aid research into the effects of age on speech, noisy speech processing, visual speech processing, viseme-based speech synthesis, and lip synchronisation utilising audio-visual speech asynchrony, among other topics. The segmentation and labelling of recorded audio and video speech are the most crucial task once the database is recorded. The segmentation is done semi-automatically, starting with a spectral subtraction method to remove the noisy background content while recording. The labelling process, which involves utterance, speaker, and noise information with alphanumeric characters, was carried out automatically. Any real-world speech-based application must consider its performance in various noisy environments with varying signal-to-noise ratios (SNR). This research created a realistic environment by

incorporating several noisy signals such as white Gaussian noise, pink noise, and red noise, with noise levels ranging from 20 to -20 dB. The durational analysis of phonemes and allophones from the audio and visual speech signal is covered in the next part, which includes a comprehensive statistical analysis. Duration modelling is the preliminary task in phoneme or viseme level speech processing applications like speech recognition and speech synthesis system. Based on the durational analysis, audio-visual asynchrony is estimated for phonemes and allophones to study the coarticulation nature of the Malayalam language.

The fourth chapter delves into an in-depth investigation into the visual cues, or viseme, which can be applied to bi-modal speech application in Malayalam. Viseme is a visual language unit that describes distinct speech movements of the visual speech articulators. This chapter has three main sections; the first one is lip region extraction. Lip segmentation and tracking are the most important functions of a lip-reading system since the lip is the most active articulator and holds the most visual information. In the initial stage, the lip region is extracted manually using 36 landmark points. Later, the lip region is extracted by utilising the hue channel information from the HSV colour model. The Lip region is also made rotational and translational invariant by utilising lip corner coordinates information. The second section aims to identify the viseme set in the Malayalam language using a linguistic involved data-driven approach. The relevant frames were chosen based on articulatory norms and linguistic expertise. Two preceding and following frames from the linguistically chosen frame were also selected to encode the time evolution of the underlined phoneme's visual speech. Seven geometric features and 20 Discrete Cosine Transform (DCT) coefficients are extracted from each image to represent the visual speech signals. Then the viseme map is developed by categorising visual feature vectors using the K-means clustering and Gap statistic methods. The last section deals with the creation of allophone-to-

viseme mapping. This chapter concludes with the necessity of developing an allophone-to-viseme mapping linguistically to address world-level visual speech analysis issues.

The fifth chapter investigates the analysis of audio speech signals in various noisy environments, which is a core component of today's speech-based applications. The most challenging problem for any speech recognition system is to extract noise-resistant features even in ambient noise. The most popular acoustical features used by research groups are fundamental frequency (F0), formant frequencies (F1 and F2) and Mel frequency cepstral coefficients (MFCCs). An improvisation of these features that is noise-tolerant even at negative noise levels is proposed. To achieve noise robustness, a noisy speech signal in time-domain is transformed into the autocorrelation domain and then subjected to certain operations to extract these features. The proposed features performance is compared with prominent algorithms in the speech processing field in different noisy speech signals at various noise levels, and its dominance is verified experimentally.

The sixth chapter discusses an audio-visual speech recognition system that employs support vector machine (SVM) classifiers. SVMs, unlike some HMM modifications that reduce the empirical risk on the training set, also reduce the structural risk, resulting in improved generalisation ability even with minimal training data. In the presence of noise, the maximum margin solution allows SVMs to dominate most nonlinear classifiers, which has long been a challenge in speech recognition. Various issues that arise before and after the integration of audio, and visual speech information is discussed in detailed. The SVM classifier's main feature is selection of optimal hyperspace parameters such as margin width from penalty parameter C and margin shape, which are determined experimentally using grid search. The SVM method is modified by implementing nested cross-validation so that each observation is either training,

validation, or test data. The audio-visual fusion strategy is one of the most crucial aspects of any speech recognition system. Stream weights of each stream is estimated from its corresponding reliability measures, thereby favouring the reliable stream depending on the noise levels. This work adopted a hierarchical fusion method in which viseme is recognised first, then the underlined modified phonemes in that viseme class is recognised. Thus, it will reduce the computational complexity without computing the likelihood of each phoneme, even in intense background noise. The algorithms proposed in previous chapters for building an audio-visual speech recognition system for the Malayalam language are validated through experiments.

The seventh chapter concludes the thesis by summarizing the conclusions and suggestions to extend the research work. Reference is provided after this chapter and the details of the author's publications at last.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Many human-machine communication technologies make use of speech recognition. However acoustic noise, which is common in real-world applications, makes automatic speech recognition challenging. There are two approaches to deal with this problem. The first is to extract noise-resistant audio features from the distorted audio speech input. The second is to employ the visual modality, such as highlighting key characteristics of the speaker's face or lips, to improve recognition. As a result, several attempts have been made in recent years to extract noise-tolerant audio features and incorporate visual information into the speech recognition process. Thus, significant milestones in this field must be highlighted.

Although speech recognition research began in the 1950s, the 1980s marked a significant increase in the technology associated with ASR. In the 1980s, speech research was portrayed by a technological transition away from template-based methods to statistical modelling methodologies. Hidden Markov Models were employed in the development of more realistic automatic continuous speech recognition systems [5]. The introduction of Mel frequency Cepstral Coefficients (MFCCs) [6], one of the most notable metrics that characterise the audio speech signal, is another significant advance in ASR. Pollack and Sumbly [7], on the other hand, started the theoretical work for merging visual cues with noisy acoustic speech in 1954. The McGurk effect also demonstrated the favourable benefits of visual speech information on speech perception. Petajan created the first visual speech recognition system in 1984. Artificial Neural Networks were used to experiment with automatic

speech recognition in the late 1980s, and reliable results were claimed. During this timeframe, many effective systems were developed; in 1990, Carnegie Mellon University released Sphinx, an open source-based voice recognition system [8]. The introduction of convolutional neural networks, which fuelled deep learning, was another advance in this subject in 1994. Deep learning research advances slowly at first, but the demand for fast computational power sparks interest, and a significant breakthrough was made in 2009 [9]. Google's speech recognition engine delivered ground-breaking results through its Google Voice-based search service in 2011. All of today's commercially available speech recognition systems were designed with language-related properties in mind, allowing for trustworthy results in noise-free or low-noise conditions. Implementing a speech recognition system for useful applications in loud situations and public areas is still a research problem.

This research aims to create an audio-visual speech recognition system that primarily relies on the visual speech unit (viseme) to recognise the underlying phoneme in noisy environments. The background and current scenario of its connected methodologies or approaches must be thoroughly investigated to implement this system. This chapter discusses a comprehensive review of relevant works. Section 2.2 summarises the growth of a known audio-visual speech database. Section 2.3 reviews the work in phoneme and allophone modelling and audio-visual speech asynchrony. Section 2.4 gives a detailed list of available viseme sets and their attributes. Section 2.5 describes a brief review of lip segmentation techniques. Section 2.6 presents the performance of an audio-visual speech recognition system in intense background conditions. Section 2.7 concludes the literature review.

2.2 Review on known Audio-Visual Speech Database

Over the years of development and unceasing research in speech-based applications has endorsed the dearth of standard audio-visual speech databases.

The need for diversity in resources and massive storage capacity is the main challenge that produces the remarkable difference in the statistic between audio-only and audio-visual speech databases. Creating an exhaustive and methodically arranged audio-visual speech database built up a worldwide acceptance in the research community. An audio speech signal can capture at high quality at a relatively low cost. However, a high-quality visual speech can capture the dynamic movements of the lip accurately, which is relatively large. Massive storage space is needed for the storage and distribution of audio-visual speech database than the audio-only speech databases. Instead of giving a detailed description of the characteristics of the existing database, a tabular representation of the available audio-visual speech databases is presented. It offers a clear picture of parameters and comparison among different others. Table. 2.1 summarize the historical background of the audio-visual speech database in terms of gender distribution, speech corpus, hardware setup and some distinctive features. The history begins with Petajan 1988 for lip-reading digit recognizer. This work presents a historical sketch in different aspects of the available audio-visual speech database from 1988. Even though many papers have reviewed some of the audio-visual speech databases, this work gives an overall view of creating the databases in resourced and under-resourced language, and technological advances occur in the recording devices and approaches. A new face in this realm gets a proper insight into the quantity and quality of basic requirements and resources needed to build an audio-visual speech database for a task in their mother language.

Table 2.1 Summary of known Audio-Visual Speech Database

Database – Year	Speaker (F, M)	Corpus - Repetition	Video Parameters (Pixel size, fps)	Audio Parameters (Sampling frequency)	Special Features
TULIPS1 1995 [10]	12 (9,3)	<ul style="list-style-type: none"> • First four English digit – twice. 	<ul style="list-style-type: none"> ▪ 100x75, 30 fps. ▪ Mouth region. 	11.1 kHz.	<ul style="list-style-type: none"> ➤ Isolated digit recognition.
DAVID 1996 [11]	124	<ul style="list-style-type: none"> • English language. • Isolated digits. • English-alphabet E-set. • Video-conference control commands. • ‘VCVCV’ nonsense utterances. 	<ul style="list-style-type: none"> ▪ 640x480, 30 fps. ▪ Full face. ▪ Frontal view. 	—	<ul style="list-style-type: none"> ➤ Speech/Person recognition. ➤ Contain 4 corpus with different research theme. Complex background and variable illumination. ➤ Contain individual speaker and more than one speaker during recording. ➤ Lack of head pose and facial expression variations.
M2VTS 1997 [12] Multi Modal Verification for Teleservices and Security applications	37	<ul style="list-style-type: none"> • French language. • Numbers (0 to 9) – 5 times. 	<ul style="list-style-type: none"> ▪ 286x350, 25 fps. ▪ Full face. ▪ Frontal view. 	48 kHz	<ul style="list-style-type: none"> ➤ Speech verification, face recognition. ➤ Mostly French Speakers. ➤ Head rotation (left, right, up and down). ➤ Presence of glasses and hats.

<p>XM2VTSDB 1999 [13] Extended M2VTS Database</p>	<p>295</p>	<ul style="list-style-type: none"> • Three sentences (numbers and word) – twice. 	<ul style="list-style-type: none"> ▪ 720x576, 25 fps. ▪ Full face. ▪ Frontal view. ▪ 2 Cameras used. 	<p>32 kHz</p>	<ul style="list-style-type: none"> ➤ Personal identity verification. ➤ Head rotation (left, right, up and down). ➤ Recorded in extremely controlled condition. ➤ Text dependent. ➤ http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/
<p>AMP/CMU 2001 [14] Advanced Multimedia Processing Lab</p>	<p>10 (3,7)</p>	<ul style="list-style-type: none"> • 78 Isolated words (date and time, month, day and miscellaneous) – 10 times. 	<ul style="list-style-type: none"> ▪ 720x480, 25 fps. ▪ Full face. ▪ Frontal view. 	<p>16 kHz</p>	<ul style="list-style-type: none"> ➤ Lip reading. ➤ Presence of glasses and hats. ➤ Recorded in controlled situation.
<p>AV Letters 2002 [15]</p>	<p>10 (5, 5)</p>	<ul style="list-style-type: none"> • English language. • Isolated letters (A to Z) – 3 times. • Total 780 utterances. 	<ul style="list-style-type: none"> ▪ 376x288, 25 fps. ▪ Full face. ▪ Frontal view. 	<p>22.05 kHz</p>	<ul style="list-style-type: none"> ➤ Speech recognition.

CUAVE 2002 [16] Clemson University Audio-Visual Experiments	36 (19, 17) Speaker pairs - 20	<ul style="list-style-type: none"> • English language. • Isolated digits. • Connected digits. • Total 7000 utterances. 	<ul style="list-style-type: none"> ▪ 720x480, 29.97 fps. ▪ Shoulder and head. ▪ Frontal and Profile view. 	16 kHz	<ul style="list-style-type: none"> ➤ Speaker independent digit recognition. ➤ Speaker independent database. ➤ Contain individual speaker, speaker pairs and moving speakers. ➤ Head movement in side-to-side, back-and-back. ➤ Presence of glasses, facial hairs and hats. ➤ Fit to one DVD-data disk. ➤ http://universal.elra.info/product_info.php?cPath=25&products_id=2228
VidTIMIT 2002 [17]	43 (19, 24)	<ul style="list-style-type: none"> • English language. • 10 TIMIT sentences per speaker. • First 2 sentences are same for all but remaining 8 are unique. 	<ul style="list-style-type: none"> ▪ 512x384, 25 fps. ▪ Frontal view. 	32 kHz	<ul style="list-style-type: none"> ➤ Multi-modal person verification. ➤ Data acquisition with 3 sessions. ➤ Extended head rotation. ➤ Change in speaker appearance and voice during each session. ➤ Variability in camera zoom factor and background noise. ➤ http://conradsanderson.id.au/vidtimit/

DUTAVSC 2002 [18]	8 (1, 7)	<ul style="list-style-type: none"> • Dutch language. • POLYPHONE corpus. • Phonetically rich sentences. • Connected digits. • Spelling. • Application driven utterance. 	<ul style="list-style-type: none"> ▪ 384x288, 25f fps. ▪ Frontal view. ▪ Lower face view. 	44 kHz	<ul style="list-style-type: none"> ➤ Audio visual speech recognition.
BANCA 2003 [19]	52 (26, 26) for each language class.	<ul style="list-style-type: none"> • 4 Languages-English, French, Italian and Spanish. • Numbers. • Names. • Addresses. • Date of birth. 	<ul style="list-style-type: none"> ▪ 720x576, 25 fps. ▪ Shoulder and head. ▪ Frontal view. ▪ 2 Cameras used. 	* 32 kHz. * 2 Microphone used.	<ul style="list-style-type: none"> ➤ Multi-modal identity verification. ➤ Recorded in controlled, degraded and adverse condition. ➤ Text independent. ➤ Lack of head pose and facial expression variations. ➤ http://www.ee.surrey.ac.uk/CVSSP/banca/

<p>AVICAR 2004 [20] Audio-Visual Speech in a Car.</p>	<p>100 (50, 50)</p>	<ul style="list-style-type: none"> • English language. • Isolated digits. • Isolated letters. • Phone numbers. • TIMIT sentences. • Total 59,000 utterances. 	<ul style="list-style-type: none"> ▪ 720x480, 30 fps. ▪ Full face. ▪ 4 Frontal views. ▪ 4 Camera arrays. 	<ul style="list-style-type: none"> * 48 kHz. * 8 Microphone arrays. 	<ul style="list-style-type: none"> ➤ Speech recognition in car. ➤ 60% American English others Latin American, European, East Asian and South Asian. ➤ Recorded in 5 noisy condition (automotive noise). ➤ http://www.isle.illinois.edu/st/AVICAR/
<p>AV-TIMIT 2004 [21]</p>	<p>223 (106, 117)</p>	<ul style="list-style-type: none"> • 450 TIMIT-SX sentences. • Each speaker utter 20 sentences. • First sentences are common and other 19 sentences are different. 	<ul style="list-style-type: none"> ▪ 720x480, 30 fps. ▪ Full face. ▪ Frontal view. 	<ul style="list-style-type: none"> 16 kHz. 	<ul style="list-style-type: none"> ➤ 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.

<p>AVOZES 2004 [22] Audio Video OZtralian English Speech</p>	<p>20 (10, 10)</p>	<ul style="list-style-type: none"> • Australian English language. • Digits. • Words. • Phrases. • Continuous word. • Total of 56 sequences per speaker without repetition. 	<ul style="list-style-type: none"> ▪ 720x480, 29.97fps. ▪ Full face. ▪ Stereo view. 	<p>48 kHz.</p>	<ul style="list-style-type: none"> ➤ Modular approach database- each module addresses specific task. ➤ 6 Modules- 1 general module (speaker independent) and 5 speaker specific module. ➤ Native speakers of Australian English. ➤ Presence of glasses, facial hairs and lip highlighter. ➤ Speaker personal data acquisition mode. ➤ http://users.cecs.anu.edu.au/~roland/avozes.html
<p>MANDARIN CHINESE 2004 [23]</p>	<p>225</p>	<ul style="list-style-type: none"> • Chinese language. • Continuous speech. • Total 17,000 utterances. 	<ul style="list-style-type: none"> ▪ 720x576, 25 fps. ▪ 768x576, 25 fps. ▪ 7 Cameras used. 	<p>* 48 kHz. * 12 Microphones used.</p>	<ul style="list-style-type: none"> ➤ Audio-visual speech/speaker recognition and 3Dface modelling. ➤ Recorded in 2 sessions.

VALID 2005 [24]	106 (29, 77)	<ul style="list-style-type: none"> • XM2VTS speech corpus. 	<ul style="list-style-type: none"> ▪ 720x576, 25 fps. ▪ Full face with shoulder. ▪ Frontal view. 	32 kHz.	<ul style="list-style-type: none"> ➤ 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.
UWB-04- HSCAVC 2006 [25] University of West Bohemia- 2004-Hundred Speakers Czech Audio- Visual Corpus	100 (61, 39)	<ul style="list-style-type: none"> • Slavonic language (Czech and Russian). • 200 Sentences (50 shared and 150 unique). 	<ul style="list-style-type: none"> ▪ 720x576, 25 fps. ▪ Full face. ▪ Frontal view. 	* 44 kHz. * 2 Microphones used.	<ul style="list-style-type: none"> ➤ Audio-visual speech recognition. ➤ Visual speech parameterizations. ➤ Recorded in laboratory condition. ➤ Constant illumination and static head position. ➤ Mean age is 22.
GRID 2006 [26]	34 (16, 18)	<ul style="list-style-type: none"> • English language. • Command sentences. • Each sentence contains six-word sequence. • Total corpus size 34,000. 	<ul style="list-style-type: none"> ▪ 720x576, 25 fps ▪ Full face. ▪ Frontal view. 	25 kHz.	<ul style="list-style-type: none"> ➤ Speech recognition. ➤ Recorded in lab environment with blue background. ➤ Mean age is 27. ➤ http://spandh.dcs.shef.ac.uk/gridcorpus/

AV Letters 2 2008 [27]	5	<ul style="list-style-type: none"> • English language. • 26 Isolated letters – 7 times. 	<ul style="list-style-type: none"> ▪ 1920x1080, 50 fps. ▪ Full face. ▪ Frontal view. 	8 kHz.	<ul style="list-style-type: none"> ➤ Speech recognition. ➤ High-definition version of AV Letter database.
UWB-07-ICAVR 2008 [28] University of West Bohemia-2007-Impaired Conditions audio visual speech Recognition	50 (25, 25)	<ul style="list-style-type: none"> • Czech language. • 200 Sentences (50 shared and 150 unique). • Total 10,000 continuous utterances. 	<ul style="list-style-type: none"> ▪ 720x576, 50 fps (high quality). ▪ 640x480, 30 fps (low quality). ▪ 2 Cameras used. 	<ul style="list-style-type: none"> * 44 kHz. * 2 Microphones used. 	<ul style="list-style-type: none"> ➤ 6 types of illumination condition ➤ Average age is 22.
IV2 2008 [29]	300	<ul style="list-style-type: none"> • 15 French sentences. 	<ul style="list-style-type: none"> ▪ 780x576, 25 fps (high quality). ▪ 640x480, 25 fps (low quality). ▪ Full face. ▪ Frontal and profile view. 	2 Microphone used.	<ul style="list-style-type: none"> ➤ 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.

DXM2VTS 2008 [30] Damascened XM2VTS	295	<ul style="list-style-type: none"> • XM2VTS database. • Additional videos containing several degradation level of background noises. 	<ul style="list-style-type: none"> ▪ 720x576, 25 fps. ▪ Full face. ▪ Frontal view. ▪ 2 Cameras used. 	32 kHz.	<ul style="list-style-type: none"> ➤ Face recognition. ➤ Internal video distortion (blur, salt and pepper and rotation). ➤ External video distortion (zooming and dynamic background noise). 	
IBM Smart-Room 2008 [31]	38	<ul style="list-style-type: none"> • Connected digit strings. • Total 1661 utterances. 	<ul style="list-style-type: none"> ▪ 368x240, 30 fps. ▪ Full face. ▪ Frontal and profile view. 	<ul style="list-style-type: none"> * 22 kHz. * 2 	Microphones used.	➤ Lip-reading system.
HIT-AVDB-II 2008 [32] Harbin Institute of Technology Audio Visual Speech Database II	30 (15, 15)	<ul style="list-style-type: none"> • Digits. • Chinese poems. • Tongue twisters of Chinese and English. • Greek alphabets. • Music notes. • Mandarin vowels. 	<ul style="list-style-type: none"> ▪ 720x576, 25 fps. ▪ 4 Cameras used. ▪ 4 views- frontal, profile, 30⁰ and 60⁰. 	—	<ul style="list-style-type: none"> ➤ Visual speech, Biometrics, lip tracking and multi view. ➤ Database witness emotions, fast mouth movements and tunes. ➤ Recorded in 3 sessions in different day time to capture varying speaker appearances and background. ➤ Presences of spectacles and hair ornaments. 	

OuluVS 2009 [33]	20 (3, 17)	<ul style="list-style-type: none"> • English language. • 10 daily use short phrases – 9 times. • Total 817 sequences. 	<ul style="list-style-type: none"> ▪ 720x576, 25 fps. ▪ Full face. ▪ Frontal view. 	No audio	<ul style="list-style-type: none"> ➤ Visual only speech recognition.
WAPUSK20 2010 [34]	20 (9, 11)	<ul style="list-style-type: none"> • 100 GRID database sentences. • Total 2000 sentences. 	<ul style="list-style-type: none"> ▪ 640x480, 32 fps. ▪ Full face. ▪ Frontal view. 	<ul style="list-style-type: none"> * 16 kHz. * 4 audio channels. 	<ul style="list-style-type: none"> ➤ Stereoscopic video. ➤ Speakers – 2 England, 1 Greece, 1 Kazakhstan and 1 Spain. ➤ All other native German speakers. ➤ Mean age is 29.
AVA II 2010 [35]	14 (7,7)	<ul style="list-style-type: none"> • Persian language. • Phonemes, Phonemic combinations (cv, vc, vcv), 20 sentences and digits. 	<ul style="list-style-type: none"> ▪ 720x576, 25 fps. ▪ 3 cameras used. ▪ Frontal view, profile view and lip area. ▪ Full face with shoulder and lip region. 	<ul style="list-style-type: none"> * 48 kHz. * 2 microphones used. 	<ul style="list-style-type: none"> ➤ 23 consonants and 6 vowels. ➤ Speakers have Tehrani accent. ➤ Recorded in studio environment with lighting system and blue curtain.

BL-Database 2011 [36] Blue Lips- Database	17 (8, 9)	<ul style="list-style-type: none"> • French language. • 238 sentences. • Diphone rich utterances. 	<ul style="list-style-type: none"> ▪ 576x720, 25 fps (front view). ▪ 640x480, 30 fps (side view). ▪ 640x480, 30 fps (depth camera). ▪ Full face. ▪ Frontal and side view. 	<ul style="list-style-type: none"> * 44.1 kHz. * 2 <p>Microphones used.</p>	<ul style="list-style-type: none"> ➤ Audio-Visual speech recognition. ➤ Recorded in 2 sessions- First sessions for 2D analysis and second session for 3D analysis of mouth movement. ➤ Native French speakers. ➤ Blue lipstick used.
UNMC-VIER 2011 [37]	123 (49, 74)	<ul style="list-style-type: none"> • 11 XM2VTS sentences. • Sequence of numerals. 	<ul style="list-style-type: none"> ▪ 708x640, 25 fps (3 high quality cameras). ▪ 320x240, 29 fps (low quality camera). ▪ 320x240, 15 fps (low quality camera). ▪ 5 Cameras used. ▪ Full face. ▪ Frontal and left profile view. 	<ul style="list-style-type: none"> * 48 kHz (From high quality camera). * 44 kHz (From low quality camera). * 32 kHz (From low quality camera). * 22 kHz (Audio device). 	<ul style="list-style-type: none"> ➤ Audio-Visual speech/speaker recognition. ➤ Contain multiple visual variation in the same video recording. ➤ Visual variations- illumination, expression, head pose, background and resolution. ➤ Head rotation- left-to-right and up and down. ➤ Spoken in normal and slow speech pace. ➤ Recorded in controlled and uncontrolled environment with different devices.

AusTalk 2011 [38]	1000	<ul style="list-style-type: none"> • Australian English. • Multiple read and spontaneous speech tasks. 	<ul style="list-style-type: none"> ▪ 640x480, 48 fps 	—	<ul style="list-style-type: none"> ➤ Applied for Speaker verification, Audio-Visual speech Recognition, Forensic Speaker recognition. ➤ http://austalk.edu.au
MoBio 2012 [39]	152 (52, 100)	<ul style="list-style-type: none"> • English language. • 32 questions (short response questions, short response free speech, set speech, and free speech). 	<ul style="list-style-type: none"> ▪ 640x480, 16-30 fps. 	48 kHz.	<ul style="list-style-type: none"> ➤ Mobile-based biometric system. ➤ Database almost captured from mobile devices. ➤ High variability in pose and illumination. ➤ http://www.idiap.ch/dataset/mobio
LILiR 2012 [40] Language Independent Lip Reading	20	<ul style="list-style-type: none"> • Resource management corpus. • Total 200 sentences per speaker. 	<ul style="list-style-type: none"> ▪ 5 cameras used (2 HD and 3 SD). ▪ Full face. ▪ 5 views- frontal, profile, 30⁰, 45⁰ and 60⁰. 	—	<ul style="list-style-type: none"> ➤ Continuous speech recognition.
AVAS 2013 [41] Audio-Visual Arabic Speech	50	<ul style="list-style-type: none"> • Arabic language. • 36 daily words. • 13 casual phrases. 	<ul style="list-style-type: none"> ▪ 640x480, 30 fps. ▪ Full face. ▪ Frontal view. 	48 kHz.	<ul style="list-style-type: none"> ➤ Audio-Visual speech/speaker recognition. ➤ Visual variations-4 illumination condition and 5 head pose variations. ➤ First database in Arabic language.

†Oriya Digit Database 2013 [42]	15 (10,5)	<ul style="list-style-type: none"> • Oriya language. • Digits-4 times. 	<ul style="list-style-type: none"> ▪ 1024x768, 30 fps. ▪ Frontal view. ▪ Full face. 	16 kHz	<ul style="list-style-type: none"> ➤ Digit recognition. ➤ Lip tracking. ➤ 3 type of noise added to speech signal.
AGH 2015 [43]	166 (one third female)	<ul style="list-style-type: none"> • Polish language. • Isolated words and Numbers. • Total 1,17,450 words. 	<ul style="list-style-type: none"> ▪ 1920x1080, 50 fps. 	44.1 kHz.	<ul style="list-style-type: none"> ➤ Automatic speech recognition and Text-to-speech systems. ➤ Largest audio-visual Polish corpus.
OuluVS2 2015 [44]	53 (13, 40)	<ul style="list-style-type: none"> • English language. • Continuous digits. • Phrases. • TIMIT sentences. 	<ul style="list-style-type: none"> ▪ 1920x1080, 30 fps (From 5 HD camera). ▪ 640x480, 100 fps (From frontal HS camera). ▪ Six cameras used. ▪ Full face. ▪ 5 views- frontal, profile, 30⁰, 45⁰ and 60⁰. 	High quality audio.	<ul style="list-style-type: none"> ➤ Multi-view audio-visual database. ➤ No native English speakers. ➤ Speakers- European, Chinese, Indian/Pakistan, Arabian and African. ➤ Neutral facial expression and static head pose. ➤ Simultaneous recording by 6 cameras from 5 different views. ➤ http://www.ee.oulu.fi/research/imag/OuluVS2/

TCD-TIMIT 2015 [45]	62 (30, 32)	<ul style="list-style-type: none"> • 6913 phonetically rich TIMIT sentences. 	<ul style="list-style-type: none"> ▪ 1920x1080, 30fps. ▪ Cameras used. ▪ 2 views- frontal and 30^o. 	16 kHz.	<ul style="list-style-type: none"> ➤ Phone recognition in continuous speech. ➤ Two class of speakers- 3 professional lip speakers (female) and non-lip speakers. ➤ Presence of glasses and piercings. ➤ https://sigmedia.tcd.ie/TCD-TIMIT/
†AMAUV 2015 [46] Aligarh Muslim University Audio Visual	100	<ul style="list-style-type: none"> • Hindi language. • 10 sentences out of which 2 sentences are common to all speaker. 	<ul style="list-style-type: none"> ▪ 640x380, 25 fps. ▪ Frontal view up to shoulder. 	44.1 kHz	<ul style="list-style-type: none"> ➤ Audio-Visual speech recognition. ➤ Phonetically balanced Hindi database.
†vVISWa 2016 [47] Visual Vocabulary of Independent Standard Words	58 (20, 38)	<ul style="list-style-type: none"> • Marathi, Hindi and English languages. • Isolated words–10 times. • Continuous words–10 times. • Total 2, 96,960 words. 	<ul style="list-style-type: none"> ▪ 720x576, 25 fps. ▪ Cameras used. ▪ 3 views- frontal, profile and 30^o. 	—	<ul style="list-style-type: none"> ➤ Multi-pose audio visual speech recognition system for 3 languages. ➤ Speakers- native (20F & 28M) and non-native (Iraq and Yemen) (10M). ➤ Presence of glasses and caps. ➤ Induced mode of data acquisition contains speakers with lipstick.

†Kannada and Telugu Digit Database 2017 [48]	2 speakers for each language	<ul style="list-style-type: none"> • Kannada, Telugu and Indian English languages. • Digits–2 times. 	<ul style="list-style-type: none"> ▪ 640x480, 25 fps. ▪ Frontal view. ▪ Lip region. 	No Audio	➤ Visual Speech Recognition.
MODALITY 2017 [49]	35 (9, 26)	<ul style="list-style-type: none"> • English language. • 168 commands. 	<ul style="list-style-type: none"> ▪ 1080x1920, 100 fps. ▪ Cameras used. ▪ Full face. ▪ Partial front views. 	<ul style="list-style-type: none"> * 44.1kHz. * Array of 8 microphones used. 	<ul style="list-style-type: none"> ➤ Audio-visual speech recognition. ➤ 31 hours of recording. ➤ Speakers- native and non-native English speakers. ➤ Noise varying recording setup. ➤ http://www.modality-corpus.org
AVID 2017 [50]	10 (5,5)	<ul style="list-style-type: none"> • Indonesian language. • 1040 sentences. 	<ul style="list-style-type: none"> ▪ 1280x962, 48 fps. ▪ Full face. ▪ Frontal view. 	44.1 kHz	<ul style="list-style-type: none"> ➤ Audio-visual speech recognition system. ➤ First database in Indonesian language.
NTCD-TIMIT 2017 [51]	56	<ul style="list-style-type: none"> • Irish accent. • 5488 different TIMIT sentences. 	<ul style="list-style-type: none"> ▪ 1920x1080, 30fps. ▪ Frontal view. 	16 kHz.	<ul style="list-style-type: none"> ➤ Noisy version of TCD-TIMIT database. ➤ Six noises. ➤ https://doi.org/10.5281/zenodo.260228

LRS 2017 [52] Lip Reading Sentences	—	<ul style="list-style-type: none"> • English language. • 1,00,000 natural sentences from BBC television 	<ul style="list-style-type: none"> ▪ 120x120, 25 fps ▪ Frontal and Non-frontal views. 	—	<ul style="list-style-type: none"> ➤ Visual speech recognition. ➤ Phrase and sentence recognition. ➤ Talking face synthesis. ➤ https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html
Audio-Visual Lombard Speech 2018 [53]	54	<ul style="list-style-type: none"> • 2700 Lombard and 2700 plain reference utterances. • Extension of GRID corpus. 	<ul style="list-style-type: none"> ▪ 720x480, 24 fps (from frontal web cam). ▪ 864x480, 30 fps (from side webcam). 	48 kHz	<ul style="list-style-type: none"> ➤ Analysis of Lombard effect. ➤ http://spandh.dcs.shef.ac.uk/avlombard/ ➤ Lombard speech is usually more intelligible than plain speech.
AVSpeech 2018 [54]	Wide range	<ul style="list-style-type: none"> • Various languages • Public instructional YouTube videos. • 4700 hours of video segments. 	<ul style="list-style-type: none"> ▪ 25 fps. ▪ Full face. ▪ Head pan up to 50°. ▪ Head tilt up to 30°. 	16 kHz	<ul style="list-style-type: none"> ➤ Speech separation, Video captioning and Speech recognition. ➤ Multi-speaker with background noise. ➤ 3 to 10 s video segments. ➤ Age up to 100. ➤ https://looking-to-listen.github.io/avspeech/explore.html

AVSD 2019 [55] Arabic Visual Speech Database	22 (14,8)	<ul style="list-style-type: none"> • Arabic language. • 10 daily communication sentences. 	<ul style="list-style-type: none"> ▪ 1980x1080, 30 fps. ▪ Smartphone camera. ▪ Frontal view. 	No Audio	<ul style="list-style-type: none"> ➤ Visual Speech Recognition. ➤ Indoor illumination and simple background.
3D Audio Visual Speech Corpus 2020 [56]	5 (3,2)	<ul style="list-style-type: none"> • American English. • 224 sentences from CRM corpus -2 repetition. • 50 sentences from IEEE corpus. 	<ul style="list-style-type: none"> ▪ 3D 180x180x2 stereoscopic mode (half sphere). ▪ 360x180 stereoscopic mode (full sphere). ▪ 5.7 k video resolution, 30 fps. 	* 48 kHz. * 2 microphones used	<ul style="list-style-type: none"> ➤ Virtual reality applications. ➤ 360⁰ degree and 180⁰ stereoscopic recording. ➤ Black and green background. ➤ https://fb.sharepoint.com/:f/s/FRLAudioResearch/EpU2AeUdvDBBvF589aYDOEEBeREku01AkIQWID8V4H3m8g?e=Dq4FWR.
RGB-D 2020 [57]	20	<ul style="list-style-type: none"> • 20 English phrases. 	<ul style="list-style-type: none"> ▪ 640x480, 30 fps. ▪ 3D Head Pose Angles- Yaw, pitch and roll. 	No Audio	<ul style="list-style-type: none"> ➤ Visual speech recognition. ➤ Employed kinetic facial tracking. ➤ Facial points, facial outline, RGB data, depth data, mapping between RGB and depth data, and 2D and 3D face representations of the face along with the 3D head orientation

RUSAVIC 2021 [58]	20	<ul style="list-style-type: none">• Russian language.• 50 phrases related to driving condition.• 10 recording sessions.	<ul style="list-style-type: none">▪ 1920x1080, 60 fps.▪ 2 smartphones.	48 kHz.	<ul style="list-style-type: none">➤ Audio-visual speech recognition in vehicle.➤ First smartphone placed near the left side of driver.➤ Second smartphone placed 20° to the right.➤ https://mobiledrivesafely.com/corpus-rusavic
----------------------	----	---	---	---------	--

† Audio-Visual speech database in Indian languages.

A diversity of standard databases has been reported, and most of them claim to be beneficial for the specific task. The most favourable speech material for the speech recognition task is continuous speech compared to isolated speech for the speaker verification task. The speaker recognition task requires a large size and high variability of speaker population when compared to the speech recognition task. Multi-view visual speech performs better in lip reading tasks. So, the database invented should serve more than one goal to be useful for various research works. The resolution and frame rate of the camera is chosen to employ a trade-off involving computational complexity and high-quality visual speech information. The database should be captured with uniform distribution to avoid gender imbalance. The main criteria for speech database construction are to have a large phonetically balanced speech corpus uttered by many unique speakers in an uncontrolled environment. It is mandatory to figure out the peculiarities of the language of the database and its linguistic background and compare it with other groups of languages, which help resolve the issues that arose during the creation of the database in under-resourced languages. The reported audio-visual speech databases in Indian languages are still not in the full-fledged form in different aspects. Thus, creating an audio-visual speech database in the Malayalam language that addresses most of these requirements has an immense influence on the research community.

2.3 Review on Phoneme and Allophone Durational Modelling and Audio-Visual Speech Asynchrony

Durational analysis of phonemes and allophones are analysed for speech recognition and speech synthesis task in various languages. The variance in phoneme duration is mostly due to its relative position and the characteristics of neighbouring phonemes. Rule-based approaches, statistical approaches, model-based approaches, and their combinations are used to characterise this

variability. Kamrunnahar Swarna et al. used the rule-based approach for the duration modelling of diphones in the Bangla text to speech system [59]. Subachan. Bo Chen et al. investigated the duration of Chinese phonemes using an LSTM RNN by incorporating the linguistic information using cross-entropy [60]. Mohamed O.M. Khalifa et al. utilised HsMM (Hidden semi-Markov Model) for durational analysis of Arabic phoneme and recognition [61]. Sovilj-Nikic et al. Serbian phone duration model is Tree-based machine learning approach is proposed [62]. Magdalena Igras et al. uses statistical analysis to model the phoneme durations in Polish [63]. Yonas Demeke et al. built duration model for Amharic phonemes using Classification and Regression Trees (CART) [64]. Kalu U. Ogbureke et al. proposed a joint venture of HMM and Multilayer Perceptron to model American English phonemes for better improvement of the perceived quality of synthesised speech [65]. Jan Romportl et al. modelled the phoneme duration for the Czech text-to-speech system, ARTIC, using the CART method [66]. Alexandros Lazaridis et al. CART machine learning approach is utilised for durational modelling of Greek phonemes [67]. Giedrius Norkevičius et al. Lithuanian phoneme duration analysis was carried out by using a decision tree-based algorithm [68]. Janne Pylkkönen et al. duration modelling using HsMM is employed to Finnish phone for continuous speech recognition [69]. Ömer Şayli et al. investigated the duration of Turkish phonemes using the statistical method for text-to-speech synthesis [70].

Durational analyses of phonemes in various Indian languages were also reported. Somnath Roy et al. duration analysis of Hindi phonemes is presented for implementing prosody in text-to-speech systems [71]. D. Govind et al. demonstrated the significance of duration in the context of phonological aspects of Assamese [72]. B. Lakshmi Kanth et al. used statistical analysis to characterise the duration of Hindi, Tamil and Telugu [73]. They discriminated against these languages using vowel, nasal and stop sounds. Deepa P. Gopinath

et al. carried out the preliminary durational analysis of Malayalam phonemes for a text-to-speech system. They also analysed the positional variation of consonants in a word [74]. K. Sreenivasa Rao et al. used the SVM regression model for modelling the durational analysis of Hindi, Telugu and Tamil syllables [75]. N. Sridhar Krishna et al. proposed a preliminary work using the CART method for modelling the duration of Hindi and Telugu phonemes [76]. Samudravijaya K et al. studied the duration analysis of Hindi stop consonants [77]. S. R. Savithri identified factors affecting the duration of Tamil vowels in the initial position of words [78].

In addition to phoneme durational analysis, the acoustical properties of allophones were explored in various languages. Piotr Koziński et al. allophone based speech recognition system for the Polish language is presented [79]. Instead of using the entire allophone set, most frequent allophones were used for this purpose. Fayçal Imedjdouben et al. used a rule-based method to convert phonemes to allophones for synthesising the Arabic speech [80]. Ji Xu et al. automatic allophone deriving approach is utilised to recognise Korean speech [81]. Wafa Barkhoda et al. proposed three Kurdish speech synthesis systems based on diphone, syllable and allophone and compared their performance using various tests [82]. Diphone based speech synthesiser showed the most natural one while all systems intelligibilities are acceptable. Long Nguyen et al. used the most frequent allophones to recognise the Japanese speech [83]. Pavel A. Skrelin described an allophone based Russian speech synthesiser [84]. Vivek P et al. presented the rule set for Malayalam vowel allophones based on the position and neighbourhood information and analysed its duration properties [85].

Audio-Visual speech asynchrony is one of the major issues to be addressed in audio-visual speech processing. Aciel Eshkya et al. used ultrasound to visualise the tongue during speech production. A self-supervised

neural network is employed to address asynchrony between ultrasound and audio speech signals. Gerasimos Potamianos et al. proposed a two-stream ConvNet architecture to determine the lip-synchronisation error [4]. Etienne Marcheret et al. Two DNN-based audio-visual synchrony detectors, one based on scattering and the other on DCT features, were used with identical architectures [86]. Kshitiz Kumar et. al. proposed a time-evolution model for audio-visual features in linear prediction and proposed the using canonical correlation analysis to multidimensional audio-visual features [87]. Enrique Argones Ru'a et al. The degree of synchronisation between the lips and the voice was measured using coupled hidden Markov models and co-inertia analyses [88].

The durational analysis of phonemes, allophone and audio-visual speech asynchrony is performed in various languages by employing various methods. However, the corresponding analysis in Indian languages, especially in Malayalam, is less addressed. As Malayalam is a language with a rich phoneme and allophone set, durational analysis of these basic units and audio-visual speech asynchrony can better understand language and performance improvement in speech-based applications.

2.4 Review on known Viseme set

Many researchers have analysed the importance of the phoneme to viseme mapping. The phonemes which have almost the same visual mouth appearance grouped to a single viseme class. In literature, many mappings reported, range of visemes in a language varies between 10 and 20. The number and nature of viseme are language dependent. Hence a language-specific exploration is needed for establishing the viseme set for a language. Traditionally there are three approaches for obtaining visemes from a many to one mapping: linguistic knowledge-based, perception experiments with human subjects and data-driven approach. Some authors blend linguistic and

perception experiments based on approaches and name them as subjective assessments. Using a subjective approach, viseme classes are defined through linguistic knowledge and prediction of phonemes having a similar visual appearance. Viseme classes created by clustering of phonemes based on features extracted from the mouth region is the highlight of a data-driven approach. Most of the work has reported in European languages, but in India, only a few such as Hindi and Marathi have been studied in visual speech. The centre of discussion so far shows attempts have not yet reached the realm of successful development of visual speech technology in the Indian language. Besides, only a few viseme maps have studied for the coarticulation effects of visual speech, which might be the lack of database containing all contextual variations in the language concerned. This research work is going to be the initial study in Malayalam phoneme to viseme mapping based on linguistic involved data-driven approach and allophone-to-viseme mapping based on a data-driven approach alone. Table 2.2 summarizes a brief background of established viseme mapping in terms of language, methodology and some special features.

Table 2.2 Summary of known Viseme sets

Author	Year	Linguistic Information	Implementing Method	Special Features
Woodward et al.	1960 [89]	<ul style="list-style-type: none"> • 24 initial consonants of English to 4 viseme classes. 	<ul style="list-style-type: none"> * Perception experiment with human subjects. 	<ul style="list-style-type: none"> ➤ Studied on a black and white video of a speaker face.
Fisher et al.	1968 [90]	<ul style="list-style-type: none"> • 23 initial consonants to 5 viseme classes. • 20 final consonants to 5 viseme classes. 	<ul style="list-style-type: none"> * Multiple-choice intelligibility test. * Confusion matrix. 	<ul style="list-style-type: none"> ➤ Studied visual perception of initial and final consonants. ➤ Mapping is done subjectively. ➤ Fisher introduced the term viseme which is a compound word of visual and phoneme.
Franks et al.	1972 [91]	<ul style="list-style-type: none"> • Initial consonants. 		<ul style="list-style-type: none"> ➤ Consonants are only studied.
Binnie et al.	1976 [92]	<ul style="list-style-type: none"> • 20 English consonants to 9 viseme classes. 	<ul style="list-style-type: none"> * Human testing. * Confusion matrix. 	<ul style="list-style-type: none"> ➤ Consonants are only studied. ➤ 20 English consonants were combined with the vowel /a/ to form 20 CV syllables. ➤ 34 female observers have participated in the testing process. ➤ Mapping is done subjectively.
Jeffers et al.	1980 [91]	<ul style="list-style-type: none"> • American English. 	<ul style="list-style-type: none"> * Pure linguistic approach. 	

	<ul style="list-style-type: none"> • 43 phonemes to 11 viseme classes. 		
Montgomery et al. 1983 [93]	<ul style="list-style-type: none"> • American English. • 15 vowels and diphthongs. 	<ul style="list-style-type: none"> * 3 methods- Perceptual analysis using confusion matrix, Physical measurements (height, width, area, acoustical and visual duration) and Correlation between the two. 	<ul style="list-style-type: none"> ➤ Study was restricted to vowels only.
Goldschen et al. 1994 [94]	<ul style="list-style-type: none"> • 19 vowel and diphthong phonemes to 16 viseme classes. • 36 consonant phonemes to 18 viseme classes. 	<ul style="list-style-type: none"> * HMMs using Forward-Backward algorithm. 	<ul style="list-style-type: none"> ➤ 67 sentences from TIMIT database. ➤ Used optical information from the oral cavity shadow. ➤ Continuous Optical Automatic speech recognition. ➤ Facial Animation.
Lander et al. 1999 [95]	<ul style="list-style-type: none"> • 35 phonemes to 12 classic Disney mouth position. 	<ul style="list-style-type: none"> * Linguistic approach 	
Neti et al. 2000 [96]	<ul style="list-style-type: none"> • 42 phonemes to 12 viseme classes (excluding silence). 	<ul style="list-style-type: none"> * Mixture of linguistic and data driven approach. * Decision tree based HMM state clustering method. * Models are trained using DCT visual features. 	<ul style="list-style-type: none"> ➤ IBM ViaVoive database was used.
Pandzic et al. 2002	<ul style="list-style-type: none"> • 24 phonemes to 14 viseme classes. 	<ul style="list-style-type: none"> * Based on Face Animation Parameters. 	<ul style="list-style-type: none"> ➤ MPEG-4 facial animation.

Lee et al. 2002 [97]	<ul style="list-style-type: none"> • 41 phonemes to 14 viseme classes. • 7 vowel, 6 consonant and 1 silence viseme. 	<ul style="list-style-type: none"> * Assumed to be linguistic approach. * HMM modelling. Used context-independent recognition units (phone model). * Produced a sequence of viseme symbols from speech waveform. 	<ul style="list-style-type: none"> ➤ TIMIT speech database. ➤ Two approaches in building viseme recognizer-viseme HMMs and phoneme HMMs.
Hazen et al. 2004 [21]	<ul style="list-style-type: none"> • American English. 50 phonemes to 14 viseme classes. • There are 54 phonemes, but 4 phonemes were merged to get 50 phonemes. 	<ul style="list-style-type: none"> * Data-driven approach. * Agglomerative hierarchical clustering algorithm. * Bottom-up clustering using maximum Bhattacharyya distances. * 96-dimension stacked PCA feature vectors. 	<ul style="list-style-type: none"> ➤ AV-TIMIT speech database. ➤ Viseme was represented by three consecutive frames with middle frame describe the static viseme of each phoneme. ➤ Before clustering some phonemes were merged
Aschenberger et al. 2005 [98]	<ul style="list-style-type: none"> • German language. • 42 phonemes to 15 viseme classes. 	<ul style="list-style-type: none"> * Linguistic approach. 	<ul style="list-style-type: none"> ➤ Applied for speech synthesis.
Melenchón et al. 2007 [99]	<ul style="list-style-type: none"> • Spanish language. • 12 allophones to 6 viseme classes. 	<ul style="list-style-type: none"> * Data-driven approach. * 12 PCA coefficients were used as feature vector. 	<ul style="list-style-type: none"> ➤ Three speakers utter 12 Spanish sentences.
Chitu et al. 2009 [100]	<ul style="list-style-type: none"> • Dutch language. • 40 phonemes to 18 viseme classes. 	<ul style="list-style-type: none"> * Confusion matrix. 	

Damien et al. 2009 [101]	<ul style="list-style-type: none"> ● Arabic language. ● 28 phonemes to 10 viseme classes. 	<ul style="list-style-type: none"> * Data-driven approach. * Geometrical features used. 	<ul style="list-style-type: none"> ➤ Four speakers utter four types of word sequences.
Yu et al. 2010 [102]	<ul style="list-style-type: none"> ● 50 words to 60 classes of visual speech units (VSU). 	<ul style="list-style-type: none"> * Data-driven approach. * Used Expectation Maximization Principal Component Analysis (EM-PCA) as feature extraction method. * Based on HMM classification. 	<ul style="list-style-type: none"> ➤ Introduced new term “Visual Speech Unit (VSU)” which include transition information between consecutive visemes. ➤ Two speakers utter a total of 50 words.
Chelali et al. 2012 [103]	<ul style="list-style-type: none"> ● Arabic language. ● 28 consonant phonemes to 11 viseme classes. 	<ul style="list-style-type: none"> * Data-driven approach. * Statistical parameters of lips and geometric features. 	<ul style="list-style-type: none"> ➤ Viseme Recognition. ➤ Ten native Algerian speakers. ➤ Video- 576*720, 25 fps. ➤ Front region.
Mattheyses et al. 2013 [104]	<ul style="list-style-type: none"> ● Dutch language. ● Many-to-many phoneme-to-viseme mapping. 	<ul style="list-style-type: none"> * Data driven approach. * AAM (Active Appearance Model)-based representation of mouth region. * Tree- based and k-means clustering approach was used. 	<ul style="list-style-type: none"> ➤ Coarticulation effect was studied. ➤ Applied for visual speech synthesis.
Seko et al. 2013 [105]	<ul style="list-style-type: none"> ● Japanese language. ● 40 phonemes to 14 viseme classes (excluding silence). 	<ul style="list-style-type: none"> * HMM modelling. 	<ul style="list-style-type: none"> ➤ CENSREC-1-AV database was used.

Aghaahmadi et al. 2013 [106]	<ul style="list-style-type: none"> • Persian language. • 23 consonants to 7 viseme classes. 	<ul style="list-style-type: none"> * Data-driven approach-PCA. * Subjective assessment. 	<ul style="list-style-type: none"> ➤ Studied the effect of coarticulation and the phoneme position in syllable. ➤ Coarticulation effect-middle image in bi-viseme (CV syllable) and tri-viseme (CVC syllable). ➤ AVA II database was used. ➤ Viseme Recognition. ➤ In Hindi, total 61 phonemes.
†Varshney et al. 2014 [107]	<ul style="list-style-type: none"> • Hindi language. • 20 consonant phonemes to 5 viseme classes. 	<ul style="list-style-type: none"> * Linguistic approach. 	<ul style="list-style-type: none"> ➤ AVA II database was used. ➤ Viseme Recognition. ➤ In Hindi, total 61 phonemes.
Bear et al. 2015 [108]	<ul style="list-style-type: none"> • 46 phonemes to visemes ranging from 2 to 45. 	<ul style="list-style-type: none"> * Viseme classes were obtained based upon the mapping of articulated phonemes, which was confused during phoneme recognition, into viseme groups. 	<ul style="list-style-type: none"> ➤ Designed Speaker-dependent viseme classes. ➤ Studied on LiLIR dataset. ➤ 12 British speakers utter about 1000 words totally.
Taylor et al. 2015 [109]	<ul style="list-style-type: none"> • Created many-to-many mapping. • Approximately 50000 visual speech gestures – 150 	<ul style="list-style-type: none"> * Clustered the speech gestures identified by AAM (Active Appearance Model) of jaw and lips. 20 Dimension feature vector entirely describe the shape and appearance 	<ul style="list-style-type: none"> ➤ KB-2K database was used. ➤ A single actor recites 2542 phonetically balanced sentences from TIMIT database.

Setyati et al. 2015 [110]	<p>dynamic viseme classes.</p> <ul style="list-style-type: none">• Indonesian language.• 49 phonemes to 12 viseme classes.	<p>information. Dynamic visemes were learned entirely from visual data.</p> <p>* Linguistic approach.</p>	<ul style="list-style-type: none">➤ Applied for automatic redubbing of video.➤ Used Blend shape models for analysing the facial images.➤ 10 speakers were used for this study.
†Brahme et al. 2016 [111]	<ul style="list-style-type: none">• Marathi language.• 44 phonemes to 13 viseme classes.• Including silence.• 10 vowel phonemes to 5 viseme classes.• 34 consonant phonemes to 7 viseme classes	<p>* Linguistic approach.</p>	<ul style="list-style-type: none">➤ Visual Speech Recognition.➤ First work in Marathi language.

† Viseme mapping in Indian languages.

2.5 Review on Lip Segmentation

Being lips is the active articulator, segmenting the lip region is the primary task in visual speech analysis. One of the successes behind the AVSR task is to extract the lip region from its surroundings efficiently. Lip segmentation and tracking can be broadly classified into colour-based and model-based approaches. The colour-based approach uses colour uniformity or discontinuity, whereas the model-based approach uses a set of parameters to represent the lip shape mathematically.

Ashley D. Gritzman et al. presented a detailed analysis of 33 colour transforms, including 21 channels from seven different colour spaces and 12 more [112]. They further expanded the research to include oral cavity segmentation and created a rating of colour channels based on their potential for segmentation. Using optimal thresholding based on bacterial foraging optimisation, Mohamad Amin Bakhshali et al. segmented the lip area [113]. A new colour space (IHLS) was also proposed, which is less computationally demanding and error-prone. Ashley D. Gritzman et al. offer a method termed adaptive threshold optimisation (ATO), which uses shape information feedback to determine the threshold [114]. For lip segmentation, Shu-Hung Leung et al. use fuzzy clustering in the CIELab and CIELuv colour space combined with distance information [115]. Simon Lucey et al. employed chromatic temporal information and a fuzzy-based thresholding algorithm to solve difficulties with the traditional thresholding method [116]. Meng Li et al. adopted a three-stage technique in which the CIE Lab and LUX colour spaces were used first, followed by a Gaussian model using hue and saturation values. The lip region is extracted using a morphological 38 filter in the final stage [117].

Nicolas Eveno et al. proposed a jumping snake that makes use of various distinctive points in the lip region, as well as a cubic-natured parametric model in the segmentation process [118]. M.Li'evin et al. used spatiotemporal

Bayesian segmentation with hue and motion information to locate and segment the mouth accurately [119]. Lastly, an active contour is used to precisely outline the lips. Tony F. Chan et al. presented active contours without edges that detect the inner contour automatically regardless of where the initial curve is on the lip [120]. Pierre Garcon et al. used a statistical model of shape with local appearance gaussian descriptors [121]. In a multi-speaker task, this strategy can be generalised to account for intra-person appearance variability. The Active Shape Model (ASM) was utilised by Juergen Luetin et al. to extract visual speech [122]. K.L. Sum et al. 14-point ASM is used to extract the outer lip shape information, and the cost function is computed using fuzzy clustering analysis [123]. The Active Appearance Model (AAM) was utilised by Piotr Dalka et al. to statistically describe lip shape and texture [124]. In the face recognition challenge, M Iqtait et al. applied ASM and AAM to extract visual speech features [125].

Researchers have given lip tracking top priority since the accuracy of recognition relies heavily on the proper tracking of spoken lips. Many literary works, especially in European contexts, have adopted the red colour dominance of lips. However, in the Indian context, there is no significant variance in lip and skin colour tone, and the presence of facial hairs further complicates the capture and analysis of lip dynamics. In Indian languages, there are very few works reported. The lip-skin discrimination power varies depending on the colour space and the ethnicity of the database used. Recent lip-tracking experiments have taken a hybrid approach, which could lead to better segmentation findings.

2.6 Review on AVSR systems in Intense background Noise

Speech processing in the Malayalam language is still in its infancy stage, with only a few works focusing only on audio speech. Lavanya B.Babu et al. proposed a continuous Malayalam speech recognition system in the Kaldi

platform [126]. They extracted MFCC features and its transformed version using LDA and built triphone and mono phone-HMM models and compared its performance. Smrithy K Mukundan et al. introduced a speaker-independent word recognition system using MFCC features and a hidden Markov toolkit [127]. Raji Sukumar A1 et al. presented an intelligent query processing system in Malayalam [128]. Cini Kurian et al. presented a speaker-independent connected Malayalam digit recogniser using PLP and HMM and obtained 99.5% accuracy [129]. Anu V Anand et al. developed a large vocabulary continuous speech recognition system for visually impaired people using MFCC features and HMM [130]. Vimal Krishnan V R et al. proposed an isolated word recognition task using wavelet transforms and ANN, which obtained an 89% recognition rate [131].

Few attempts have been made to build an AVSR system in Indian languages like Hindi [46], [132], [133], Oriya [42], etc. Since the first AVSR system in 1984, researchers have focused on improving the audio and visual speech feature extraction approach [134]–[136] and employing various schemes for modelling and fusing techniques [45], [58], [137], [138]. Table 2.3 summarises the performance of various ASVR systems employed in the intense background (negative SNR) conditions in terms of features, fusion and stream weight.

Table 2.3 Summary of AVSR systems employed in negative SNR Conditions.

Author Year	Speech Features	Special Features (Purpose, Database, Fusion & Weight Strategies, Model)	Performance			
Wentao Yu et al. 2021 [137]	<ul style="list-style-type: none"> • Audio-40 log Mel Features + 2 Pitch features + Probability of Voicing. • Visual-43 IDCT coefficients. 	<ul style="list-style-type: none"> * AVSR in Vehicle cabin. * LRS2 database. * Decision Fusion. * Model and Signal-based Reliability measures. * BLSTM and LSTM neural networks. 	<ul style="list-style-type: none"> ➤ Additive Noise ➤ Word error rate (%). 			
				Clean	10.32	7.84
				9 dB	15.84	10.78
				6 dB	13.97	10.25
				3 dB	19.17	14.93
				0 dB	21.25	16.35
				-3 dB	21.26	17.89
				-6 dB	27.22	23.11
				-9 dB	33.30	27.55
Ali S.Saudia et al. 2019 [139]	<ul style="list-style-type: none"> • Audio-Gabor Audio features. • Visual-Gabor video features 	<ul style="list-style-type: none"> * Continuous digit recognition. * CUAVE database. * Late integration. * Signal based-Zero crossing rate, short-time energy, energy entropy, spectral roll-off, spectral centroid and spectral flux. * Synchronous Multi-stream HMM. 	<ul style="list-style-type: none"> ➤ NOISEX database-white gaussian noise. ➤ Visual-only 69.23%. 			
				Clean	98.89	98.89
				20 dB	97.98	97.78
				15 dB	96.67	96.67
				10 dB	94.17	95.83

				5 dB	92.22	94.44
				0 dB	91.11	93.33
				-5 dB	81.11	85.71
				-10 dB	75.56	80
Ian McLoughlin et al. 2019 [140]	<ul style="list-style-type: none"> • Audio-Spectrogram, Cochleogram and constant-Q transform-based images. 	<ul style="list-style-type: none"> * Audio-only event classification problem. * Real World Computing Partnership (RWCP) Sound Scene Database. * CNN Classifiers. 	<ul style="list-style-type: none"> ➤ NOISEX-92 database-4 noise. ➤ Combination of Cochleogram and Spectrogram performs well. 			
					Clean	99.33%
					20 dB	99.50%
					10 dB	99.24%
					0 dB	96.96%
					-5 dB	73.02%
					-10 dB	34.44%
Hendrik Meutzner et al. 2017 [135]	<ul style="list-style-type: none"> • Audio-23 Mel filter bank features, 13 MFCC, 32 ratemap features and 13 GFCC. • Visual-63 DCT coefficients. 	<ul style="list-style-type: none"> * AVSR. * CHiME challenge dataset. * Early and State-based integration. * Stream weight-Improved Minima Controlled Recursive Averaging. * DNN. 	<ul style="list-style-type: none"> ➤ Highly non-stationary background noise. ➤ Ratemap feature performs better. ➤ Keyword accuracy (%). ➤ Video-only 71.34% 			
					Audio-only AVSR	
				9 dB	94.47	96.82
				6 dB	92.86	95.02
				3 dB	89.71	95.02
				0 dB	86.48	93.47
				-3 dB	80.36	90.21
				-6 dB	73.98	88.32

<p>Ahmed Hussen Abdelaziz et al. 2017 [51]</p>	<ul style="list-style-type: none"> • Audio-13 MFCC + first and second derivatives. • Visual-DCT + PCA. 	<ul style="list-style-type: none"> * NTCD-TIMIT database. * AVSR. * Direct and Separate integration. * Speaker independent acoustic DNN-HMM hybrid model. 	<ul style="list-style-type: none"> ➤ 6 Noises from NOISEX-92 and CHiME challenge dataset. ➤ Phone error rate (%). ➤ Visual-only 68%. 																								
<table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th></th> <th>Audio-only</th> <th>AVSR</th> </tr> </thead> <tbody> <tr><td>Clean</td><td>21.6</td><td>22.5</td></tr> <tr><td>20 dB</td><td>45.4</td><td>45.8</td></tr> <tr><td>15 dB</td><td>56</td><td>55.6</td></tr> <tr><td>10 dB</td><td>67</td><td>62.1</td></tr> <tr><td>5 dB</td><td>76.8</td><td>64.3</td></tr> <tr><td>0 dB</td><td>85.4</td><td>65.7</td></tr> <tr><td>-5 dB</td><td>90.6</td><td>65.6</td></tr> </tbody> </table>					Audio-only	AVSR	Clean	21.6	22.5	20 dB	45.4	45.8	15 dB	56	55.6	10 dB	67	62.1	5 dB	76.8	64.3	0 dB	85.4	65.7	-5 dB	90.6	65.6
	Audio-only	AVSR																									
Clean	21.6	22.5																									
20 dB	45.4	45.8																									
15 dB	56	55.6																									
10 dB	67	62.1																									
5 dB	76.8	64.3																									
0 dB	85.4	65.7																									
-5 dB	90.6	65.6																									
<p>Naomi Harte et al. 2015 [45]</p>	<ul style="list-style-type: none"> • Audio-12 MFCC. • Visual-132 DCT coefficients + first derivative + second derivative. 	<ul style="list-style-type: none"> * Continuous AVSR. * TCD-TIMIT database. * Early Integration. * HMM. 	<ul style="list-style-type: none"> ➤ Additive White Gaussian noise. ➤ Visual-only 34.54%. 																								
<table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th></th> <th>Audio-only</th> <th>AVSR</th> </tr> </thead> <tbody> <tr><td>40 dB</td><td>50.88</td><td>37.24</td></tr> <tr><td>30 dB</td><td>44.63</td><td>35.95</td></tr> <tr><td>20 dB</td><td>32.20</td><td>32.41</td></tr> <tr><td>10 dB</td><td>21.60</td><td>27.55</td></tr> <tr><td>0 dB</td><td>15.49</td><td>22.3</td></tr> <tr><td>-10 dB</td><td>8.73</td><td>18.33</td></tr> </tbody> </table>					Audio-only	AVSR	40 dB	50.88	37.24	30 dB	44.63	35.95	20 dB	32.20	32.41	10 dB	21.60	27.55	0 dB	15.49	22.3	-10 dB	8.73	18.33			
	Audio-only	AVSR																									
40 dB	50.88	37.24																									
30 dB	44.63	35.95																									
20 dB	32.20	32.41																									
10 dB	21.60	27.55																									
0 dB	15.49	22.3																									
-10 dB	8.73	18.33																									
<p>Kwanchiva Thangthai et al.</p>	<ul style="list-style-type: none"> • Audio-39 MFCC with first and second derivatives. 	<ul style="list-style-type: none"> * AVSR in noisy environment. * Resource management-3000 database. 	<ul style="list-style-type: none"> ➤ Word accuracy(%). ➤ Visual only 84.67%. 																								

Mihai Gurban et al. 2009 [136] • Audio-13 MFCC + first and second derivative. * AVSR CUAVE database.
 • Visual-192 DCT coefficients. * Model based- Entropy.

- Additive white Gaussian noise.
 - 25 dB 96%
 - 20 dB 96%
 - 15 dB 95%
 - 10 dB 93%
 - 5 dB 90%
 - 0 dB 85%
 - 5 dB 78%
 - 10 dB 67%

E.K. Patterson et al. 2002 [16] • Audio-16 Mel frequency discrete wavelet coefficients. * Isolated word recognition. CUAVE database.
 • Visual-Geometric features and Fourier descriptors. * HMM.

- NOISEX database.
- Visual-only 87%.

	Audio-only	AVSR
Clean	100	100
18 dB	96.5	99.3
12 dB	80.9	96.9
6 dB	52.6	92.9
0 dB	29.1	89.8
-6 dB	21.0	88.4

2.7 Conclusions

The research on recent developments in the AVSR system is examined from several aspects. A thorough examination of the audio-visual speech database is conducted. A study on phoneme and allophone durational analysis and audio-visual speech asynchrony is also provided. The identification of viseme sets in various languages are discussed in depth. A review of lip segmentation methods is conducted. Finally, reviewed the current developments in audio-visual speech recognition tasks in noisy environments.

CHAPTER 3

AN AUDIO-VISUAL SPEECH DATABASE IN MALAYALAM – “MOZHI”

3.1 Introduction

Back in the last couple of decades, human interact with a machine in every aspect of their life. For better human-computer interaction, the device must perform just like a human interacts with his surroundings. In the speech processing aspect, human interaction with its surroundings is bimodal. The audio signal from the mouth is the primary source for recognizing speech. Still, it is well established that incorporating visual cues from the mouth has improved the efficiency of speech perception levels. Even though the visible part of speech contains less speech information than the acoustic part, it helps better understand when the acoustic element is less informative. The idea of visual speech in speech processing is employed in resourced languages like English. However, under-resourced languages like Malayalam (a regional language of India) never explored the usefulness of visual speech in speech processing systems. The non-availability of an audio-visual speech database for Malayalam is one major reason that keeps researchers away from such studies. Researchers working in languages, which are less addressed by the speech research group, are forced to do their research work with self-collected speech data, which brings problematic conditions when comparing their results with others.

Over the years of development and unceasing research in audio-visual speech processing has endorsed the deficiency of a standard audio-visual speech database. The demand for diversity in resources and large storage capacity is the main challenge that produces the remarkable hindrance in developing the bi-modal speech database. Creating an exhaustive and

methodically arranged audio-visual speech database builds up worldwide acceptance in the research community. A diversity of standard databases has been reported, and most of them claim to be beneficial for the specific task. The database required to develop should have many features to be useful for various research works. The main criteria needed for speech database construction is to have a large phonetically balanced speech corpus uttered by many unique speakers in an uncontrolled environment. To develop an efficient speech-based application system, a long duration of the annotated recording of the speech utterance is necessary for both training and testing schemes. It is mandatory to figure out the peculiarities of the language of the database and its linguistic background and compare it with other language groups that help to resolve the issues that spring up during the creation of the database.

An audio-visual speech database in Malayalam was captured in various environments for a variety of research goals. It is probably the first of its sort in Malayalam. This database was recorded in 3 categories with unique speakers in each stage. The first recording category contains an audio-visual speech database recorded from 25 females and five male speakers speaking 50 Malayalam isolated phonemes and 207 connected words comprising of all allophonic variations in a controlled environment. Each isolated phonemes and word in the audio and video domain are appropriately segmented and labelled. The second recording category creates an audio-only speech database captured in two different conditions, one in a secluded and noiseless environment and another in the acoustically realistic environment. Finally, in the third category, the audio and visual speech signal from 5 female speakers uttering isolated phonemes and real-time isolated words in an acoustically and visually realistic environment is recorded. It took three years to prepare this database. This chapter also discusses the different simulated noisy signals and the audio-visual asynchrony in the Malayalam language.

This chapter's content is organised as follows. Section 3.2 displays a detailed outline of this database architecture, such as different categories of recording, language material, hardware setup, and environmental conditions. Section 3.3 discusses the segmentation and labelling process involved in the creation of this database. Section 3.4 explains different simulated noisy signals used in this work. Section 3.5 discusses the audio-visual asynchrony in the Malayalam language. Section 3.6 concludes the work.

3.2 Database Design

The purpose of a speech database depends intensely on the language material, speaker population, and the quality of the recording, which should be optimally adapted to attain the prime goal. It is better to consider a continuous speech corpus for speech recognition and isolated words for speaker verification. Many speakers are required for the verification task than compared to the recognition task. Extracting minute variations in the visual features of the speaker may improve the recognition task, which can be implemented by using a high definition camera rather than cameras with high fps (frames per second) under better illumination conditions.

A new audio-visual speech database in Malayalam is induced. The database is recorded by native speakers of northern Kerala. The visual speech database is mainly a female-oriented database by keeping in mind the low contrast between the lip and skin colour tone in the Indian context and the presence of facial hairs in the significant male population in Kerala. However, to avoid gender discrimination, five male speakers are included in the visual speech database, which carries the visual complexity like facial hairs partially covering the lip region. This database is recorded in 3 categories with unique speakers in each stage. The first category is the audio-visual speech database recorded in a controlled environment. This category includes 25 female and five male speakers uttering 50 Malayalam isolated phonemes and 207

connected words comprising all allophonic variations five times each. The second category consists of an audio-only speech database which has two sub-categories. The first sub-category is a clean audio speech database recorded in a closed environment uttering 50 Malayalam isolated phonemes five times each by ten male and 20 female speakers aged 21-25. The second sub-category is an audio speech database recorded in an acoustically realistic environment uttering of five Malayalam short vowel phonemes ten times each by ten males and ten females in each group ranging from 5 to 60, a total of 560 males and 560 females. The third category contains an audio-visual speech database captured in an uncontrolled environment to help in developing a visual speech processing system in real-world ambient conditions. This category includes five female speakers uttering 50 Malayalam isolated phonemes and real-time isolated words five times each with a complex background. It includes lighting variations and multiple speakers in the background.

3.2.1 Language Material

India has 23 constitutionally recognised official languages. Hindi and English are treated as official language by the Central Government. Malayalam is a Dravidian language spoken across Kerala, Lakshadweep, and Mahe has been spoken by 38 million people worldwide and designated a status of classical language in 2013. Due to its lineage deriving from both Tamil and Sanskrit, the Malayalam alphabet has many letters among the Indian Language orthographies. Malayalam is a language that is used by masses and less accessed from the computational point of view. So, developing a speech-based application in Malayalam benefits from enjoying the recent technological revolution with the native language rather than in English.

It is mandatory to figure out the peculiarities of the language of the database and its linguistic background and compare it with other groups of languages that help to resolve the issues that arose during the creation of the

database. The language basic units like phoneme, allophone and viseme are the essential terminology used in any speech-oriented application. Phonemes are the relatively distinct and fundamental utterances of a language [142]. The number of phonemes varies between dialects and languages; for instance, British English has 44, Indian English has 38 [91], and Malayalam has 50. The participation of articulators in the speech production system is used to classify the phonemes. In Malayalam, the language components like phoneme and allophone is linguistically categorized according to articulation points and manners as in [<http://www.cmltemu.in/phonetic/#/>], an inclusive Malayalam phonetic archive owned by Thunchath Ezhuthachan Malayalam University (TEMU), Kerala, India. The audio file in this archive is utilized to understand the pronunciation of phonemes and words.

3.2.1.1 Speech Production

When humans speak, the air is forced out of their lungs through the trachea, the larynx, and the vocal tract, with two openings (mouth and nose) and variable constrictions at different places for different sounds. The glottis is the beginning of the vocal tract, around 17 cm long and extends to the lips. The vocal folds, commonly known as vocal cords, are two tiny muscular folds in the larynx that may be opened or closed. Glottis is the space between two folds. Vocal cords are tensed muscular tissues which vibrate when the folds are partially closed and do not vibrate when the folds are apart. The nasal cavity is a chamber between the velum and the nostrils. Sounds produced with vocal cords vibration is called voiced sounds, and sounds produced when folds are apart (no vibration) are called unvoiced sounds. During vibration, airflow is chopped into quasi-periodic pulses, which are then modulated in frequency by passing through the pharynx, mouth cavity, and nasal cavity—depending upon the position and manner of different articulators (lips, teeth, tongue, alveolar, and palates) different sounds are produced. Linguistically, phonemes are

broadly classified into vowel phonemes, diphthong phonemes and consonant phonemes. Fig. 3.1 shows the schematic diagram of the human speech production system.

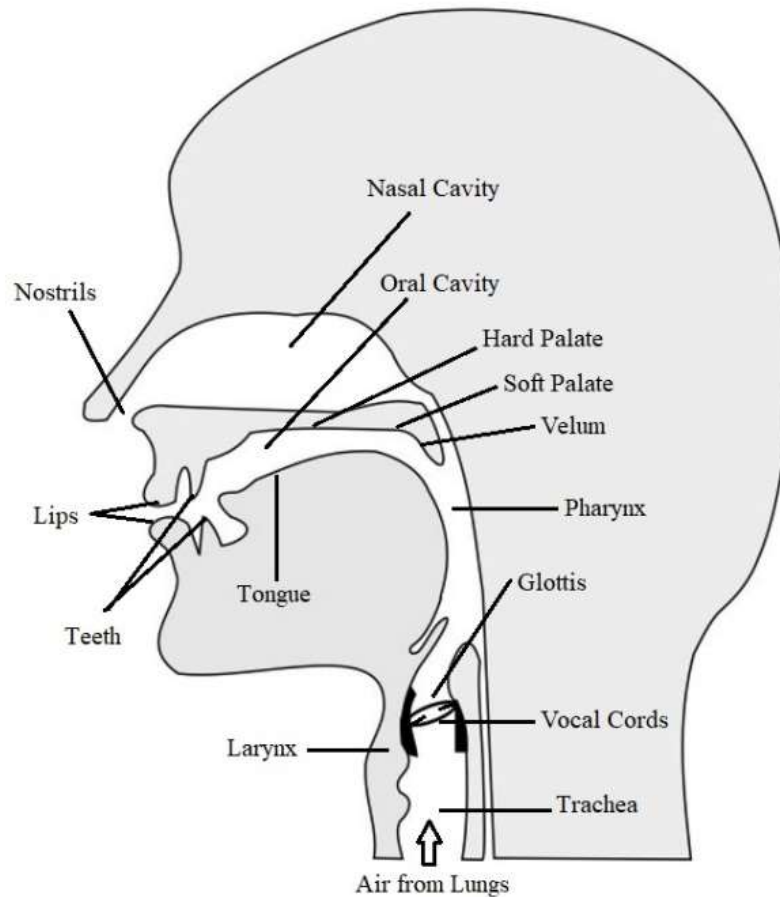


Fig. 3.1 Schematic Diagram of Speech Production System

3.2.1.2 Vowel Phonemes

The vocal tract, driven by quasi-periodic air pulses produced by the vocal cord's vibrations, produces vowel phonemes. Vowels are voiced sounds with less constriction in the vocal tract and generally have a long duration and are louder than other classes of phonemes. The tongue plays a vital role in the creation of vowel sounds (vowel phonemes). Vowels are classified based on

the height (High, Mid, and Low) and position (Front, Central, and Back) of the highest part of the tongue and the shape of the lips (rounded or not). The height (Low, Mid, and High) and location (Back, Central, and Front) of the highest portion of the tongue and the shape of the lips are used to classify vowels. Because the lips are the most visible element of visual speech, their shape is vital in classifying visemes (as discussed in section 4.7.1). Classification of vowel phonemes is listed in table 3.1. Fig. 3.2 displays the short vowel phonemes in the time domain.

Table 3.1 Linguistic Classification of Vowel Phonemes

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

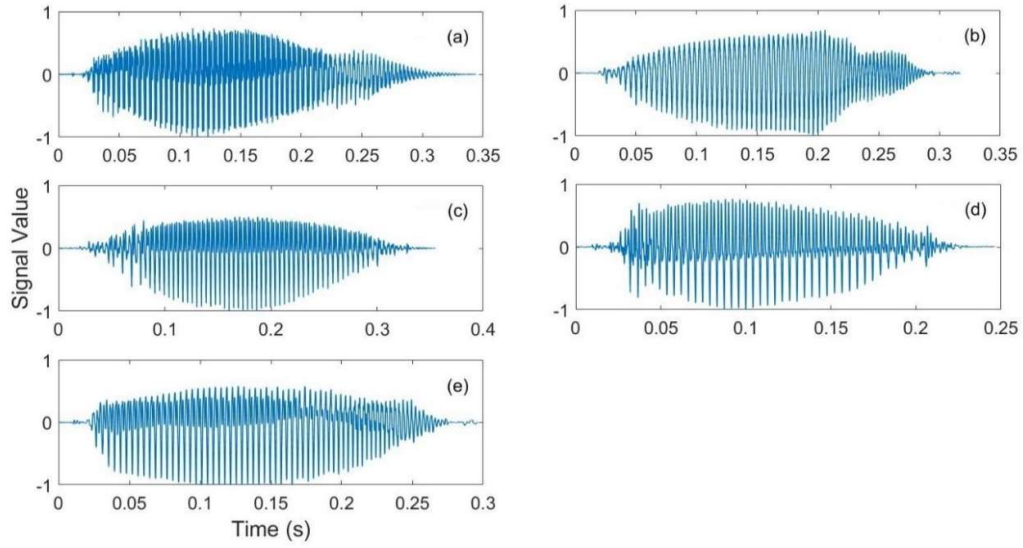


Fig. 3.2 Time Domain Representation of Short Vowel Phonemes: (a) അ /a/, (b) ഇ /i/, (c) എ /e/, (d) ഒ /o/ and (e) ഉ /u/.

3.2.1.3 Diphthong Phonemes

Diphthong phonemes are voice sounds generated by smoothly switching between two vowel configurations of the vocal tract. In Malayalam, there are two diphthongs: ഐ /ai/ and ഔ /au/. Diphthong phoneme ഐ /ai/ is produced by starting with the vowel phoneme അ /a/, and before ending it, the tongue changes its position to the configuration of the vowel phoneme ഇ /i/. Diphthong phoneme ഔ /au/ is produced by uttering the vowel phoneme അ /a/ and then glides to the vowel phoneme ഉ /u/. Fig. 3.3 displays the diphthong phonemes in the time domain.

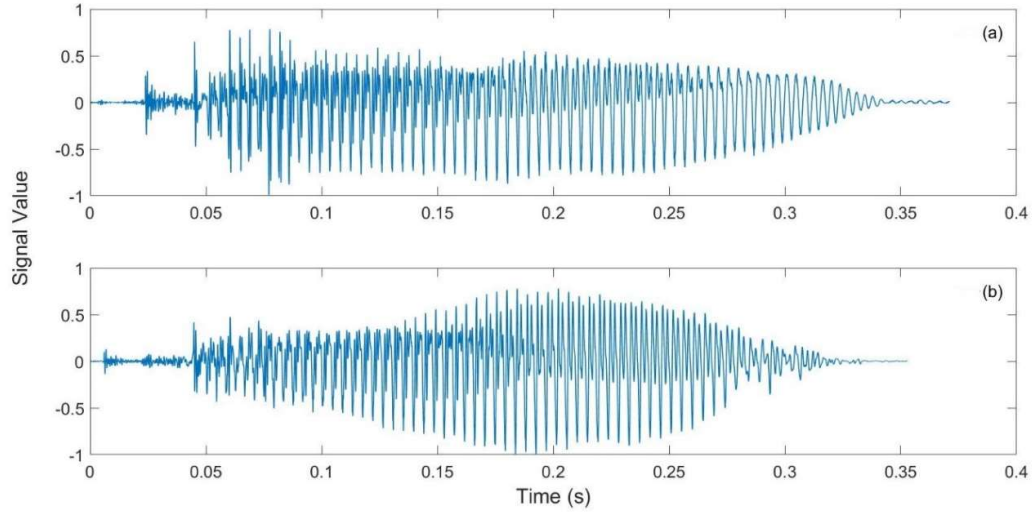


Fig. 3.3 Time Domain Representation of Diphthong Phonemes: (a) ഈ /ai/ and (b) ഔ /au/.

3.2.1.4 Consonant Phonemes

Consonant phonemes are produced by restricting the flow of air by articulators and maybe voiced or unvoiced. Consonants can be distinguished by the place of articulation (the point of maximum restriction) and manner of articulation (partial or complete restriction). In Malayalam, consonant phonemes appear as Consonant-Vowel (CV) unit termed as syllable ($\text{ക} < \text{ka} > = \text{ക} / \text{k} / + \text{അ} / \text{a} /$). Before analyzing the speech signal, consonant phonemes must be segmented from the CV unit. Bilabial, labiodental, dental, alveolar, retroflex, palatal, velar, and glottal are consonant phonemes based on the place of articulation. Consonant phonemes are classified based on the point of articulation, as shown in fig. 3.4.

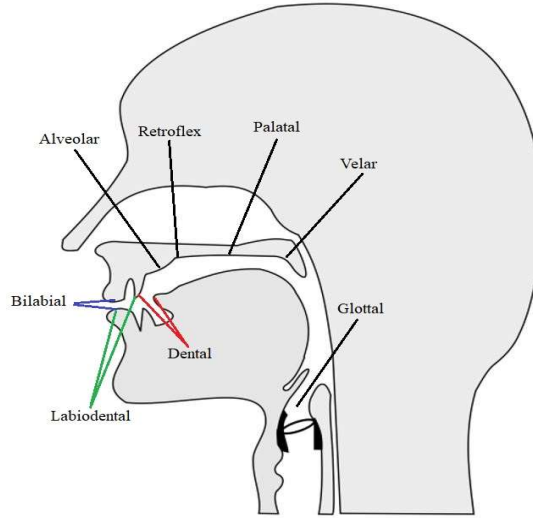


Fig. 3.4 Place of Articulation

Bilabial consonant phonemes are created by restricting the air flow by the two lips coming together. In Malayalam, there are five bilabials, പ് /P/, പ് /p^h/, ബ് /b/, ഭ് /b^h/ and മ് /m/ as shown in fig. 3.5. The consonant boundary is indicated by the red line.

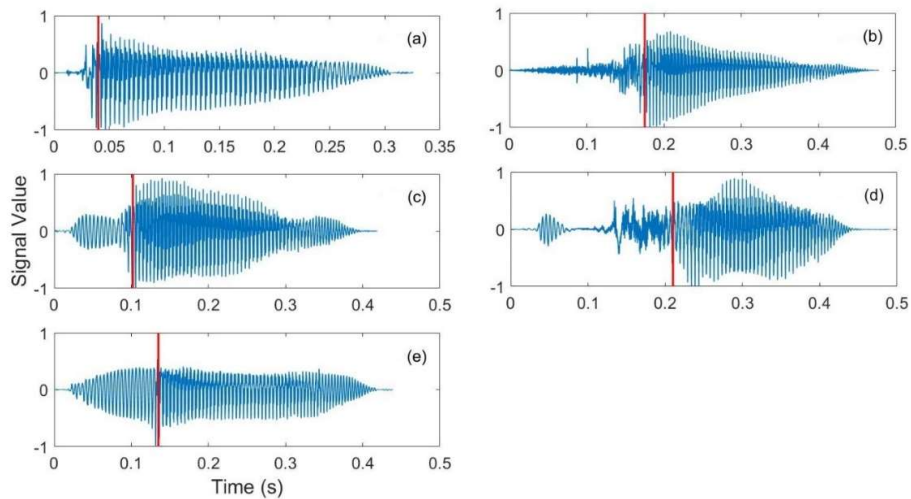


Fig. 3.5 Time Domain Representation of Bilabial Consonant Phonemes: (a) പ് /P/, (b) പ് /p^h/, (c) ബ് /b/, (d) ഭ് /b^h/ and (e) മ് /m/.

Labiodental consonant phonemes are produced by restricting the airflow by touching the upper teeth with the lower lips and then releasing the air by opening the mouth. വ് /v/ is the only labiodental consonant phoneme in Malayalam, and its time-domain representation is shown in fig. 3.6.

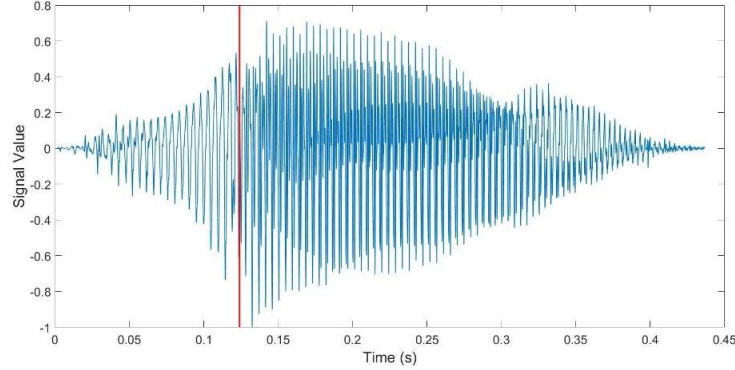


Fig. 3.6 Time Domain Representation of Labiodental Consonant Phoneme: വ് /v/.

Dental consonant phonemes are produced by placing the tip of the tongue behind the teeth. The consonant phonemes ത് /t/, ത് /t^h/, ദ് /d/, ദ് /d^h/ and ന് /n/ belongs to this category.

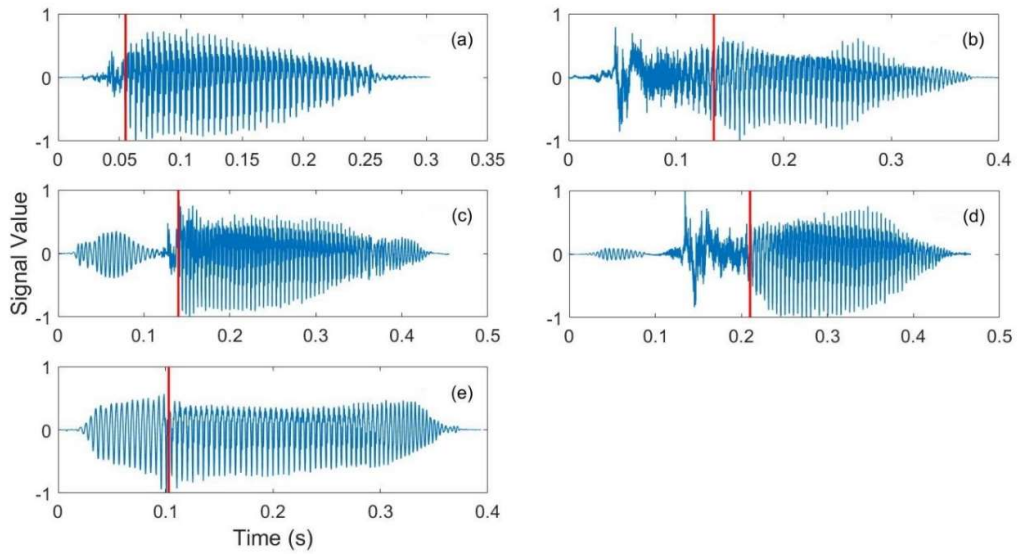


Fig. 3.7 Time Domain Representation of Dental Consonant Phoneme: (a) ത് /t/, (b) ത് /t^h/, (c) ദ് /d/, (d) ദ് /d^h/ and (e) ന് /n/.

Alveolar consonant phoneme is produced by limiting the airflow by pressing the tongue tip on the alveolar ridge, which is the part of the mouth's roof immediately behind the upper teeth. The consonant phonemes like റ്റ /r̥/, ന്ന /n/, സ്സ /s/, ര്ര /r/, റ്റ /r̥/ and ല്ല /l/ belong to this class.

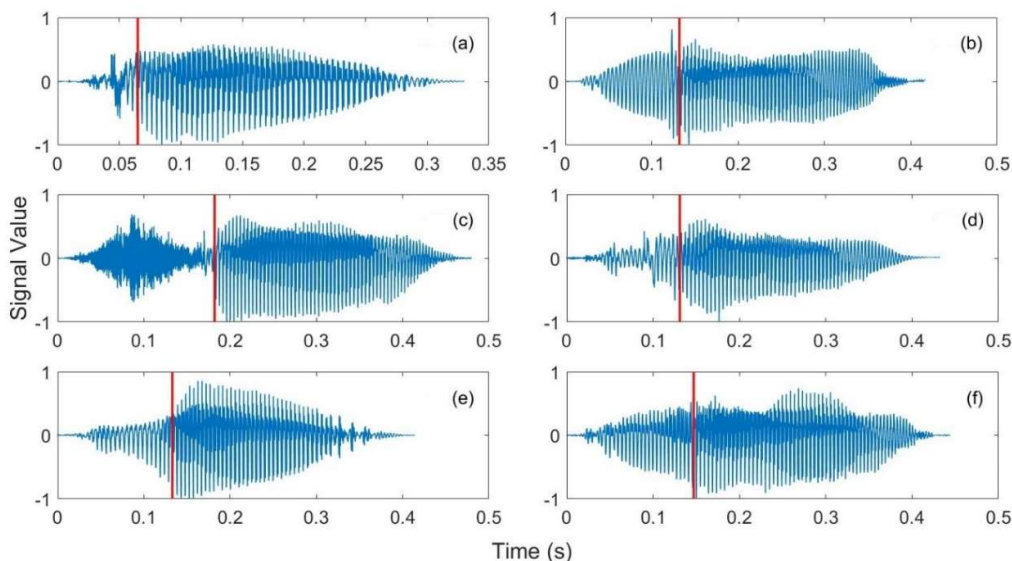


Fig. 3.8 Time Domain Representation of Alveolar Consonant Phonemes: (a) റ്റ /r̥/, (b) ന്ന /n/, (c) സ്സ /s/, (d) ര്ര /r/, (e) റ്റ /r̥/ and (f) ല്ല /l/.

The retroflex sound is produced when the tongue articulates between the alveolar ridge and the hard palate. It is the largest consonant phoneme group in Malayalam, which includes ട് /t̠/, ട് /t̠ʰ/, ഡ് /d̠/, ഡ് /d̠ʰ/, ണ് /ɳ/, ണ് /ɳ̠/, ള് /ʃ/ and ള് /ʃ̠/ (fig. 3.9). Palatal consonant phonemes are made by a constriction between the tongue's tip and the mouth's roof (palate). The consonant phonemes like ച് /ç/, ച് /çʰ/, ജ് /j/, ത് /t̪ʰ/, ന് /ɲ/, ശ് /ʃ/ and യ് /j/; its time-domain representation is shown in fig. 3.10. When the back of the tongue is in contact with the velum (soft palate), velar consonant phonemes are produced. The phonemes that belong to this categories are ക് /k/, ക് /kʰ/, ഗ് /g/, ഗ് /gʰ/ and ണ് /ŋ/ as shown in fig. 3.11. Glottal consonant phoneme was made by closing the far back of the oral cavity (glottis). Only glottal consonant phoneme in

Malayalam is ഹ് $/h/$, whose time-domain representation is shown in fig. 3.12.

Based on the point of articulation, 38 consonant phonemes are grouped into eight groups.

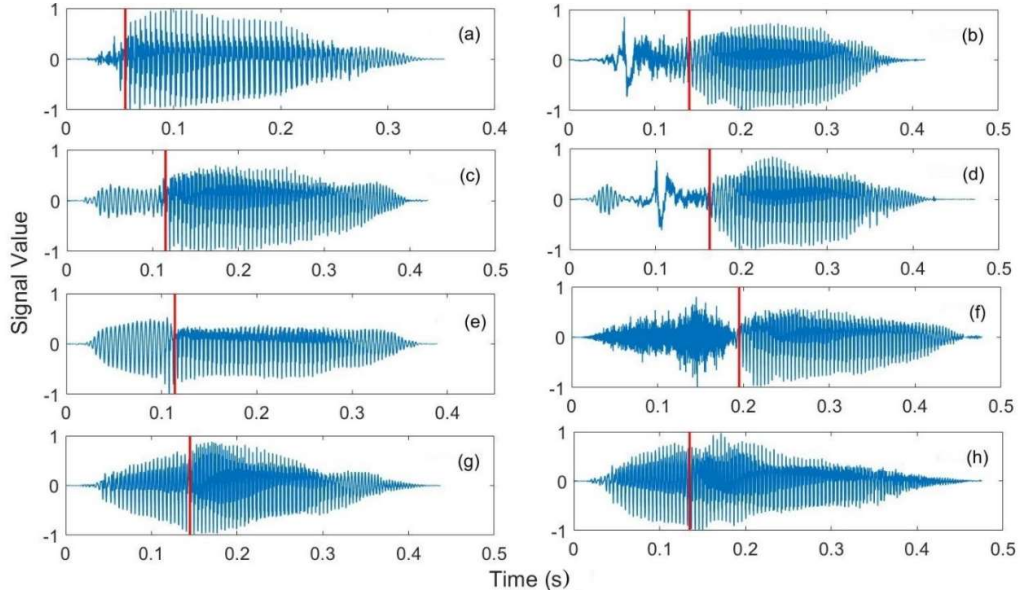


Fig. 3.9 Time Domain Representation of Retroflex Consonant Phonemes: (a) ട് $/t/$, (b) റ് $/t^h/$, (c) ഡ് $/d/$, (d) ഡ് $/d^h/$, (e) ണ് $/n/$, (f) ണ് $/ɳ/$, (g) ള് $/l/$ and (h) ഴ് $/z/$.

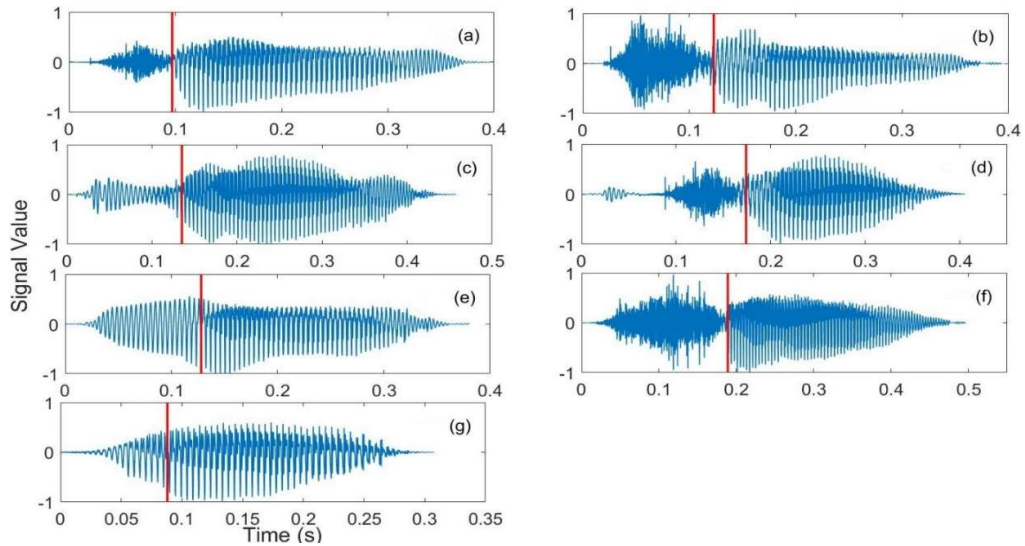


Fig. 3.10 Time Domain Representation of Palatal Consonant Phonemes: (a) ച് $/c/$, (b) ച് $/c^h/$, (c) ജ് $/j/$, (d) ഞ് $/j^h/$, (e) ണ് $/ɲ/$, (f) ശ് $/ʃ/$ and (g) യ് $/y/$.

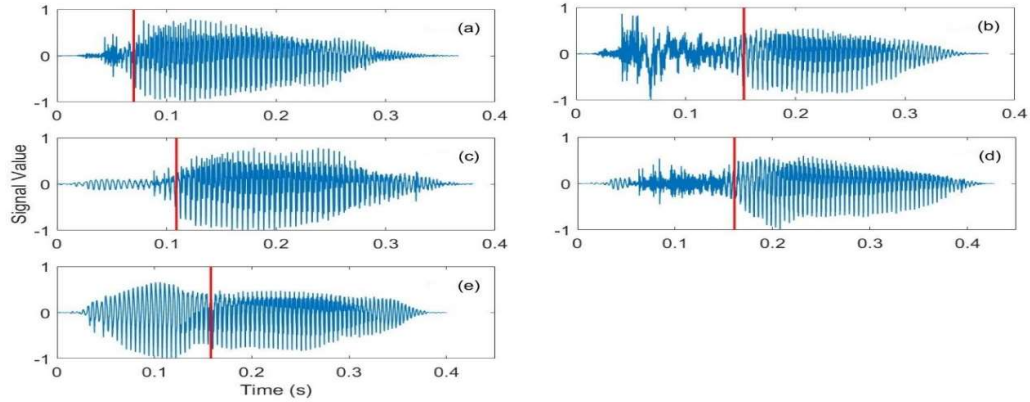


Fig. 3.11 Time Domain Representation of Velar Consonant Phonemes: (a) ക്ക /k/, (b) ക്ക^h /k^h/, (c) ഗ്ഗ /g/, (d) ഗ്ഗ^h /g^h/ and (e) ണ്ണ /ŋ/.

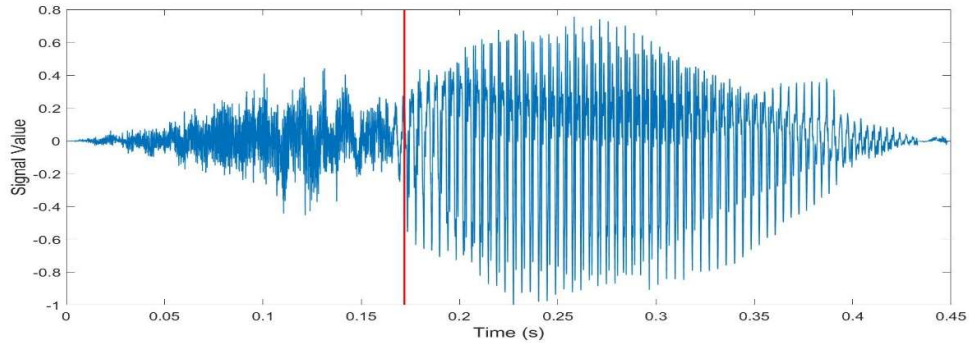


Fig. 3.12 Time Domain Representation of Glottal Consonant Phoneme: ഹ്ഹ /h/.

Point of articulation refers to the point at which the constriction is made in the oral cavity. Manner of articulation represents the extend of constriction made for a consonantal gesture. In Malayalam, manner of articulation was broadly classified into four classes: Plosives (Stops), Nasal, Fricatives and Semivowels.

Plosives (Stops) are made by completely blocking the airflow for a short duration termed as closure, and then the air is released, termed as a release. Based on the state of vocal cords during the closure time, plosives are divided into unvoiced plosives and voiced plosives. During the production of unvoiced plosives, vocal cords do not vibrate along with the complete closure of the vocal tract, creating short-duration silence followed by the abrupt release of air. On

the other hand, the vocal cords vibrate when the vocal tract is closed, making a short low-frequency signal during the closing time. Unvoiced and voiced plosives are subdivided into two groups based on whether aspiration (stable glottis airflow) is present or not.

Nasal sounds are the results of coupled action of an oral and nasal tract. They are caused by glottal excitement and complete narrowing in the vocal tract somewhere in the mouth cavity. However, the velum is lowered, which opens the nasal passage to allow airflow through the nostrils.

Fricative sounds are produced by partially constricting the airflow through the vocal tract. It is generated by a continuous airflow that is turbulent at the constriction site and characterised by a hissing tone, stimulating the vocal tract.

Semivowels are vowel-like sounds produced by gliding transition in vocal tract area function between the adjacent phonemes, like the vowels and diphthongs. It is divided into four groups: Trill/flapped, lateral, approximant and glide. Flap is a quick flip of the tongue against the alveolar ridge. Lateral sound is produced by raising the tip of the tongue towards the roof of the mouth so that the air flows along both sides of the tongue. Approximant sounds are made when the tongue moves close to the roof of the mouth but not close enough to cause turbulence. Thus, approximant falls between the characterized of vowel and fricative sounds. Finally, in glide sound, airflow is frictionless and modified by the tongue and lips position.

Thus, combining the place of articulation and manner of articulation is enough to classify the Malayalam consonant phonemes uniquely. In Malayalam, all vowels are voiced sounds, while there are 23 voiced consonants (Voiced Plosives, Nasals, and Semivowels) and 15 unvoiced consonants (Unvoiced Plosives and Fricatives). A comprehensive classification of Malayalam consonant phonemes is shown in table 3.2.

Table 3.2 Linguistic Classification of Malayalam Consonant Phonemes

Place of Articulation	Manner of Articulation									
	Plosive					Semivowel				
	Unvoiced Unaspirated	Unvoiced Aspirated	Voiced Unaspirated	Voiced Aspirated	Nasal	Fricative	Trill/ Flapped	Lateral	Approximant	Glide
Bilabial	പ്/P/	ഫ്/pʰ/	ബ്/b/	ഭ്/bʰ/	മ്/m/					
Labiodental										വ്/v/
Dental	ത്/t/	ഥ്/tʰ/	ദ്/d/	ധ്/dʰ/	ന്/n/					
Alveolar	റ്/r/				സ്/s/	ര്/r/	ല്/l/			
Retroflex	ട്/ɽ/	ഠ്/ɽʰ/	ഡ്/d/	ഢ്/dʰ/	ണ്/n/	ഷ്/ʃ/	ള്/ɭ/			
Palatal	ച്/c/	ഛ്/ç/	ജ്/j/	ഝ്/jʰ/	ഞ്/n/	ശ്/ʃ/				യ്/y/
Velar	ക്/k/	ഖ്/kʰ/	ഗ്/g/	ഘ്/gʰ/	ങ്/ŋ/					
Glottal										ഹ്/h/

3.2.1.5 Allophones

In the Malayalam language, there are ten vowel phonemes, two diphthongs, and 38 consonant phonemes. Based on articulation, among the ten vowels, four are classified as front vowels, two as central vowels and four as back vowels. Based on the relative position of articulators, the consonant phonemes are categorized as bilabial, labiodental, dental, alveolar, retroflex, palatal, velar, and glottal. An allophone is a version of a phoneme that is phonetically distinct [84]. The place or phonetic surroundings in the word generally characterises the allophones of the same phoneme. Because these differences do not assist in identifying one word from another, speakers of a language sometimes have trouble recognising the phonetic differences between allophones of the same phoneme. There are 207 words accommodating 106 Malayalam allophones, which include 75 consonant allophones, 28 vowel allophones, and three allophones corresponding to diphthongs. Table 3.3 & 3.4 shows the list of Malayalam vowel phonemes and consonant phonemes with its corresponding allophones. The consonant phonemes are arranged in a front to the back manner (from lip to glottal) of the articulator's position.

Table 3.3 Malayalam Vowel and Diphthong Phonemes with its Allophonic variations

SI. No.	Vowel Phoneme IPA	Vowel Allophone IPA	SI. No.	Vowel Phoneme IPA	Vowel Allophone IPA
1	ഇ /i/	[i] [y ⁱ] [y ⁱ]	7	ഉ /u/	[^w u] [u ^w] [u] [ə] [ə*] [u ^v] [U]
2	ഈ /i:/	[y ⁱ :] [i:]	8	ഊ /u:/	[^w u:] [u]
3	എ /e/	[y ^e] [e ^v] [E]	9	ഒ /o/	[^w O] [O]
4	ഏ /e:/	[y ^e :] [e ^r :] [e:]	10	ഓ /o:/	[^w O:] [O:]
5	അ /a/	[Λ] [A]	11	ഐ /ai/	[ai] [ei]
6	ആ /a:/	[a:] [a]	12	ഔ /au/	[au]

Table 3.4 Malayalam Consonant Phonemes with its Allophonic variations

SI. No.	Consonant Phoneme IPA	Consonant Allophone IPA	SI. No.	Consonant Phoneme IPA	Consonant Allophone IPA	SI. No.	Consonant Phoneme IPA	Consonant Allophone IPA
1	പ് /P/	[p] [β] [b] [P]	14	സ് /s/	[s]	27	ച് /c ^h /	[c ^h] [C ^h]
2	ഫ് /p ^h /	[p ^h]	15	ര് /r/	[r]	28	ജ് /j/	[J] [j]
3	ബ് /b/	[B] [b]	16	റ് /r̄/	[r̄]	29	ത് /j ^h /	[J ^h]
4	ഭ് /b ^h /	[b ^h] [ḡ ^h]	17	ല് /l/	[l] [d]	30	ന് /n/	[n]
5	മ് /m/	[M] [m] [ṁ]	18	ട് /t/	[t̄] [t] [T]	31	ശ് /ʃ/	[ʃ]
6	വ് /v/	[w] [v] [t]	19	ഠ് /t ^h /	[t ^h] [T ^h]	32	യ് /y/	[y]
7	ത് /t/	[t ^h] [ṭ] [ḍ]	20	ഡ് /d/	[d]	33	ക് /k/	[k] [kj] [y̥] [g]

								[t]
								[K]
								[k ^h]
8	ഥ് /t ^h /	[t ^h]	21	ഡ് /d ^h /	[d ^h]	34	ഖ് /k ^h /	[K ^h]
								[K ^h]
		[d]	22	ണ് /ɳ/	[ɳ]	35	ഗ് /g/	[G]
9	ദ് /d/	[d]						[g]
10	ഡ് /d ^h /	[d ^h]	23	ഷ് /ʃ/	[ʃ]	36	ഘ് /g ^h /	[g ^h]
11	ന് /n/	[n]	24	ള് /l/	[l]			[ɳ]
		[n]						[ɳj]
12	റ്റ് /ɾ/	[d]	25	ഴ് /z/	[z]	37	ഞ് /ɳ/	[ɳ<]
		[t]						[ɳ>]
								[ɳ']
					[c]			
13	ന് /n/	[n ^h]	26	ച് /c/	[c]	38	ഹ് /h/	[H]
		[n]			[ɟ]			[h]
					[ʃ]			
					[C]			

3.2.2 Database Acquisition Hardware

Cost-effective recording devices were used for documenting the audio-visual speech in this database. The video signal is recorded with two handy cameras in the first and third categories of recording. High-quality visual utterances are captured using Sony handy camera HDR-CX405. The video signal is recorded at a frame rate of 25 fps and a resolution of 1280 x 720. The audio signal is sampled at 44100 Hz. This camera is used to capture the dynamical variation of the mouth region with audio and video signals wrapped in MP4 format. Low-quality visual signals are captured using Sony handy camera DCR-SX45E. The video signal is captured at a frame rate of 25 fps with a resolution of 720 x 576 in MPG format. The audio signal is sampled at 48000 Hz. This camera is employed for capturing the speaker’s face in both categories with audio and video signals wrapped in MPG format. Both cameras are mounted on a tripod stand with a bubblehead to adjust the horizontal position of the camera. Two LED video lamps with varying intensity ranging up to 400W is used in the recording environment to capture the in-depth information from the speaker’s mouth. Two different types of audio capturing hardware are used in the second category of recording. First, a standard headphone with a microphone (with foam cover) captures sound in a controlled environment. Second, an ordinary mobile headset with a microphone was employed in the acoustically realistic environment. The recorded audio files are saved as a wav file at a sampling frequency of 16 kHz. Before processing, all audio speech data is resampled to 16 kHz.

3.2.3. Recording Setup

The database entitled “MOZHI” consists of 3 categories of recording. A single audio-only speech database acquisition category and two audio-visual speech database acquisition categories. Recording in each stage begins with an explanation of the objectives to be achieved. During each utterance, the

speakers are advised to close their mouths at the beginning and end of each utterance. Speakers start their utterance only a few seconds after the recording device starts; this strategy is utilised in all categories to capture the background noise and channel distortion at the beginning of the speech. Most of the data is recorded in a studio-like environment with provisions to record the signal in controlled and uncontrolled conditions.

The first recording category contains an audio-visual speech database recorded to highlight the visual features of the speaker’s mouth in a linguistically rich environment. Speaker is instructed to keep their head in quasi-static condition and exhibit a neutral facial expression during recording. Two handy cameras (high quality and low quality) are set to record the frontal view of the speaker. The cameras are mounted on a tripod stand about 100cm away from the speaker's face. The low-quality camera is placed a little bit vertically above the high-quality camera. The low-quality camera is zoomed in to capture only the speaker’s face and shoulder. The high-quality camera is zoomed in to capture the mouth region, from chin to tip of the nose. Both cameras are operated by a single speaker, which helps to maintain the synchronization between them. In addition, two LED video lamps are placed on both sides of the handy camera at equal distance, thereby aligning all in a single line with the same vertical height. The light is slightly tilted towards the speaker's face, and the intensity was adjusted to obtain uniform illumination on the speaker's face. The presence of a background screen reduces the complexity related to speaker face segmentation as in the third category of recording. This orientation is maintained throughout the recording process taken at different times. The speaker can practice elocution on the language material provided to them and modify their pronunciation. The language material is provided in printed form (which contain 50 Malayalam isolated phonemes and 106 words). Speaker is advised to widen their mouth so that the entire mouth region should be in the frame while uttering such postures. Before recording, the camera

screen is turned towards the speaker to adjust their position to achieve the desired view in both cameras. Speaker is instructed to speak in their natural manner and allow them to out of the frame while they made or felt a mistake and let them in the frame after rectifying it. The audio signals are captured using the inbuilt audio recorder in the two handy cameras synchronising with the visual cues. The audio and video are wrapped in MP4 files from the high-quality camera and MPG files from the low-quality camera. The approximate footage length is 5 minutes for isolated phonemes and 20 minutes for connected words for each speaker. These captured raw audio-visual speech databases contain redundant and unwanted information which should be removed and arranged systematically for implementation in speech-based applications. Fig. 3.13 shows the recording setup of the first category.

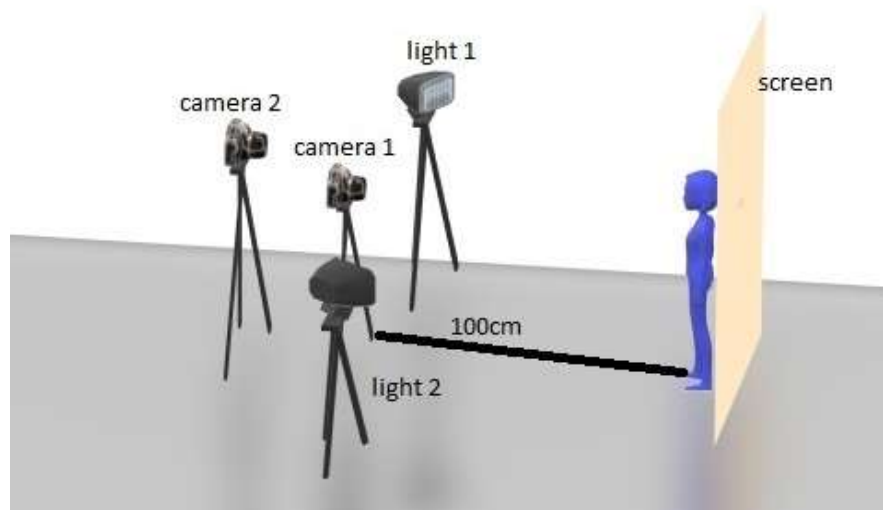


Fig. 3.13 MOZHI recording setup.

The second category consists of an audio-only speech database which has two sub-categories. The first sub-category is clean audio signal is recorded in an acoustically isolated laboratory environment using a standard headphone with a microphone near the mouth region, enabling good audio acquisition. Before recording, the channel’s gain is limited, thereby attaining fewer

background noisy signals and reducing the channel clipping issues. The speakers are requested to repeat each utterance five times and is saved as a single wav file for each phoneme and allophone. In addition, speakers are requested to repeat the utterance if necessary, either because of a mistake during articulation or if the recorded sound contains noise due to respiration or channel problem. All the speakers are post-graduate students within the age group of 21-25 years. The second sub-category is an audio speech database recorded in an acoustically realistic environment (office, school, and house) using an ordinary mobile headset with a microphone. The speakers can utter five short vowels ten times each, which the operator controls since the age group utilized for this task is 5-60. This database is used to study the effects of ageing in human speech organs and to investigate better noise-robust techniques in speech-based applications.

The third recording category is dedicated to capturing the audio-visual speech database in an acoustically and visually realistic environment. This recording is made in a lab environment with natural daylight and ordinary illumination in a room. The room may contain possible background noises from electrical equipment and human conversation inside and outside the room. The speaker is seated at the centre of the room to create illumination variation on the speakers face. This recording is done in the same manner as the first category except without LED lamps and background screen. The approximate footage length was 10 minutes for each speaker. After the recording process, the speakers signed a consent form.

The recording setup is intentionally adopted to implement a human-machine speech-based interacting system in the public domain using low cost, affordable performance and easily installable equipment is used in this recording process. The third recording category can be implemented in an open space in a public domain like a hospital, railway station, airport etc., which produces the same recording environment as in the third category. The first recording category can be implemented in a closed cabin (like a telephone

booth) with all facilities to fabricate the same recording environment, thereby ensuring better performance. Table 3.5 gives the complete picture of this database named “MOZHI- An Audio-Visual Speech Database in Malayalam Language”. As an initial work in the Malayalam language, the rest of the thesis is done with an audio-visual speech database taken from the first recording category. The last two categories of records will be utilized in future work.

Table 3.5 MOZHI Database Profile

	Modality	Utterance & Speakers	Purpose	Equipment
Category-1	Audio-Visual (Lip region)	50 phonemes-5 times each and 206 isolated words comprising of all allophones-3 times each.	Audio-visual speech processing in Malayalam language.	High quality Camera (1280 x 720) with two professional LED video light
	Audio-Visual (Face region)	Speakers- 25 females and five males.		Low quality Camera (720x 576) with two professional LED video light
Category-2	Audio-only	50 phonemes- 5 times each. Speakers- 20 females and 10 males in the age group 21-25.	Audio-only speech processing.	Good quality microphone placed closer to mouth region.
		Five Short Vowels- 5 times each. Speakers- 10 females and 10 males in each age group ranging from 5 to 60	Effects of aging on human speech organs. Effects of real-time noise in speech signal processing.	Ordinary microphone placed closer to mouth region
Category-3	Audio-Visual (Lip region)	50 phonemes and real time isolated words-5 times each.	Real-time audio-visual speech processing.	High quality Camera (1280 x 720)
	Audio-Visual (Face region)	Speakers- 5 females		Low quality Camera (720x 576)

3.3 Audio and Video Segmentation and Labelling

The major work after the recording was the processing of audio and visual speech data individually and preparing it in a standard format for further usage. The audio and visual data processing involves the segmentation and labelling of audio and video files separately for each isolated phonemes and word, thereby removing redundancy in both domains. This database contains audio files with less storage space when compared to the visual domain and video files with massive storage space, which makes manual processing an error-prone and time-consuming process. Automatic processing has its drawback while dealing with manually compromising features involving an individual difference in appearance, speaking style, pronunciation, instructions passed to the speaker, misreading etc. So, it is reliable to consider a semi-automatic viewpoint in the processing step.

The audio signals are segmented and labelled automatically, and the segmented files is analysed manually for omitting files that contain noise and wrong pronunciations. The recorded individual audio files contain the same sound repeated several times with enough separation between each other. This audio signal is passed through a spectral subtraction process, which removes the noisy background content that arises during recording. The spectral subtraction process is implemented only on the audio files recorded in a lab environment where the noise changes slowly relative to the speech signal. However, the audio files, which are recorded in an acoustically realistic condition (Category 2), are retained to study the effect of the real-time noisy speech signal in future work.

In the spectral subtraction method [143], the clean speech and noise are uncorrelated and additive in the time domain. In addition, most of the additive noise and the channel distortion change very slowly as to the speech signal.

Hence, a recorded speech signal is generally realised as a combination of convolved channel distortion and additive noise,

$$y(m) = x(m) * d(m) + n(m) \quad (3.1)$$

Where m stands for discrete-time index, $y(m)$ for recorded speech, $x(m)$ for clean speech, $d(m)$ for channel distortion, $n(m)$ for additive noise, and $*$ for convolution operator. Assuming the absence of channel distortion, the power spectrum of the noisy speech signal is the sum of clean speech power spectrum $X(k)$ and noise power spectrum $N(k)$ as in Eq. (3.2), where k is frequency bin index.

$$|Y(k)|^2 = |X(k)|^2 + |N(k)|^2 \quad (3.2)$$

According to the recording strategy, the noise power spectrum is estimated in the non-speech region. The estimated noise power spectrum is then subtracted from the power spectrum of the noisy speech signal to obtain the clean speech power spectrum as in Eq. (3.3).

$$|X(k)|^2 = |Y(k)|^2 - |N(k)|^2 \quad (3.3)$$

The inverse discrete Fourier transform (IDFT) transforms the obtained clear speech power spectrum into the time domain by combining its magnitude with the phase information gained from the noisy speech input.

$$x(m) = \sum_{k=0}^{k=M-1} |X(k)| e^{\frac{-j2\pi}{M} km} e^{j\theta_Y(k)} \quad (3.4)$$

The phase information from the noisy speech signal is represented by $\theta_Y(k)$. As noise varies randomly, spectral subtraction might produce negative results. Since the power spectrum is always positive, the negative value will impact the recovered signal, which will be visible at low signal-to-noise ratios. In a weak background noisy speech signal, this issue is unnoticeable. The block diagram of the spectral subtraction method is shown in Figure 3.14.

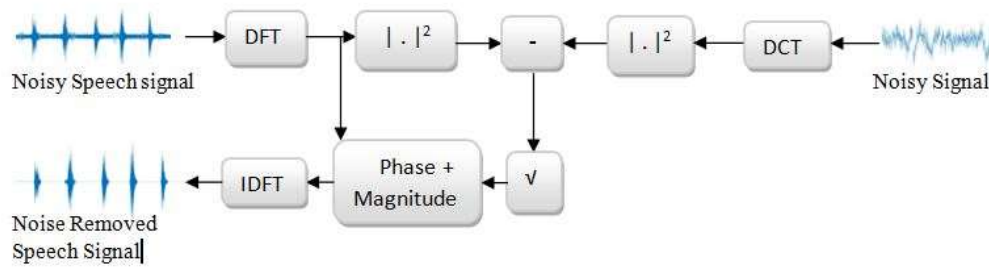


Fig. 3.14 Block Diagram of Spectral Subtraction method.

After the noise removal step, the lengthy speech signal (containing many utterances of isolated phonemes) goes through the segmentation process. In the segmentation process, the signal is segmented into 10 ms frames. The maximum of the absolute values of the samples in each frame is compared with the maximum of the absolute values of the whole signal. The first frame with a maximum of the absolute values of the samples in the frame greater than a threshold value (8% of the maximum of the absolute values of the whole signal) is found. Just before this frame, the second frame is considered the first frame of the isolated phoneme ($frame_{start}$). Then the first frame, after $frame_{start}$, with the maximum amplitude value less than a threshold value (about 8% of the maximum of the absolute values of the whole signal) is found. After this frame, the following second frame is considered the last frame of the isolated phoneme ($frame_{end}$). All the frames from $frame_{start}$ and $frame_{end}$ are saved as the first utterance. The time information of starting and ending frame is utilized to segment the corresponding visual counterpart. All frames up to $frame_{end}$ are removed from the original signal. The process is repeated until all the isolated phonemes are segmented. Fig 3.15 shows the steps involved in the segmentation of acoustic speech signals.

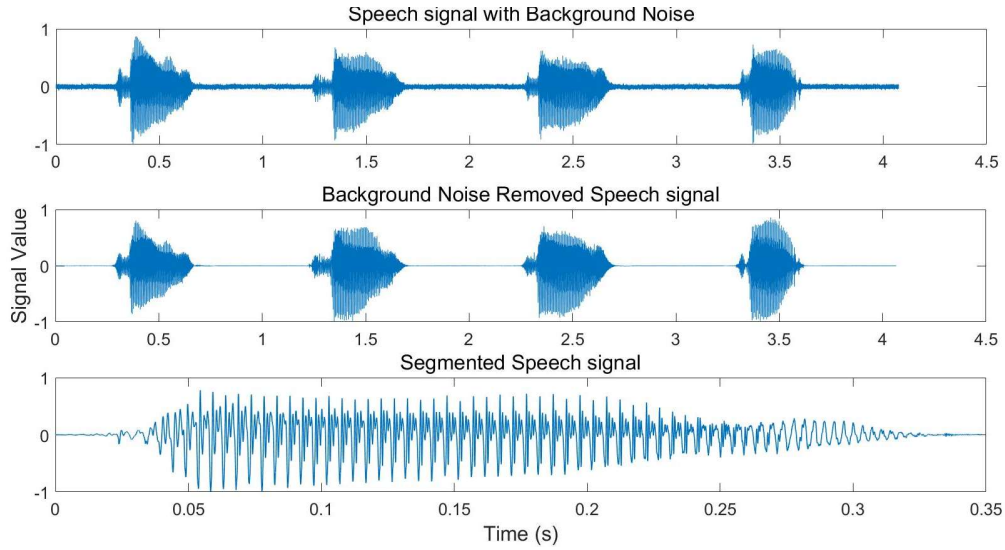


Fig. 3.15 Speech Segmentation Process.

After this process, each segmented speech signal from the same utterance is viewed acoustically and visually, and the misspelt and corrupted speech signals are removed manually. After scrutinising all segmented speech signals, it is forwarded to the labelling process, which contains all relevant information regarding the signal. The labelling process was done entirely automatically in all category of recording, which minimizes the errors occur during manual labelling.

It is decided to keep each audio and visual file in a different folder under the heading of the corresponding category, which will drop out the necessity of including category information in the labelling process. The first four characters represent the label of the isolated phonemes (the maximum characters required to label is for $\text{Ṭ}/\text{r}/$ (TTAA)). For phonemes with a label length of less than four characters, the excess characters are filled with X (like AXXX for Ṭ /a/). An additional character (single digit) is allotted to allophones for displaying the context-based variation of phonemes. The maximum context based phonemic variation is for the phoneme $\text{Ṭ}/\text{u}/$ (7 classes

as in table 3.4). The following two characters represent the repetition of utterance (from 01 to 99) since some speakers repeated the utterance more than nine times. The next character (7th for phoneme and 8th for allophone) represent the gender of the speaker (‘F’ for females and ‘M’ for males). The proceeding two characters represent the age of the speaker from 05 to 60. The next character represents the nativity of the speaker. The nativity of the speaker is represented by the first letter of the city/street. The last two characters represent the speaker number (each speaker is assigned a unique two-digit number) since the second recording category needs a considerable number of speakers. The first 6 and 7 characters are used to provide the information regarding the phoneme and allophone, respectively. The proceeding seven characters are used to represent the speaker identity. So, an audio speech file is named with 12 and 13 alphanumeric characters for phonemes and allophones, respectively. The same nomenclature is implemented for audio speech files of categories 2 and 3.

The same rule for labelling has been used for the video signal. For each speaker, two videos were captured simultaneously, with one containing only lip information and the other with complete facial information, which is specified in the format of the video file. All original audio and video data have been stored along with corresponding standardized data under the proper headings so that further examinations will still be possible. Table 3.6 displays the nomenclature of MOZHI database files.

Table 3.6 Naming rule of MOZHI Database files

	Sound	Repetition	Gender	Age	Place	Speaker Number	Total Length	Stream	Format
Phoneme	AXXX	19	M	31	V	17	12	Audio	.wav
								Video – Lip Region	.mp4
								Video - Face	.mpg
								Audio	.wav
Allophone	UXXX7	03	F	23	C	09	13	Video – Lip Region	.mp4
								Video – Face	.mpg

3.4 Simulated Noisy Signal

In signal processing, noise is considered as a signal with various frequency components at varying strengths that can distort the nature of the original signal during the recording, processing and transmission of the signals. In the real world, various noises distorted speech signals, reducing their perceptual quality and intelligibility. Low perceptual quality results in listener fatigue, and poor intelligibility yields poor performance in different speech-based applications. Vowels and consonants are the two types of speech signals. Vowel sounds have low-frequency features, whereas consonant sounds have high-frequency ones. Consonants provide most of the information in the speech signal. As a result, it is preferable to investigate noises that degrade speech in various frequency ranges. Three noise signals were used by analysing the spectral characteristics: White Gaussian noise, Pink noise, and Red noise. These noises were treated as coloured noise, which uses the idea of colour to describe its frequency response.

Any real-world speech-based application must address its performance in different noisy conditions with a different signal-to-noise ratio (SNR). The signal-to-noise ratio is a parameter used to characterises the relative strength of the signal with the noise. SNR is defined as the ratio of signal power to the noise power, often expressed in decibels (dB). SNR in terms of power is defined in linear scale as

$$SNR = \frac{P_{signal}}{P_{noise}} \quad (3.5)$$

Where P_{signal} is the power of the clear signal and P_{noise} is the power of the noisy signal. SNR in decibel (dB) scale is given by, Eq. (3.6). The SNR values of the speech signal include in this database, and their corresponding linear scale representation are shown in table 3.7.

$$SNR (dB) = 10 \log_{10} (SNR) \quad (3.6)$$

Table 3.7 SNR and Corresponding Relation between Signal Power and Noise Power

SNR in dB	SNR in linear scale
20	Signal power = 100*Noise power
10	Signal power = 10*Noise power
0	Signal power = 1*Noise power
-10	Signal power = $\frac{1}{10}$ *Noise power
-20	Signal power = $\frac{1}{100}$ *Noise power

The algorithm for the generation of the coloured noisy signal is summarised below.

Step1: Create a random signal from the Gaussian distribution of finite length.

Step2: Apply Fast Fourier Transform (FFT)

Step3: To manipulate the left half of the spectrum, remove the symmetric portion from the spectrum.

Step4: For white noise, Power spectrum density (PSD) is flat.

-Retain the spectrum without any modification.

For Pink noise, PSD =K/f, where K is a constant and f is the frequency.

-Manipulate the spectrum by dividing the spectrum amplitude with the square root of frequency indexes. Since power is proportional to amplitude squared, the power per Hz will decline at higher frequencies at the rate of about -3dB per octave or -10dB/decade.

For Red noise, $PSD = K/f^2$, where K is a constant and f is the frequency.

- Manipulate the spectrum by dividing the spectrum amplitude with frequency indexes. The power per Hz will decline at higher frequencies at the rate of about -6dB per octave or -20dB/decade.

Step5: Reconstruct the whole spectrum.

Step6: Take Inverse Fast Fourier Transform (IFFT) (Convert noisy signal from frequency domain to time domain).

Step7: Ensure zero mean and unity standard deviation.

The created coloured noise is mixed with speech signal additively at different SNR values. To generate the noisy signal of the desired SNR, convert the given SNR in dB into the linear scale and substitute in the equation below. Then combine the speech signal with the noisy signal with the appropriate SNR.

$$\text{Noise of desired } SNR = \sqrt{\frac{P_{\text{signal}}}{SNR_{\text{linear}}}} * \text{noise}[m] \quad (3.7)$$

3.4.1 White Gaussian Noise

White Gaussian noise or white noise has a uniform power distribution over all frequencies from 0 to half the sampling frequency. White noise, for example, has the same strength between 100 and 500 Hz as between 20,000 and 20,500 Hz at a sampling rate of 44,100 Hz. Thus, a sequence of statistically uncorrelated random numbers generated from a Gaussian distribution, usually with zero mean and unity variance, is characterized as white noise. When the TV or radio is tuned to an unused frequency, a human ear hears pure white noise as a hissing sound.

Fig. 3.16 visualise the properties of White Gaussian noise. Fig 3.16 (a) shows the time-domain representation of white noise with zero mean and unit

variance. The frequency response of white noise in the semi-log graph is shown in fig. 3.16 (b). In a semi-log graph, one axis is in logarithmic scale and the other in liner scale; here, the spectral magnitude is in logarithmic scale. Power Spectral Density (PSD), power per unit frequency, is the frequently used tool for noise analysis, characterising the average behaviour of fluctuating quantities. Even though the spectrum is not an exact flat for white noise, the “average power” is the same for all frequencies. The histogram of white noise matches the theoretical probability distribution function of the Gaussian random variable, which is shown in fig. 3.16 (c). The autocorrelation function measures the time fluctuating quantities related to ‘t’ and shifted time ‘t+τ’. Fig. 3.16 (d) shows the autocorrelation function of White noise equal to an impulse at zero lag, which ensures that the random variation of white noise is highly uncorrelated. Fig. 3.17 shows the effect of white Gaussian noise in the time and frequency domain.

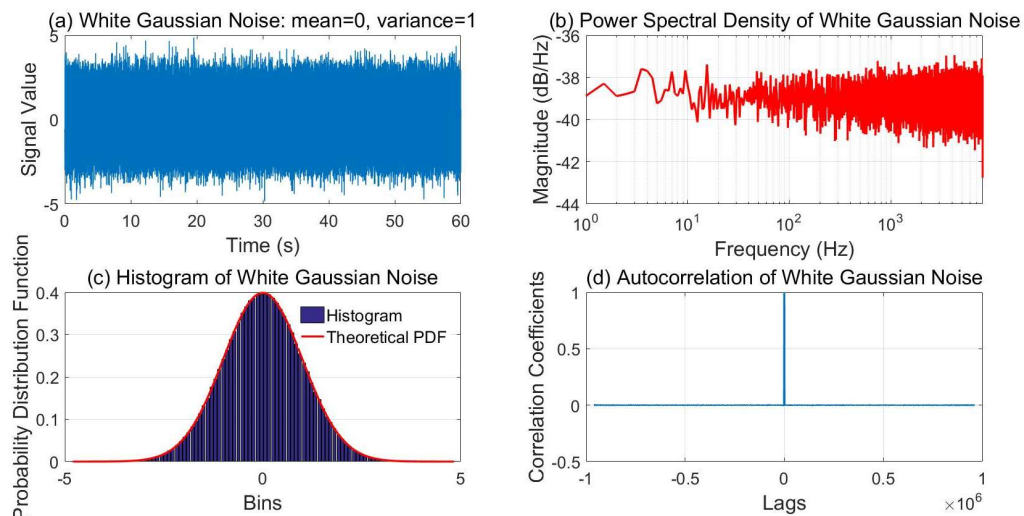


Fig. 3.16 Properties of White Gaussian Noise.

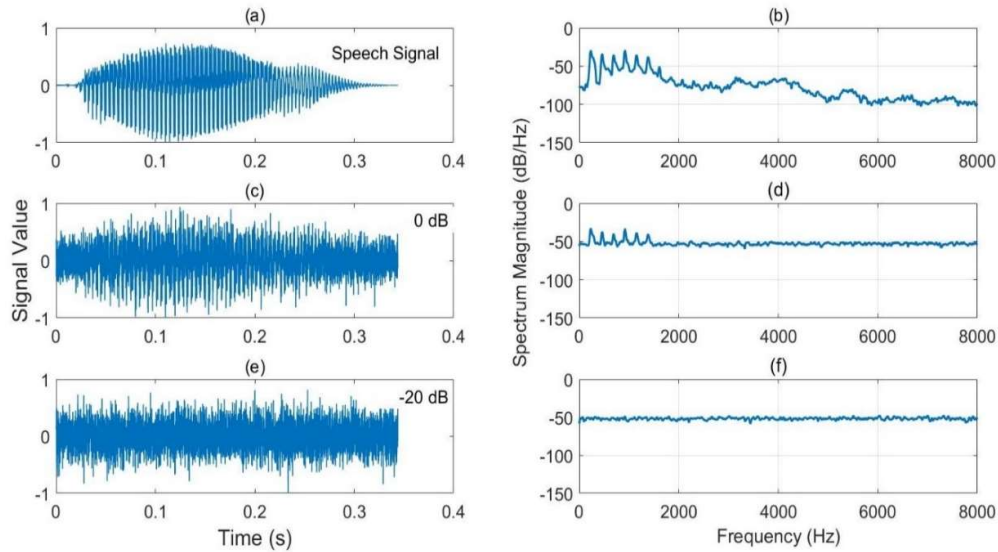


Fig. 3.17 Effect of White Gaussian Noise on Speech signal in Time and Frequency domain.

3.4.2 Pink Noise

Pink noise, sometimes known as "1/f noise," is a signal with a power spectral density that is inversely proportional to its frequency. Pink noise has the same power distribution throughout each octave in the logarithmic scale; therefore, the power between 200 and 400 Hz is the same as that between 2,000 and 4,000 Hz. Because power is proportional to amplitude squared, the power in each constant bandwidth of pink noise decreases at higher frequencies at a rate of around -3dB per octave or -10dB per decade. Thus, Pink noise is equivalent to white noise, which falls off at -3db/octave. Many physical processes in the natural world produce noise with a power distribution like pink noise; flicker noise in electronics is one of them.

Fig 3.18 (a) shows the time-domain representation of pink noise. The frequency response of pink noise in the semi-log graph is shown in fig. 3.18 (b). The spectral property of pink noise is presented by marking the rolling rate at -10dB/decade. The histogram of pink noise exactly matches the theoretical probability distribution function of the Gaussian random variable, which is

shown in fig. 3.18 (c). Pink noise can be viewed as a low-pass filtered white noise. Since the high-frequency components were removed, the random samples exhibit a less sharp transition between each other, which is displayed as in fig. 3.18 (d) when compared to 3.16 (d). Fig. 3.19 shows the effect of pink noise in the time and frequency domain.

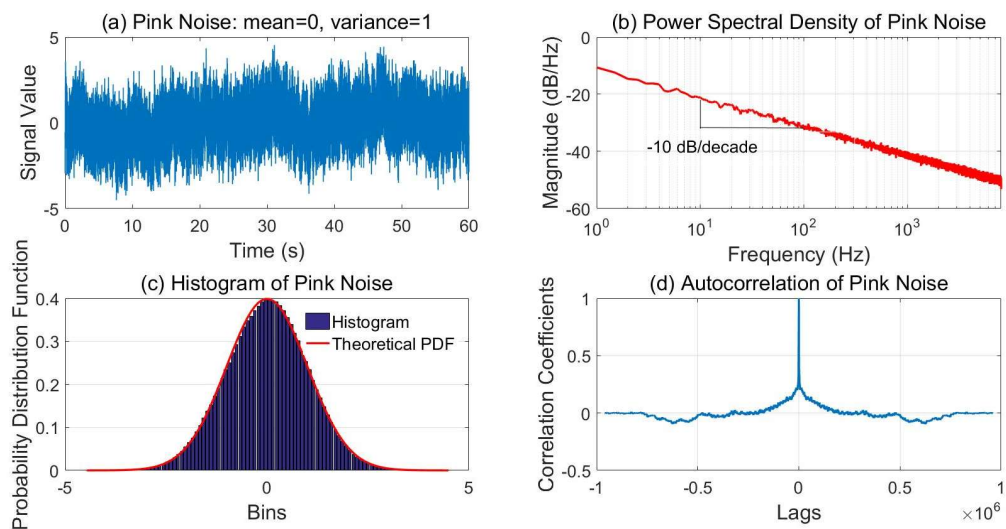


Fig. 3.18 Properties of Pink Noise

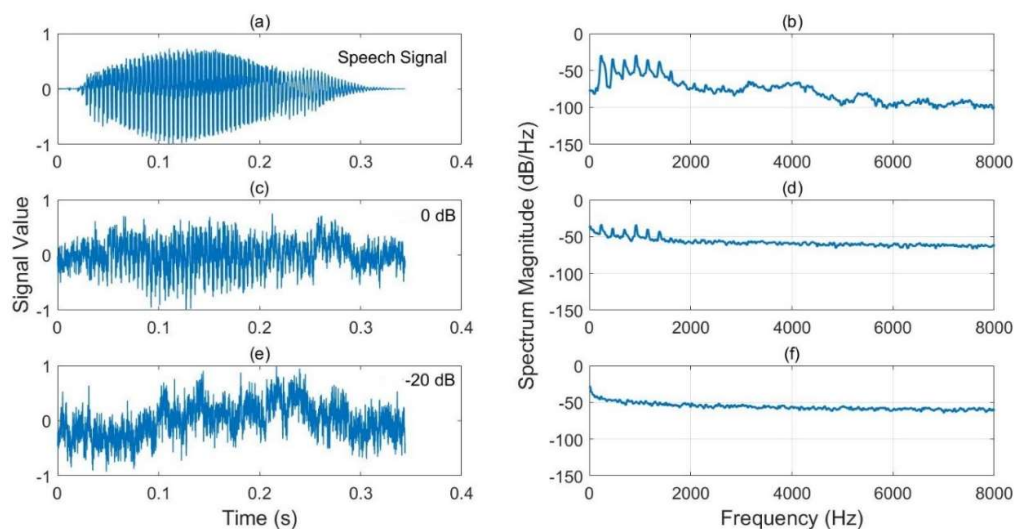


Fig. 3.19 Illustration of effect of Pink Noise on Speech signal in Time and Frequency domain

3.4.3 Red Noise

The deeper variant of pink noise, known as red, brown, or Brownian noise, has a power spectrum proportional to $1/f^2$. Therefore, lower frequencies have more energy than higher frequencies in brown noise. Brown noise is sometimes known as red noise because it has a low frequency analogous to red light. At higher frequencies, red noise rolls off at a rate of around -6dB/octave or -20dB/decade. As a result, red noise is the same as white noise, which falls off at a rate of -6dB/octave.

Fig 3.20 (a) shows the time-domain representation of red noise, which wanders up and down, but there is a clear correlation between successive values. The frequency response of red noise in the semi-log graph is shown in fig. 3.20 (b). The spectral property of red noise is presented by marking the rolling rate at -20dB/decade. The histogram of red noise deviates from the theoretical probability distribution function of the Gaussian random variable, which is shown in fig. 3.20 (c). In fig. 3.20 (d), the autocorrelation function displays a repetitive form due to the high correlation between the samples in Red noise. Finally, fig. 3.21 shows the effect of red noise in the time and frequency domain.

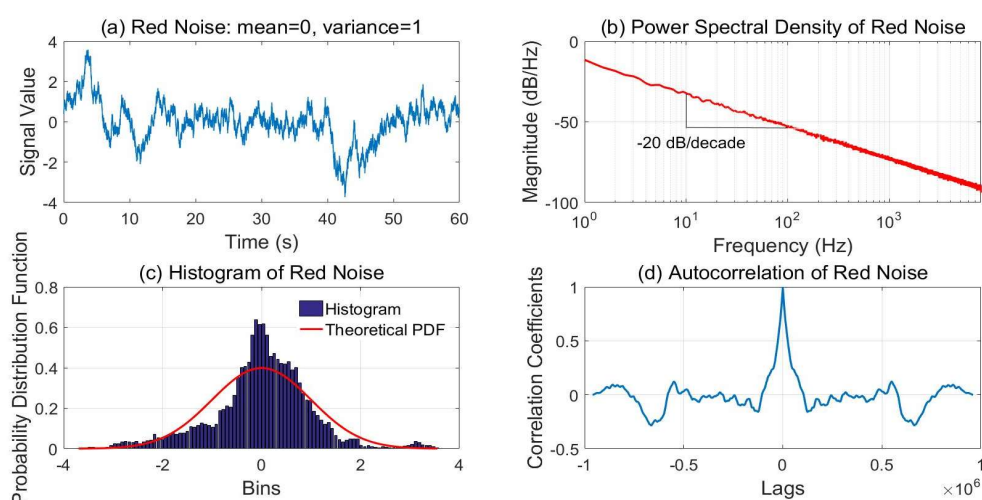


Fig. 3.20 Properties of Red Noise

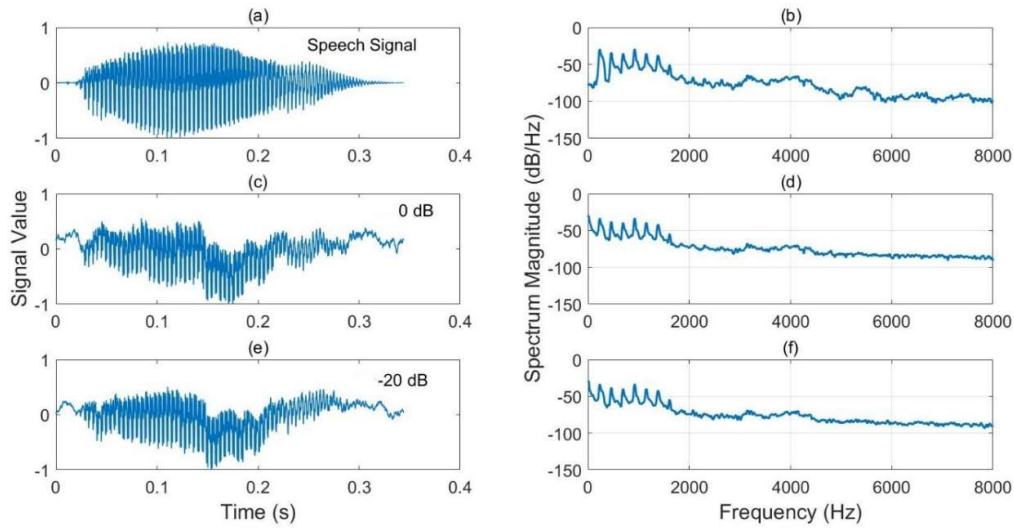


Fig. 3.21 Illustration of effect of Red Noise on Speech signal in Time and Frequency domain

Before studying the effect of the noise signal in speech, the corrupted speech signal must be properly labelled. Up to this point, each phoneme and allophones were represented by 12 and 13 alphanumeric characters, respectively (as in table 3.6). For adding noise information, additional four characters are used. The first character is represented the first letter of noise used in this database, i.e., ‘W’ for White noise, ‘P’ for Pink noise, ‘R’ for Red noise. The next three characters are representing the noise level in dB scale-like +20, +10, +00dB, -10dB, and -20dB. A corrupted phoneme file is labelled with 16 alphanumeric characters, and 17 alphanumeric characters represent its corresponding corrupted contextual variation. This labelling method helps to select the suitable subset of the database and to get all the relevant information about the speech signal.

3.5 Audio-Visual Asynchrony

The durational analysis of phonemes and allophones, its visual counterpart, and hence audio-visual asynchrony is discussed in this section. It is the first work done in Malayalam. Duration modelling is the preliminary task in phonemes or viseme level speech processing applications like speech

recognition and speech synthesis system. The segmented audio and video speech files contain silence on both sides of a vowel (V), consonant-vowel (CV) and word utterance. So, the duration of the underlined vowel and consonant phonemes must be properly time-aligned in audio and video files.

Even though the durational analysis for audio and video speech files was done separately, the correlation between the two modalities must be thoroughly examined, especially in a bi-modal speech processing task. Two major issues arise when comparing acoustical speech with visual speech: difference in frame rate and asynchrony. Based on the database, the visual speech is captured at a rate of 25 Hz (25 fps) and in the thesis, the acoustical speech was analysed in a short duration, namely ‘window’ or ‘frame’ of 25 ms with a sliding window of 10 ms, creating 100 fps. This difference must be equalized by upsampling the video frame rate. In addition, upsampling the video files is nothing but capturing the same sequence in a 100 fps camera with the same duration. The frame rate does not create an issue in this section since durational analysis is performed on the entire speech signal. It is well known that while uttering a sound, the speaker’s mouth starts moving or opening a little bit later or earlier than the acoustic signal, thereby producing an asynchrony (in the range of milliseconds) between audio and visual cues. This problem must be significantly addressed in bimodal speech-based applications. Since both modalities are captured simultaneously by the same devices, which help to consider only the natural asynchrony rather than asynchrony occur during the recording of both modalities separately. There's an inherent asynchrony between both visual and sound cues of language. Speech is generated through the closely coordinated motion of many articulators. Because of coarticulation effects and articulator inertia, the sound and visual cues might not be precisely synchronized at any certain time. After uttering a phoneme, it's not possible for the muscles of our articulatory system to instantly alter the positions of the various articulators to generate the following sound. Visual address suffers both from anticipatory and preservatory coarticulation effects. Preservatory or backward coarticulation usually means a speech gesture proceeds after uttering

a sound section, whereas the other gestures necessary to make this sound are already finished [104]. In short, the visual expressions found following the corresponding phoneme discovered in preservatory coarticulation. Besides, anticipatory or forward coarticulation occurs when a visible gesture of a language segment occurs ahead of other articulatory elements of the segment. Therefore, in anticipatory coarticulation, visual expressions have been observed before the corresponding phoneme is heard. Linguistic exploration is necessary to handle this problem since the scope and directionality of the coarticulation effect is extremely language-dependent.

To deal with this issue in the Malayalam language, we used our audio-visual speech database containing all phonemes and allophones of five speakers. Phonemes are utilized to assess the asynchrony because of articulator inertia alone. Allophones address the asynchrony by considering coarticulation effects and articulator inertia. From the recorded data, audio signals are extracted from video and stored as separate files. The video file was then converted into frames. Asynchrony is then estimated as the difference in the time duration of the acoustic and visual speech signal of the underlined phonemes or allophones. Asynchrony analysis was performed individually for phonemes and allophones to underline the coarticulation effect. The duration of each phoneme and allophones in the audio and video speech signal are estimated manually. Manual phonetic boundary identification is the most favourable approach in database creation. Even though it is time-consuming, this method is helpful when there is no clear distinction between neighbouring speech elements, especially in consonant phoneme boundary estimation. During the acoustic speech time alignment process, both waveform appearance (visual perception) and sound (auditory perception) are synchronously utilized to identify the location of the change in both attributes. The information from the articulator’s motion is explored in viseme (visual equivalent of the phoneme) boundary identification. A relatively rapid change in the appearance of visually perceivable articulators like lips, teeth and tongue is the boundary's identification mark. Due to the speaker variability, certain linguistic intuitions

are also employed to judge the boundary when the teeth' appearance hinders the tongue motion. Two labellers were utilized to do the manual segmentation of our database. For reducing individual bias, the time-aligned audio and visual speech files were rechecked by the author of this thesis, who got training from the linguistic people.

Fig. 3.22 shows the time alignment of acoustical and visual speech of vowel phoneme $\text{അ} /a/$. The first row represents the time-domain representation of speech signal with time in seconds along the x-axis and signal value along the y-axis. The vertical red line indicates the beginning and end boundary point of the underlined phoneme/ allophone. The second and fourth row shows the phoneme labels of the audio and video speech files, respectively. The third row displays the mouth appearance involved in the production of the underlined phoneme. The number beneath the frame shows the corresponding frame number in the video file.

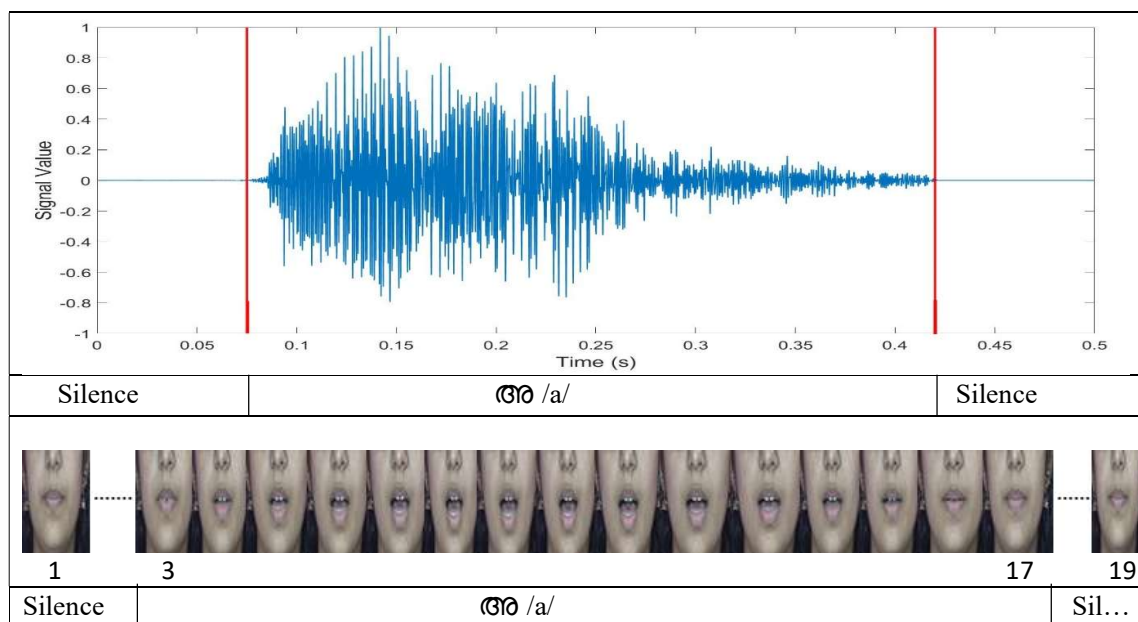


Fig. 3.22 Time alignment of vowel phoneme $\text{അ} /a/$

The dotted lines indicate the silence portion in the video file. For example, frames from 1 to 2 and 18 to 19 are silent regions. The time duration of the speech elements (phoneme/ allophone) is estimated from the time domain itself. However, for visual speech, the time duration of the underlined phoneme is estimated by multiplying the number of frames with the reciprocal of the camera’s fps. From fig. 3.22, the number of frames for uttering vowel phoneme $\text{അ} /a/$ is 15 (from 3 to 17) multiplied with the reciprocal of the fps of our camera (25 fps), creating 0.60 s.

Immense care is given while performing the time alignment of consonant phonemes. Since the consonant phonemes in Malayalam always appear like Consonant-Vowel (CV) syllables, precise boundary detection is possible by repeatedly checking the sound of the selected consonant phoneme and the overlapped vowel phoneme. Fig. 3.23 shows the time alignment of a consonant phoneme $\text{ത} /t/$.

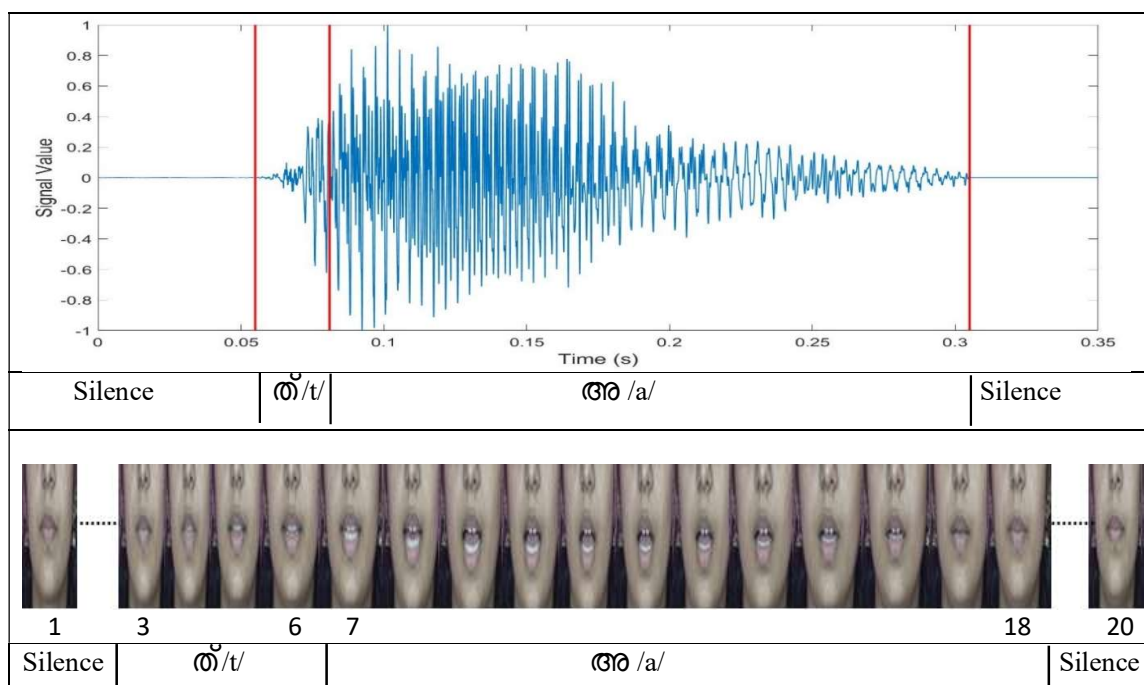


Fig. 3.23 Time alignment of consonant phoneme $\text{ത} /t/$

The coarticulation effect in Malayalam speech is embedded in Allophones. In a word, the occurrence of a speech element is highly influenced by the linguistic properties of the preceding and proceeding speech element and its relative position. എ [ye], എ[eʏ] and എ [E] are the contextual variation of the vowel phoneme എ /e/. Fig. 3.24 shows the time alignment of vowel allophone എ [ye] in the word എവിടെ [evite] with phonetic transcription എ /e/ + വ് /v/ + ഇ /i/ + ട് /t/ + എ/e/. Fig. 3.25 shows the time alignment of vowel allophone എ [eʏ] in the word പിന്നെ [pinne] with phonetic transcription പ് /P/ + ഇ /i/ + ന് /n/ + ന് /n/ + എ /e/. Fig. 3.26 shows the time alignment of vowel allophone എ [E] in the word വെളുത്ത [ve[utta] with phonetic transcription വ് /v/ + എ /e/ + ഉ് /l/ + ഉ /u/ + ത് /t/ + ത് /t/ + അ /a/.

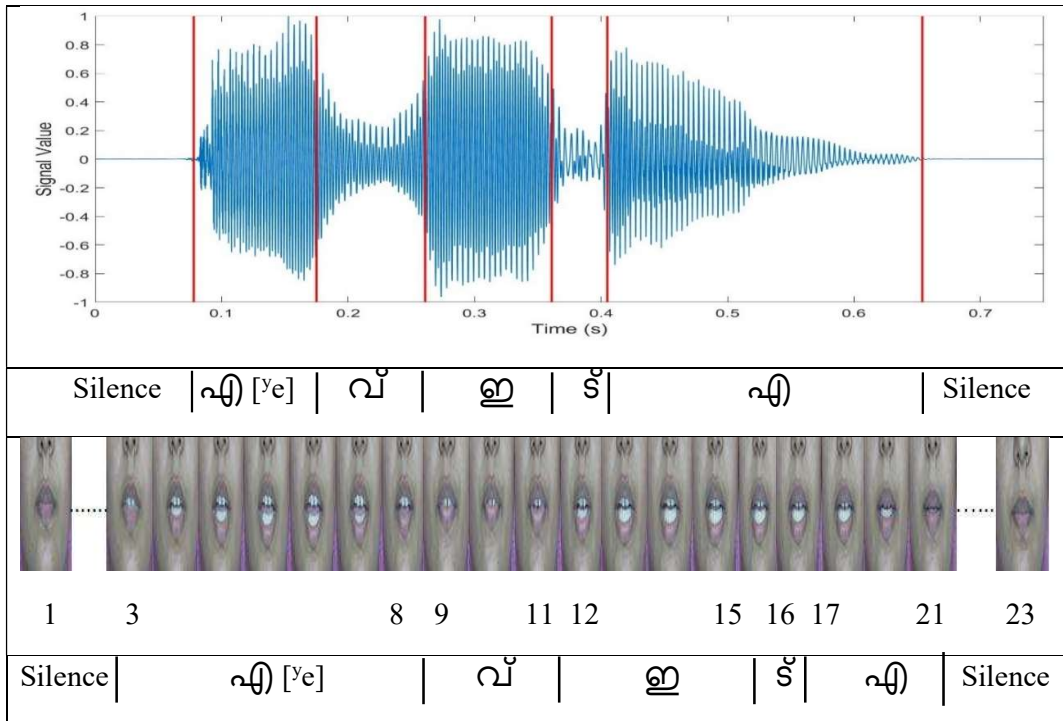


Fig. 3.24 Time alignment of vowel allophone എ [ye] in the word എവിടെ [evite]

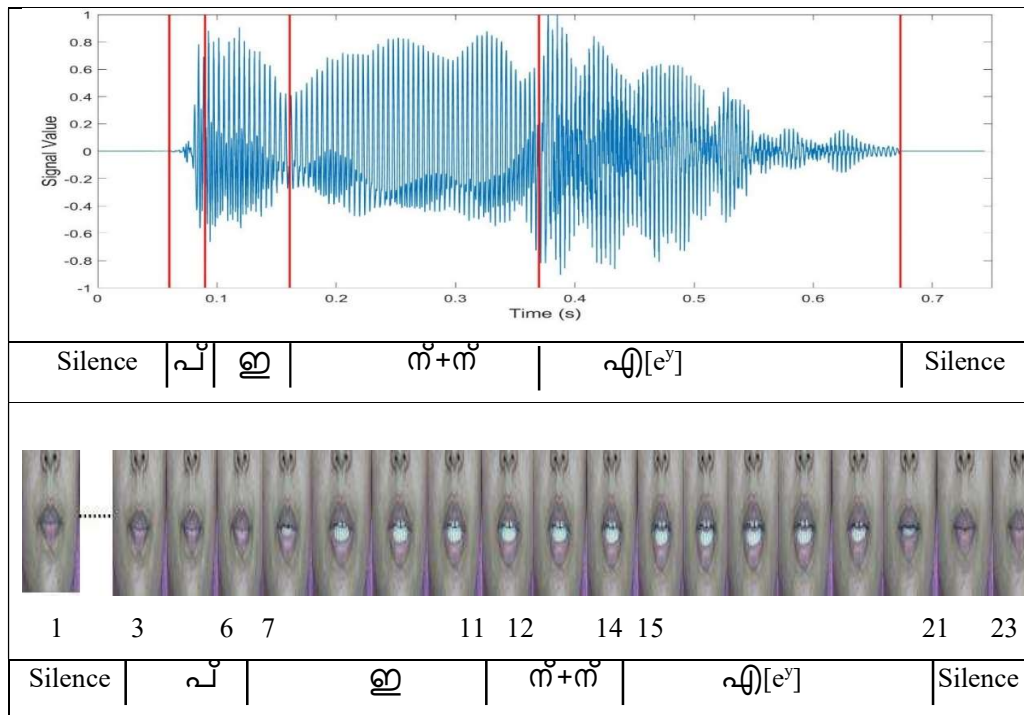


Fig. 3.25 Time alignment of vowel allophone എ [eʏ] in the word പിന്നെ [pinne]

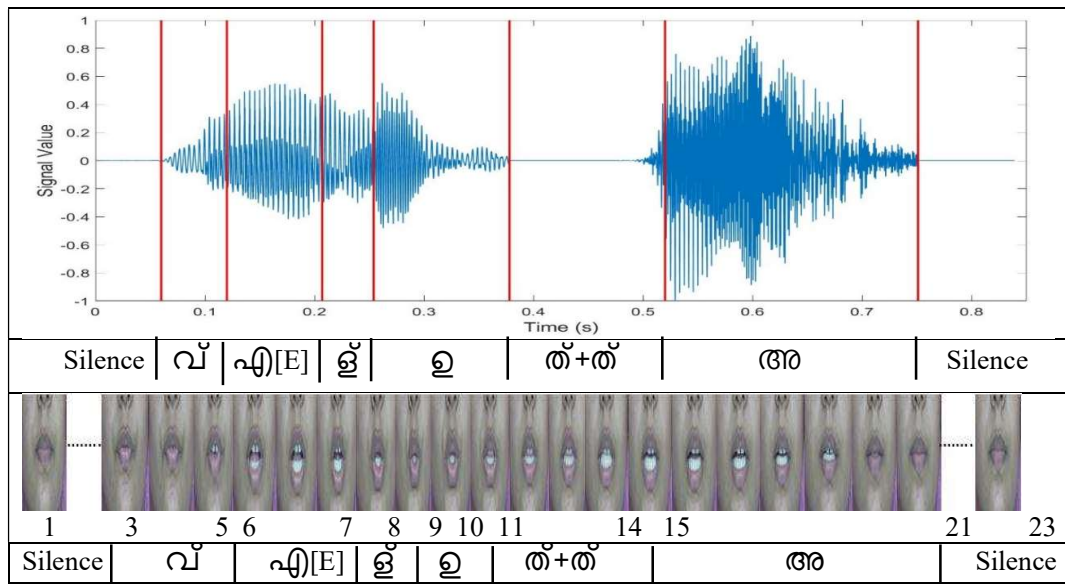


Fig. 3.26 Time alignment of vowel allophone എ [E] in the word വെളുത്ത [ve[utta]

Fig. 3.27 and 3.28 shows the graphical representation of durational statistics of Malayalam phonemes from acoustic and visual speech, respectively. The triangle in the stock plot represents the mean duration of the corresponding phoneme, and the length of the line from the top and bottom of the triangle indicates “mean + SD” (Standard Deviation) and “mean – SD”, respectively. In fig. 3.27, the short vowel phonemes roughly varied between 0.250 s to 0.450 s, and the long vowel phonemes were shifted above in between 0.450 s to 0.650 s. The consonant phonemes were distributed between 0.050 s and 0.250 s. In fig. 3.28, all the short vowel phonemes were shifted between 0.500 s and 0.650 s, and long vowel phonemes varied between 0.600 s and 0.700 s. The consonant phonemes were distributed between 0.100 s and 0.300 s. The Malayalam vowel and consonant phoneme duration from the acoustic speech is 0.411 ± 0.068 s and 0.113 ± 0.029 s, respectively. The visual speech’s Malayalam vowel and consonant phoneme duration is 0.61 ± 0.04 s and 0.20 ± 0.04 s, respectively. While comparing the fig. 3.27 with fig. 3.28, the whole plot is uplifted, especially vowel phonemes, which means, for almost all Malayalam phonemes, acoustic speech is lag the visual speech.

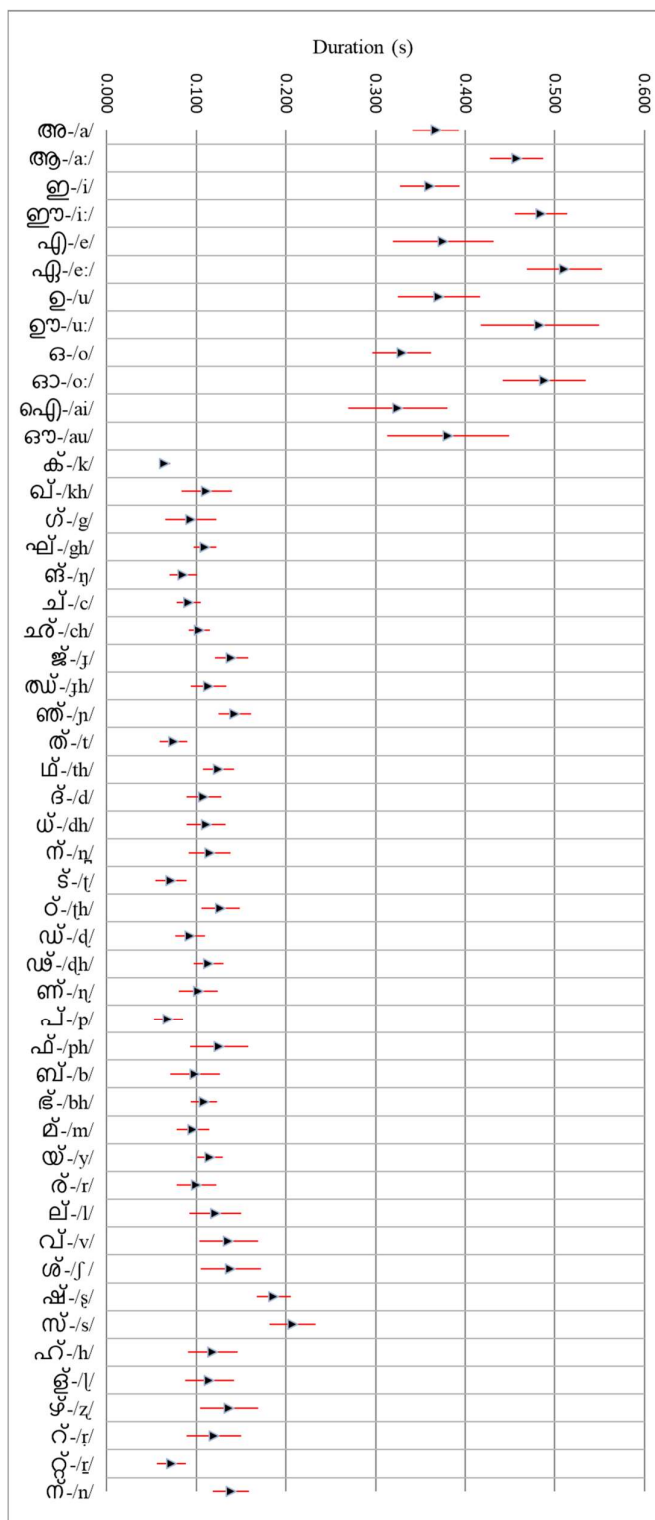


Fig. 3.27 Graphical representation of Durational Statistics of Malayalam phonemes from acoustic speech

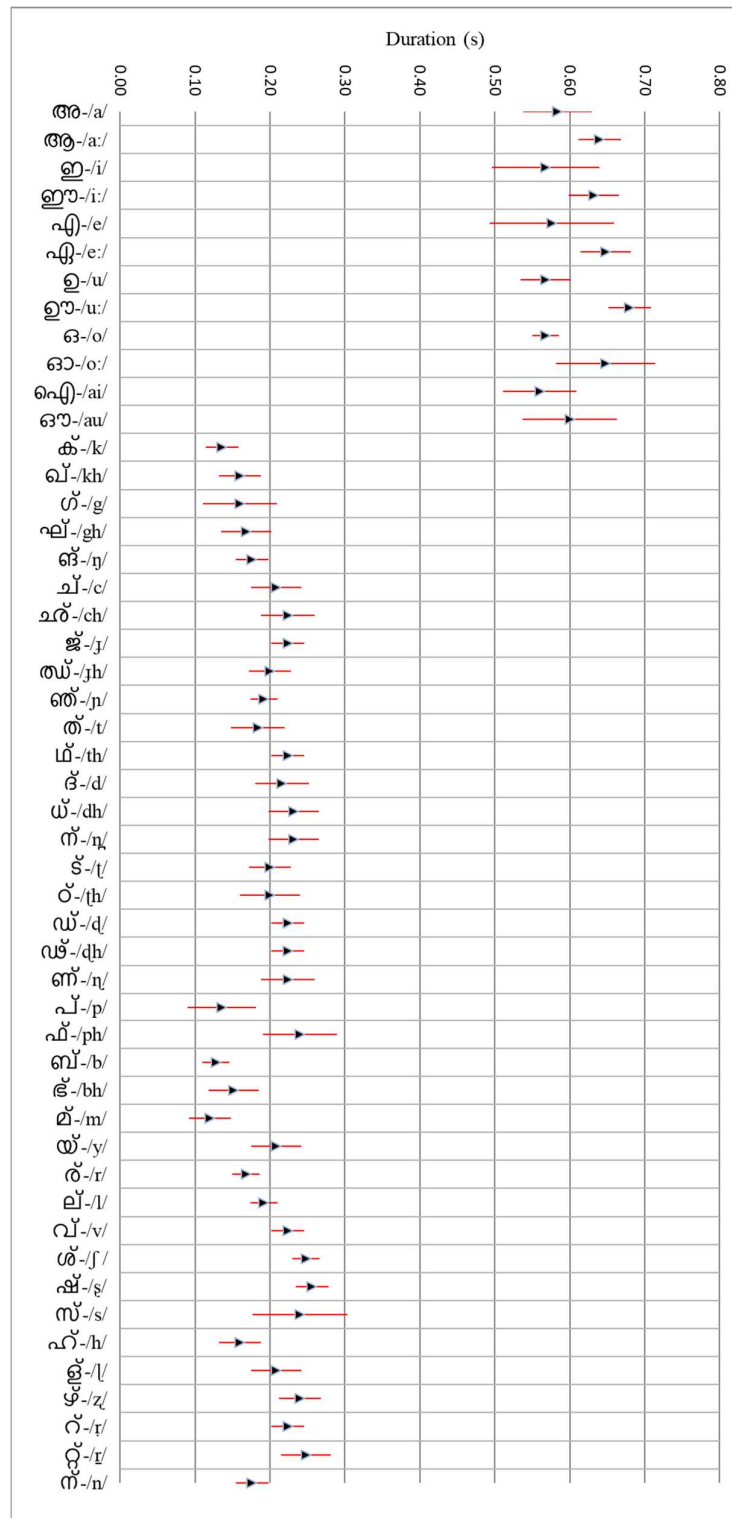


Fig. 3.28 Graphical representation of Durational Statistics of Malayalam phonemes from visual speech

Fig. 3.29 and 3.30 show the graphical representation of the durational statistics of Malayalam vowel allophones from acoustic and visual speech. The numbers associated with each phoneme represent the different types of allophones for the concerned phoneme. In fig. 3.27, the vowel phonemes varied between 0.250 s to 0.650 s but in fig. 3.29, the vowel allophones varied between 0.005 s to 0.400 s. In addition, fig. 3.28 displays the vowel phonemes varying between 0.500 s to 0.700 s but in fig. 3.30, the vowel allophones were varying between 0.005 s to 0.400 s. While comparing the fig. 3.27 with fig. 3.29 and fig. 3.28 with fig. 3.30, the whole plot is lowered due to the coarticulation effect and the articulation inertia of underlined vowel phoneme in a word. In addition, a clear durational distinguish is observed between corresponding short and long vowel allophones in acoustic speech, which is not observed in visual speech.

Fig. 3.31 and 3.32 show the graphical representation of the durational statistics of Malayalam consonant allophones from acoustic and visual speech. In fig. 3.27, the consonant phonemes from the acoustic speech varied between 0.005 s to 0.250 s but in fig. 3.31, the consonant allophones varied between 0.025 s to 0.350 s. In addition, fig. 3.28 displays the consonant phonemes from visual speech varying between 0.100 s to 0.300 s but in fig. 3.32, the consonant allophones were varying between 0.025 s to 0.400 s. While comparing the fig. 3.27 with fig. 3.31 and fig. 3.28 with fig. 3.32, the coarticulation effect and the articulation inertia of underlined consonant phonemes were less affected when compared with vowel phonemes in a word. The duration of the Malayalam vowel and consonant allophone from the acoustic speech is 0.189 ± 0.077 s and 0.121 ± 0.064 s, respectively. The visual speech's duration of Malayalam vowel and consonant allophone is 0.27 ± 0.10 s and 0.16 ± 0.06 s, respectively.

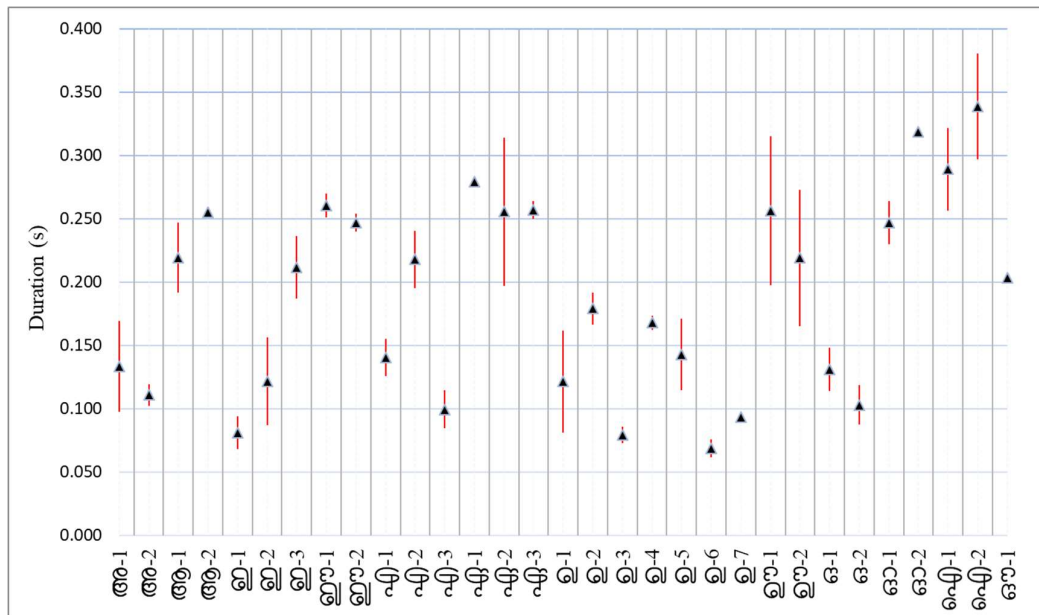


Fig. 3.29 Graphical representation of Durational Statistics of Malayalam vowel allophones from acoustic speech

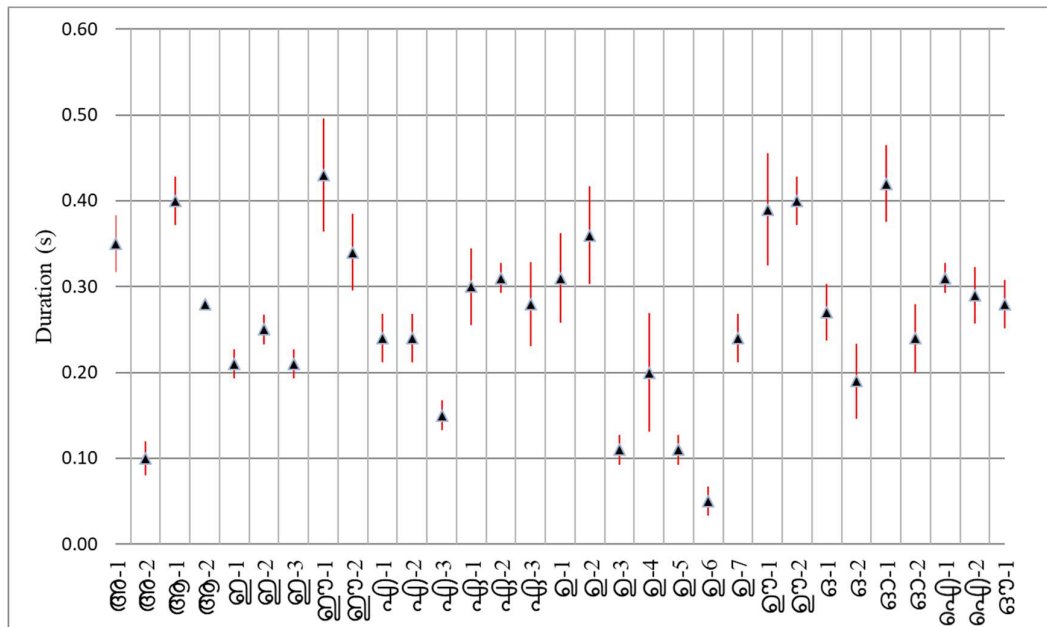


Fig. 3.30 Graphical representation of Durational Statistics of Malayalam vowel allophones from visual speech

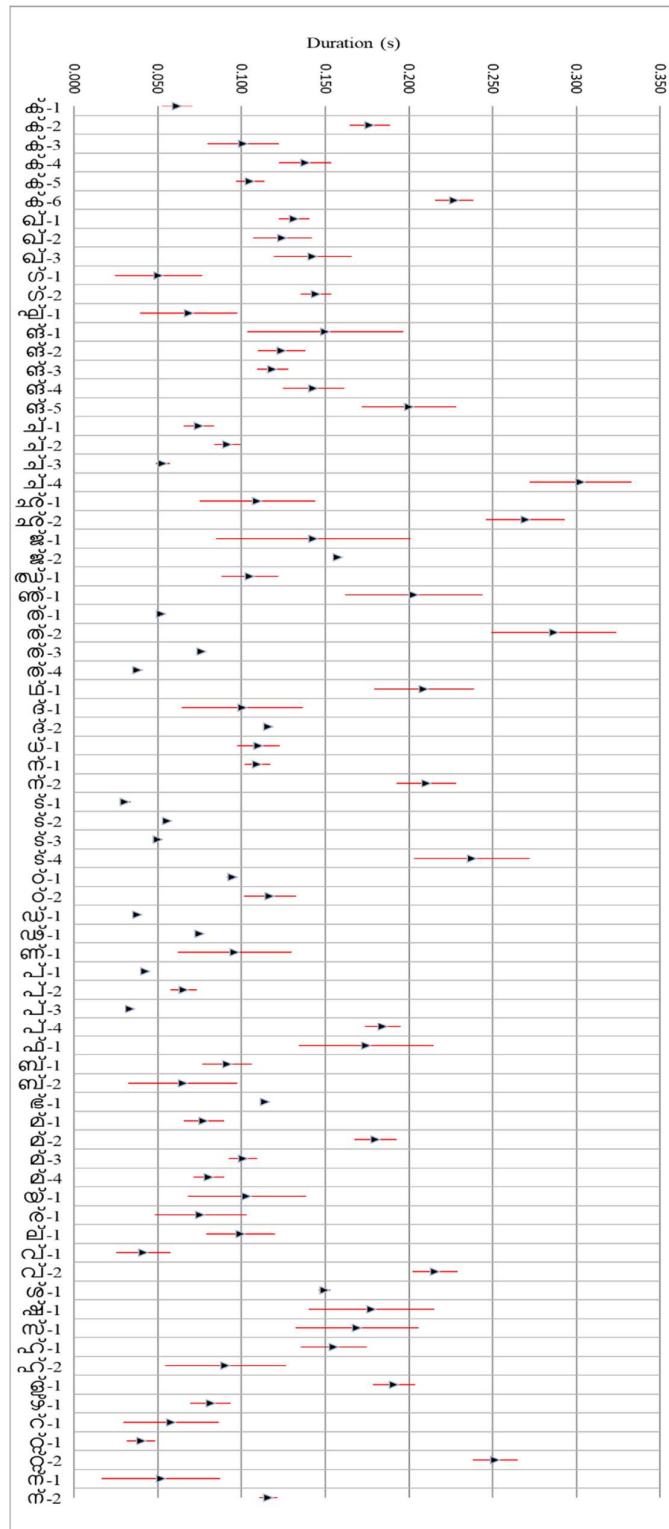


Fig. 3.31 Graphical representation of Durational Statistics of Malayalam consonant allophones from acoustic speech

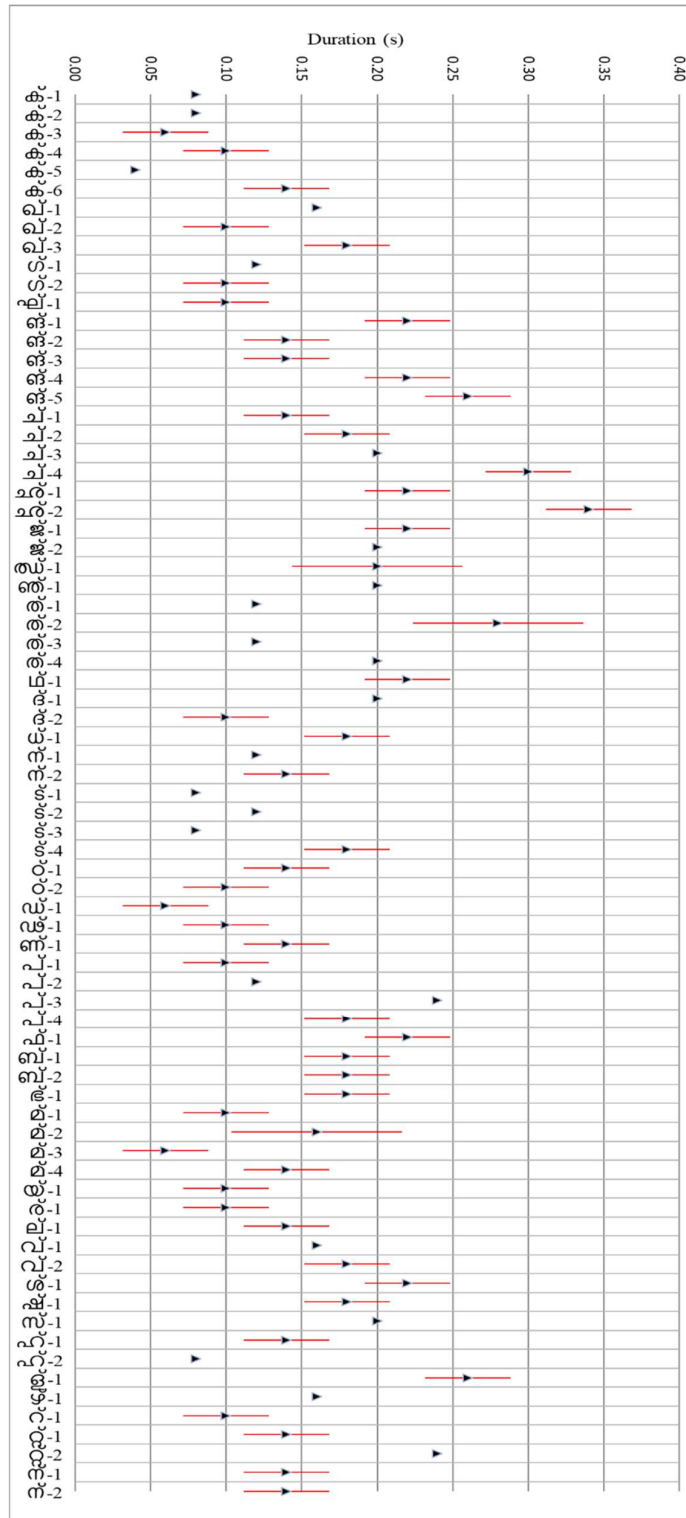


Fig. 3.32 Graphical representation of Durational Statistics of Malayalam consonant allophones from visual speech

From the durational analysis, the time duration from the audio and visual streams was estimated separately for phonemes and allophones. For almost all Malayalam phonemes, acoustic speech is lagging the visual speech. This lagging is prominent for vowel phonemes than consonant phonemes. However, due to the coarticulation effect, this lagging is feeble for vowel allophones and consonant allophones. Based on this durational analysis, audio-visual asynchrony is estimated for phonemes and allophones. The time difference between the visual and acoustic speech of the phoneme and allophone was treated as the audio-visual asynchrony of phonemes and allophones, respectively. For this, the durational difference between the speech modalities was measured for each phoneme and allophones for five speakers. After estimating the time lag, the histogram of asynchrony distributions is tabulated for phonemes and allophones as in fig. 3.33 and fig. 3.34, respectively. In both figures, early audio region and early video region are due to preservatory coarticulation and anticipatory coarticulation, respectively.

For phonemes, the histogram is centred in the visual lead region with few distributions in the audio lead region varying from -20 ms to +295 ms as in fig. 3.33. For allophones (fig. 3.34), the histogram is centred near the boundary between the synchronous and visual lead region with eye-catching distribution in the audio lead region compared to Fig. 3.33. The effect of coarticulation is reflected by extending into the early audio range of the histogram, ranging from -115 ms to 270 ms. From the analysis of this database, it is observed that anticipatory coarticulation is prominent and ranges from small early audio to large early video. The asynchrony mentioned above refers to the time delay between the recorded audio and visual speech data. The sources of this asynchrony may include the differences in response time of the hardware used to record the signals and the fps of the video recorder etc.

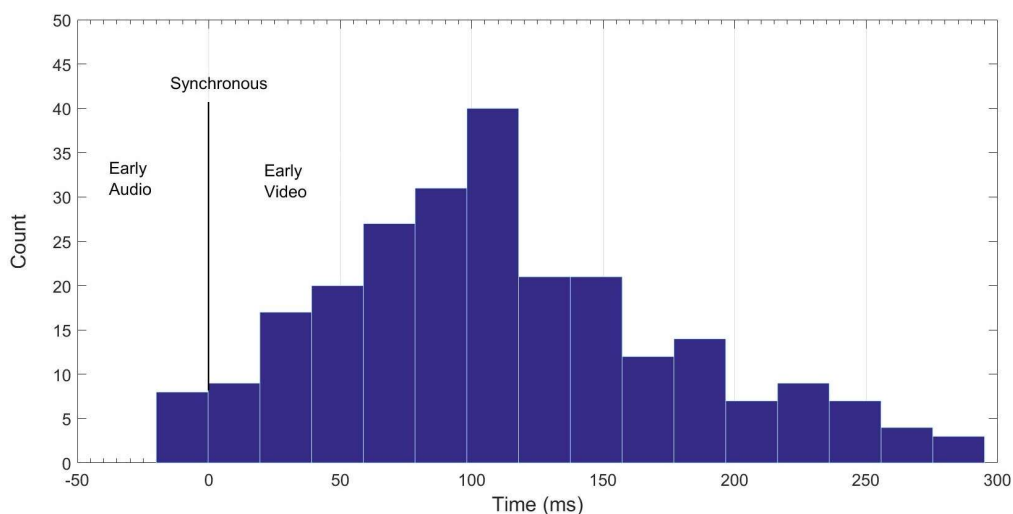


Fig. 3.33 Histogram of Asynchrony distribution in Malayalam Phonemes

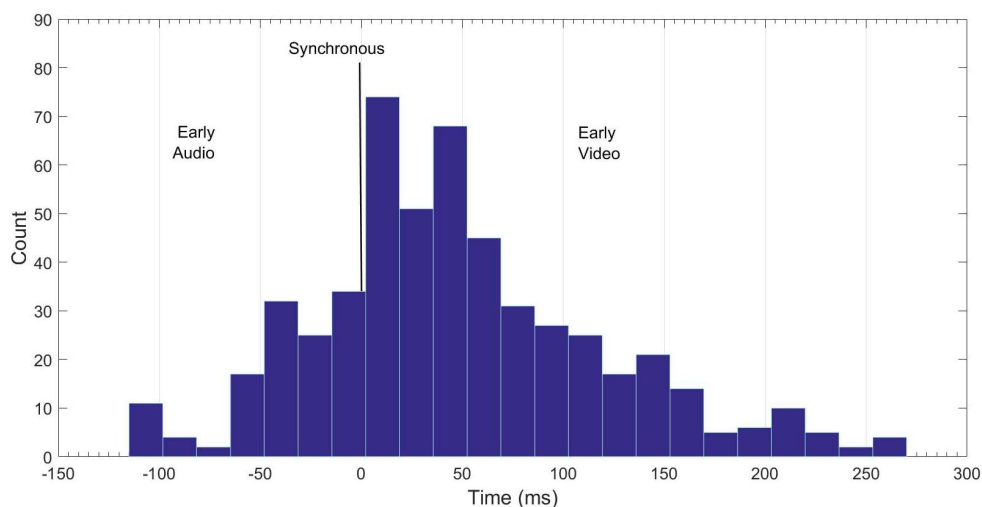


Fig. 3.34 Histogram of Asynchrony distribution in Malayalam Allophones

3.6 Conclusions

A new multimodal Malayalam audio-visual speech database is developed and presented. This database is recorded in 3 categories with unique speakers in each stage. The first category is the audio-visual speech database recorded in a controlled environment. This category includes 25 female and

five male speakers uttering 50 Malayalam isolated phonemes and 207 connected words comprising all allophonic variations five times each.

The second category consists of an audio-only speech database which has two sub-categories. The first sub-category is a clean audio speech database recorded in a closed environment uttering 50 Malayalam isolated phonemes five times each by ten male and 20 female speakers aged 21-25. The second sub-category is an audio speech database recorded in an acoustically realistic environment uttering of five Malayalam short vowel phonemes ten times each by ten males and ten females in each group ranging from 5 to 60, a total of 560 males and 560 females.

The third category contains an audio-visual speech database captured in an uncontrolled environment to help in developing a visual speech processing system in real-world ambient conditions. This category includes five female speakers uttering 50 Malayalam isolated phonemes and real-time isolated words five times each with a complex background. It includes lighting variations and multiple speakers in the background. Three additive noises (White, Pink and Red) were added to our database and is used to study its influence on speech processing applications. Durational analysis of phonemes and allophones and its visual counterpart and thereby audio-visual asynchrony has been carried out, which is the first work in the Malayalam language. The database is intended to facilitate research into the effects of age on speech, noisy speech processing, visual speech processing, viseme-based speech synthesis, and lip synchronisation utilising audio-visual speech asynchrony, among other topics.

CHAPTER 4

MALAYALAM VISEME SET IDENTIFICATION

4.1 Introduction

Speech, the most natural form of human communication, is bimodal because it combines both audio and visual signals to interpret it. Speech-based applications at public places like enquiry systems at railway stations, where clean speech is not available, is a need of today. In such situations, visual signals are necessary for decision making. For this, a bimodal speech processing technique is to be developed. However, processing the entire video signal is computationally expensive in any real-time speech processing application. Therefore, knowledge about visually separable basic units in a language (viseme) is vital in making any multimodal speech-based applications. This chapter describes an in-depth study to characterize the visual cues in Malayalam speech that can be used for any bi-modal speech application. A phoneme is the basic sound unit necessary to symbolize all words in that speech. The corresponding language unit of visual speech is termed viseme. For many years of study in visual language, it has gained considerable alterations in its definition. A viseme can be considered in terms of articulatory gestures such as mouth opening, teeth, and tongue vulnerability that must generate different phonemes [90]. An equivalent definition used extensively in literature is a viseme for a set of phonemes with a similar visual look [144]. The static viseme cannot represent the coarticulation effect of a visual address. On the other hand, the current description of viseme is a lively visual language unit that describes distinct speech movements of the visual speech articulators. The lips, tongue, teeth and jaw are the actively visible articulators used in language production [145]. Analysis of the lips, the most active visible

articulator, is crucial in the visual speech analytics framework for recognition and synthesis. Due to the appearance of the lips, many phonemes belong to a single viseme class, thereby creating a bridge between phoneme and viseme, which is termed phoneme-to-viseme mapping or many-to-one mapping. The visual equivalent of a phoneme has several features that require a comprehensive study of this phoneme-to-viseme mapping.

Lip segmentation is the prime task in a lip-reading system. However, successful lip extraction in motion, including the oral cavity, is still an unsolved problem in an actual situation. In this work, the possibility of segmenting the lip region from the speaker's face while speaking is carried out by exploiting the strength of different colour spaces in the Indian context. The low contrast between the lip and skin colour tone and facial hairs hinders capturing the lip dynamics, making it necessary to execute manual lip tracking. For developing a phoneme-to-viseme mapping, language exploration is needed since the phoneme set is language-dependent. The correlation between static visual speech unit and phoneme is carried out by developing a many-to-one phoneme-to-viseme mapping from isolated phonemes based on linguistic knowledge and a data-driven approach. The visual coarticulation effect creates a visibly different appearance for the same phoneme in a different context, creating a further subdivision in phoneme-to-viseme mapping. The contextual variation in Malayalam phonemes is modelled using allophonic characterization. The coarticulation effect in visual speech is modelled by developing many-to-many allophone to viseme mapping using a data-driven approach alone. Vowel-only and consonant only mapping is also mentioned.

The content of this work is arranged as follows. Section 4.2 discusses the substantial issues that come up while establishing a viseme set. Section 4.3 describes the various approaches utilized in the viseme set creation. Section 4.4 discusses the selection of relevant frames for the underlying phoneme. Section

4.5 analysis the strength of different colour models in the lip segmentation problem. Section 4.6 deals with the semi-automatic lip region extraction process. Section 4.7 provides a thorough description of phoneme-to-viseme mapping. Section 4.8 consolidate the visual speech analysis done and further modifications is present. Section 4.9 explains the creation of allophone-to-viseme mapping. Finally, section 4.10 concludes the work.

4.2 Challenges in Viseme set Identification

The use of visual cues in speech recognition, especially in a noisy environment, is promising. The primary requirement of the analysis of the visual speech signal is the phoneme to viseme map. The use of the viseme set shows significant improvement in noisy speech recognition.

In developing a visual speech analysing system, one must consider the problems related to speaker variability in appearance [146], pose [147], shadows and recording device [148], [149]. While developing the database (Category 1) for this study, the variations in these factors is minimised by selecting a homogeneous group of speakers and recording with the same devices in the same recording environment (Chapter 3, section 3.2). The primary task in visual speech analysis is to decode the visual information from the lips. Because the inertia of the facial muscles is greater than that of the vocal organs, the extent of lip deformation is limited. As a result, the number of visemes in a speech is always less than the number of phonemes. For developing the viseme set, the literature suggests two different approaches: linguistic and data-driven approach [104], [150]. In the linguistic approach, phonemes with the similar visual appearance of active articulators treat as a viseme. In the data-driven strategy, the visual speech is analysed by extracting essential features from the lip area and group the features based on the similarity measurement. The linguistic approach depends significantly on the perception ability of the linguistically trained individual, which accurately

represents the human lip-reading nature and is a time-consuming process. The data-driven approach provides a less time-consuming endeavour, but computational analysis of visual cues profoundly depends on the choice of visual characteristics, which may be language-dependent. In human understanding, a holistic perspective is much more important than parts (Gestalt perception theory), but in computer vision, parts (pixels) is much more significant than the whole (picture) [151]. Because of this battle, the data-driven approach alone cannot mimic human perception accurately, and linguistic knowledge alone cannot have the ability to examine substantial visual speech information. Therefore, a linguistic involved data-driven approach can make valuable of individual perception modelling from linguistic approach and the computational easiness out of a data-driven approach.

This work aims to identify the viseme set in Malayalam language using a linguistic involved data-driven approach by consciously neglecting the issues in visual speech as discussed above. The lack of a phonetically rich database in the target language is one of the significant challenges in visual speech analysis. Besides, there is no collective agreement among researchers in different aspects of the database like what is to be recorded, how many speakers are to be included, which types of speakers are to be included to generalize the whole population of interest, etc. As an initial work in Malayalam visual speech analysis area, an audio-visual speech database is created, which contains 30 native speakers (out of 30, 23 speakers are used) of Kerala by capturing the lip region of the speaker's face along with the audio. The database includes vowel and consonant-vowel syllables, which comprises 50 phonemes and 106 words that capture all contextual variations of 50 phonemes, namely the allophone, as in tables 3.3 and 3.4.

The crucial step involved in the visual speech analysis is identifying relevant static frames from the recorded video, which contains the visual

appearance of the concerned phoneme. Based on the linguistic mapping, frames with similar visual appearance were selected manually for each speaker, minimizing the error rate during data-driven analysis. The main factor in the data-driven approach is selecting relevant visual features to model human perception. It can be solved only by considering the visual speech properties of the concerned language. The tongue plays the significant role in speaking Malayalam in terms of speed and flexibility, making it distinct from other languages. The degree of presence of teeth, oral cavity, and lip shape can be modelled using geometric features of lips and deformation in the appearance of lips and tongue can be modelled through the Discrete Cosine Transform (DCT) feature. Thus, the mathematical analysis is initiated by extracting the Geometric features and DCT features from the selected frames.

The optimum number of static frames that represent the visual equivalent is a crucial factor to be studied. Since the visual length and appearance of vowel phonemes and consonant phonemes are significantly different, static frame alone may end in sparse classification from the computational point of view. Besides, it is not advisable to consider every frame in a phoneme to represent the visual speech due to computational complexity and non-uniformity in the feature-length of vowel and consonant phonemes. The time evolution of the static frame is used to solve this problem. Thus, a viseme is represented in three ways: static frame, static frame with preceding and following frames (3 frames), and static frame with two preceding and following frames (5 frames) as in fig. 4.1.

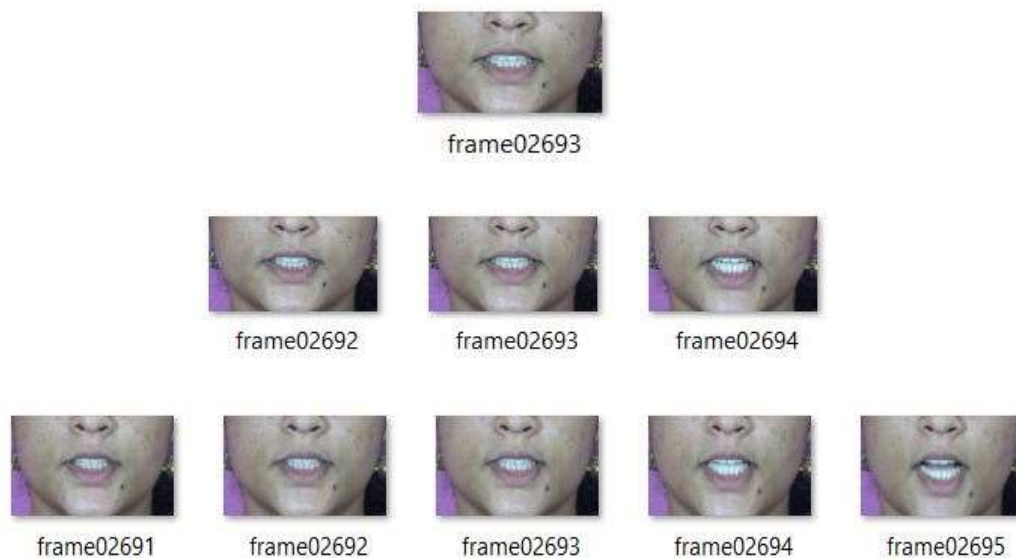


Fig. 4.1 Viseme representation using a single frame (upper), three frames (middle) and five frames (lower)

The next step is to identify the number of visemes from the visual features. Researchers conducted various techniques to establish the mapping between phoneme and viseme. Still, there are no reliable and unambiguous methods to confirm that one is better than the other. This work categorized the visual feature vectors into viseme groups by combining the K-means clustering [152] and Gap statistic methods [153]. Since K-means requires a pre-determined cluster number, the gap statistic method identifies the optimal cluster number from a range of clusters by exploring the language knowledge. The correlation between static visual speech unit and phoneme is carried out by developing a many-to-one phoneme-to-viseme mapping from isolated phonemes based on linguistic knowledge and clustering in the parametric space using different visual features. Three viseme mappings (static frame alone, three frames, and five frames) were compared with the linguistic mapping and visual speech duration, thereby identifying the best representation of visual speech in terms of frames. The coarticulation effect in visual speech is

accounted for by developing many-to-many allophone to viseme mapping using a data-driven approach alone.

A relevant methodology is presented for phoneme-to-viseme mapping. This methodology has the advantage of considering linguistic knowledge, an essential element in creating a speech-based application in the concerned language. A linguistically involved data-driven approach can make an individual perception model from a linguistic approach with computational easiness as it employs the data-driven approach. This is the first study that utilizes Gap statistics to estimate optimum cluster number from highly correlated visual speech data.

4.3 Viseme set Formation Approaches

Many researchers have analysed the importance of phoneme to viseme mapping—the phonemes which have almost the same visual mouth appearance grouped to a single viseme class. Many mappings have been recorded in the literature, with the number of visemes in a language ranging from 10 to 20 [108]. The number and nature of visemes are language-dependent. Hence a language-specific exploration is needed for establishing the viseme set for a language. Traditionally there are three approaches for obtaining viseme from a many to one mapping: linguistic knowledge-based [98], [154], perception experiments with human subjects [90], [93], [155] and data-driven approach [101],[21], [99]. Some authors blend the linguistic knowledge-based approach and the perception experiments-based approach. This approach is termed subjective assessment. In the subjective approach, viseme classes are defined through linguistic knowledge and prediction of phonemes having a similar visual appearance. A viseme class created by clustering phonemes based on features extracted from the mouth region highlights a data-driven approach. Most of the work in visual speech has been reported in European languages. Few works reported in Indian languages, such as Hindi [46], [107], [133] and

Marathi [111], have been studied in viseme mapping. Attempts to identify viseme sets have not yet reached the realm of successful development of visual speech technology in the Indian languages.

Almost all viseme maps have developed either based on linguistic approach or data-driven approach, or both. Besides, only a few viseme maps have been studied for the coarticulation effects of visual speech, which might be the lack of a database containing all contextual variations in the language concerned. This research work will be the initial study in Malayalam phoneme to viseme mapping based on a linguistic involved data-driven approach. The contextual variation of phonemes is also studied by developing an allophone-to-viseme mapping using a data-driven approach alone, which is explained in section 4.9. Before addressing this, relevant frames which represent the viseme are to be selected.

4.4 Selection of Relevant Frames

A high-quality visual speech is recorded from the speaker's mouth region with a resolution of 1280 x 720, having a frame rate of 25 fps in MP4 format (as in section 3.2.2). After documenting the multimodal speech database, the visual speech mode alone is used for viseme mapping. Considering every frame in a phoneme to represent the visual speech unit is not advisable. It leads to computational complexity and non-uniformity in the feature-length of vowel and consonant phonemes. Furthermore, selecting a single frame alone cannot represent the time evolution, which is essential to deal with the coarticulation effect. Hence, the number of frames required to represent the visual equivalent is selected, giving optimum results.

The crucial step in the mapping procedure is identifying relevant frames from the recorded video, which contains the visual look of the underlying phoneme. This work is performed by rendering the expertise of the linguistic

peoples. In Malayalam, phonemes and allophones are linguistically categorized according to articulation points and manners. However, the classification of visual speech units, visemes, of Malayalam has been not reported. In this work, visual speech is classified based on linguistic knowledge. The relevant frames were then chosen based on articulatory norms and linguistic expertise to implement a data-driven method. Two preceding and following frames from the chosen frame were also selected to encode the time evolution of the underlined phoneme's visual speech. Fig. 4.2 shows the sequential arrangement of the frame, which will capture the visual dynamics of the underlying phonemes. The linguistically chosen frame is the middle one.



Fig. 4.2 Sequential frames for phoneme - ഞ് /c/

4.5 Colour-based Approach to Lip Segmentation

Lip reading enables us to understand the spoken utterance from the speaker's face by interpreting the movement and gestures of visual articulators. In Humans, the actively moving parts of the face provide important visual speech information, especially in the mouth region. Therefore, lip segmentation and tracking are the primary tasks of a lip-reading system. Even though the use of these visual cues improves the performance of the speech recognition systems in noisy environments, a successful lip tracking algorithm from visual speech is still a challenging task. While developing an algorithm for lip tracking, one must consider the lip colour tone variations compared to its surroundings, lighting conditions, head movements, etc. Because the accuracy of lip tracking is so crucial in visual speech recognition, developing such an algorithm requires extreme caution. In most of the reported works, the

reddishness of the lips is made use in lip tracking. However, since the lips of most of the Indians are not reddish and there is no significant colour tone variation between the lips and its surrounding, such an algorithm does not give an acceptable result.

There are two approaches for segmenting lips from the mouth region: model-based approach and colour-based approach [156]–[158]. Mathematical models for lip contour are used; consequently, a set of model parameters is used for lip segmentation. Deformable templates [159], [160], Active shape [122], [123] and appearance models [125], [161] are the widely used model-based approaches. In colour-based approaches, lip and skin pixels' triplet values can be used as the base information for segmentation. The colour image segmentation usually extracts the regions from an image by exploiting colour uniformity or discontinuity. The colour-based methods are classified as simple thresholding-based methods [162], histogram-based probabilistic methods [114], [163], segmentation by clustering in the colour space [112], fuzzy clustering methods [116], etc. Thresholding-based method is used for the lip segmentation process.

The focus of this section is the segmentation of lips based on colour pixel values in different colour spaces in the Indian context. A colour model is a mathematical concept to visualize the colour spectrum in a multi-dimensional space. The colour model RGB, HSV, CIE L*a*b* [164], and more can be visualized in 3D shapes while the CMYK colour model in 4D space. A colour space maps a specific, measurable and fixed range of colours and luminance values to the colour models. sRGB is one of the colour spaces which has a comparatively smaller colour gamut than Adobe RGB. It is the most widely used colour space in digital files. RGB, HSV and CIE L*a*b* colour models are visualized with sRGB colour spaces. Adobe RGB has a broader RGB colour gamut to encompass most of the colours that printers can create. The same RGB

values may look different depends on the colour space. As a result, visualizing a colour model without an associated colour space is difficult. The different colour models represent the colour information in different ways. The lips discrimination power is different for different colour models and depends on the complexion of the speaker. Even though the transformation from RGB to some colour models is computationally expensive, the lips discrimination power is enhanced significantly. The colour models HSV and CIE L*a*b* are promising for the database of this study.

4.5.1 Image Thresholding

The method of separating the region of interest from the background using the intensity values of the pixels is known as image thresholding. Thresholding is a technique for converting colour/ greyscale data into a binary image. Pixels with intensities larger than the threshold are mapped to logical true. In contrast, pixels with intensities less than or equal to the threshold are mapped to logical false, or vice versa, depending on the situation. The thresholded (binary, B) image is defined as

$$B(x, y) = \begin{cases} 1, & \text{if } I(x, y) > T \\ 0, & \text{if } I(x, y) \leq T \end{cases} \quad (4.1)$$

Where $I(x,y)$ is the input image with pixel position represented by (x,y) and T is the threshold value.

In the HSV colour model, the Hue (H) component is used in the segmentation process with thresholding condition,

Lip Pixels, If $H > 0.70$

Skin Pixels, else

In the CIE L*a*b* colour model, the a* component is used in the segmentation process with thresholding condition,







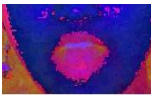













Lip Pixels, If $a^ > 3$*

Skin Pixels, else

The performance of the image thresholding method is found to be satisfactory for speakers with fair complexion and no facial hairs. However, the lip discriminating power for speakers with facial hairs or dark complexion is poor. Table 4.1 shows the performance of two colour models in the lip segmentation process with different facial features. It is found from the study that the Hue component in the HSV colour model representation is an efficient feature for the teeth information extraction of all speakers and the outer lip contour extraction of female speakers. The a^* component of the CIE $L^*a^*b^*$ colour model can be used to get the outer lips contour of speakers with fair complexion. Even though different colour space representation fails to extract the lip contour with enough accuracy, it provides vital information for teeth area computation as in section 4.7.2.1 and to isolate the lip region from its background as in section 4.8.

Researchers used a model-based technique like AAM (Active Appearance Model) and ASM (Active Shape Model) for lip contour extraction. However, the construction of these models requires an extensive training set to cover the high variability range of lips and is often a challenging task. Therefore, another possibility is the manual marking of the lip contour. Since the primary concern is to get exact information about the lip counter, the manual lip marking method is adopted. Even though this method is time-consuming, it ensures the accuracy of lip contour by using knowledge about the articulators' postures.

Table 4.1 Performance HSV and CIE L*a*b* Colour Models in Lip Segmentation Problem

Image of the speaker in RGB	Facial Hair	Complexion	Image of the speaker in HSV	Thresholded Image	Image of the speaker in CIELAB	Thresholded Image
	No	Fair				
	No	Dark				
	Yes	Fair				
	Yes	Dark				

4.6 Lip Region Extraction

The lip region is the only significant area in the representation of visual speech. Therefore, the region of interest (lip region) is to be segmented from the rest of the chosen frame, which contains irrelevant information like background screen, hairs, ornaments, *etc.*

The primary task in the lip region extraction is selecting the optimum number of landmark points that best represent the lip contour. The adjacent landmark points were joined by straight lines to form the lip contour. The number of landmark points is chosen by trial and error that they best represent the lip contour. It is found that a minimum of 36 (20 for outer lip contour and 16 inner lip contour) landmark points is required to represent the lip contour (as in fig. 4.3) without any significant deviation from the actual one. Since these points are equidistant, the contour is quasi smooth. Each landmark point is represented by the x and y-coordinates of the pixel position.

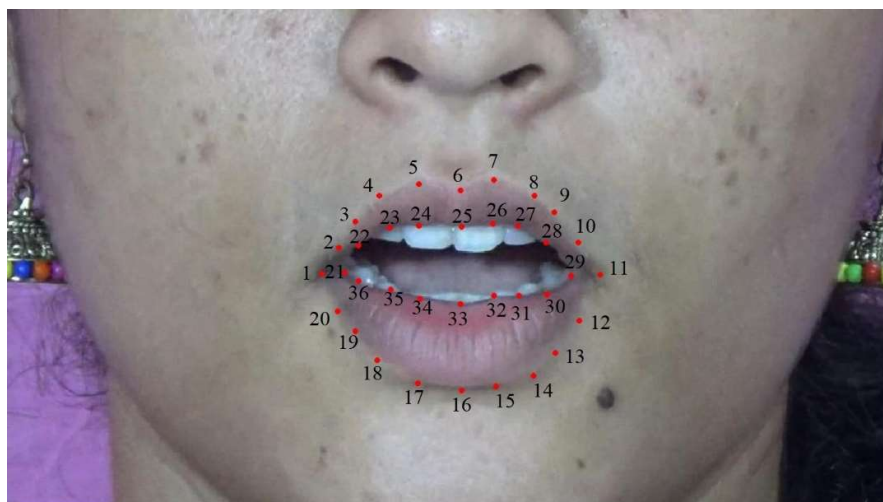


Fig. 4.3 Manual labeling of landmark points in ROI

A slight tilt in the clockwise or anti-clockwise direction may be present in the selected frame. To make the selected frame rotational invariant, rotate it in the opposite direction of the tilt. The angle θ made by the line joining the two landmark points at lip corners (1 and 11 as in fig. 4.4) with the horizontal is estimated using Eq. (4.2). The horizontal line passes through the pixel coordinates (1,360) as (a_x, a_y) and (1280,360) as (b_x, b_y) . Then the frame is rotated by an angle “ $-\theta$ ” through the midpoint of the line joining the two landmarks as in fig. 4.5. The rotated image may not have the same size as the original image.

$$\theta = \left\{ \tan^{-1} \left(\frac{b_y - a_y}{b_x - a_x} \right) - \tan^{-1} \left(\frac{y_{11} - y_1}{x_{11} - x_1} \right) \right\} * \frac{180}{\pi} \quad (4.2)$$

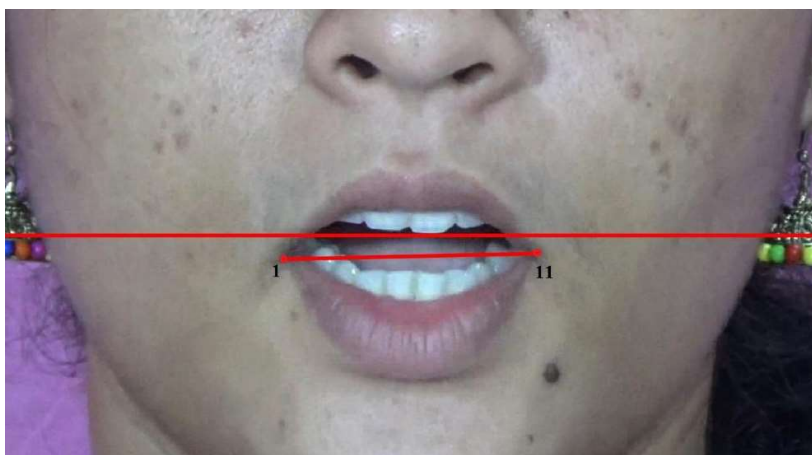


Fig. 4.4 Estimation of Angle of Tilt

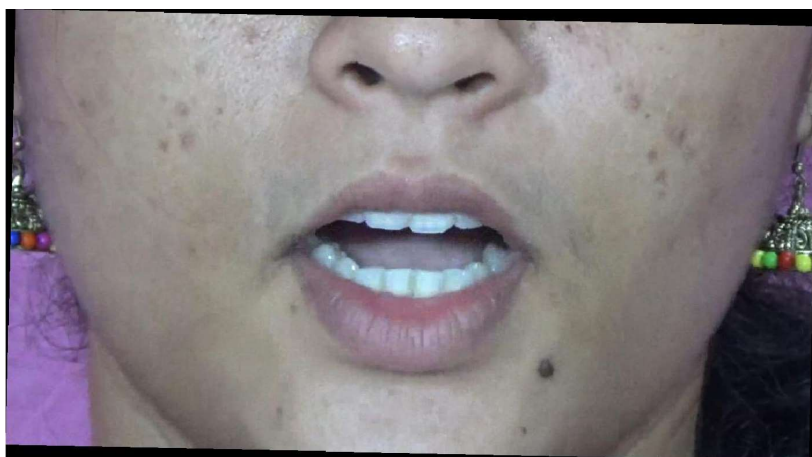


Fig. 4.5 Rotated Image

After rotating the image, the landmark points must be rotated to place them on the rotated lip region. Before rotating the points, the coordinates of the landmark points (x,y) is subtracted from the original image centre $(x_{original}, y_{original})$ so that the image centre is translated to the origin $(0,0)$. Thus, rotating a point $(x-x_{original}, y-y_{original})$ on a plane about the origin by θ degrees counter-clockwise is given by Eq. 4.3.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \text{Cos}\theta & -\text{Sin}\theta \\ \text{Sin}\theta & \text{Cos}\theta \end{bmatrix} \begin{bmatrix} x - x_{original} \\ y - y_{original} \end{bmatrix} \quad (4.3)$$

To overlay the rotated landmark points in the rotated image, the rotated coordinate points (x',y') must be modified by adding with the centre position of the rotated image $(x_{rotated}, y_{rotated})$ as (x'_{new},y'_{new}) . The rotated image along with rotated landmark points is shown in fig. 4.6. Therefore Eq. 4.3 can be modified as,

$$\begin{bmatrix} x'_{new} \\ y'_{new} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x - x_{original} \\ y - y_{original} \end{bmatrix} + \begin{bmatrix} x_{rotated} \\ y_{rotated} \end{bmatrix} \quad (4.4)$$

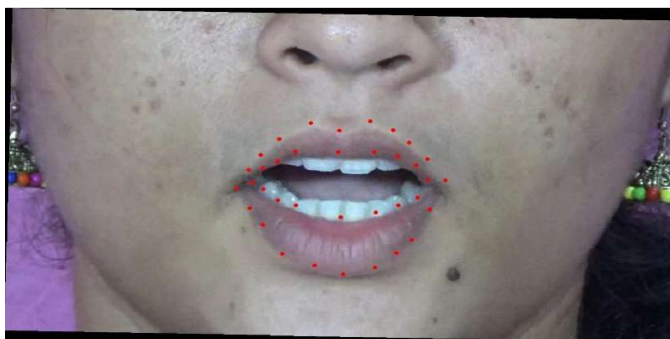


Fig. 4.6 Rotated Image along with landmark Points

From the rotation-invariant image, the translation-invariant lip region must be extracted. The area occupied by the lip contour is different for different phonemes. For example, see the visual appearance of the phoneme $\text{എ} /e/$ and $\text{ഉ} /u/$ as in fig. 4.7 and fig. 4.8, respectively. Hence, the size of the manually segmented region of interest may be different for different phonemes. So, normalization is required to make the region of interest phoneme independent. Manual segmentation will not be able to accomplish this process.



Fig. 4.7 Selected Frame of the Phoneme $\text{എ} /e/$



Fig. 4.8 Selected Frame of the Phoneme \underline{u} /u/

A semi-automatic method for the segmentation of ROI is employed. For this, ROI with minimum size and location must be identified. The minimum length and breadth of the rectangular frame that occupies the ROI for all phonemes and speakers is manually examined. It is found that a minimum length of 600 pixels and a breadth of 500 pixels. The location of the rectangular frame is selected such that its centre coincides with the centroid of the lip contour. The x and y coordinates ($X_{centroid}$ and $Y_{centroid}$) of the centroid of lip contour is estimated from x and y coordinates (x_i and y_i) of the landmark points using the equation Eq. (4.5) and Eq. (4.6) for the centroid of a finite set of points. Figure 4.9 shows the rotation and translation invariant lip region.

$$X_{centroid} = \frac{1}{36} \sum_{i=1}^{36} x_i \quad (4.5)$$

$$Y_{centroid} = \frac{1}{36} \sum_{i=1}^{36} y_i \quad (4.6)$$



Fig. 4.9 Extracted Lip Region

4.7 Malayalam Phoneme-to-Viseme / Many-to-One Mapping

In this work, linguistic and data-driven approaches were adopted for discovering the viseme set. The linguistic approach is carried out under the guidance of linguistic persons by linguistically analyzing the utterance style of the speakers. In the data-driven approach, the mathematical representation of visual speech is extracted and clustered based on the similarity measurement. Literature [165]–[167] shows a wide range of visual features based on the type of information embedded in it, which are shape-based, appearance-based and model-based. Shape-based features explicitly simulate the mouth measurement concerning height, weight, area, perimeter etc., by analyzing the pixels in the lip boundary. Geometric features, Centroid distance and Fourier descriptor [168] belong to this category. The appearance-based feature considers all pixels in the region of interest (ROI) are informative to represent the speech. The ROI is transformed into a different domain in this method, thereby capturing the most informative components. Discrete Cosine Transform (DCT) [169], [170], Discrete Wavelet Transform (DWT) [171], [172], Principal Component Analysis (PCA) [173], [174] and Linear Discriminant Analysis (LDA) [175] and the combinations [176] belong to this category. Model-based features create a mathematical model to extract visual information with high computational complexity. Active Shape Model (ASM) and Active Appearance Model (AAM) [157], [177] belongs to this class. Due to the diversity in the lip movement of the speaker's in the world, just language mining can figure out this matter. Due to the diversity in the lip movement of the speaker's in the world, only language exploration can solve this issue. The tongue plays a vital role concerning flexibility and speed for uttering Malayalam sounds, making it distinct from other languages. The geometric features of lips may be used to model the amount of teeth present, the oral cavity and lip shape, and the Discrete Cosine Transform (DCT) feature can be used to describe deformation in the appearance of lips and tongue. The visual speech attributes are then clustered to identify the visual equivalent of the phoneme. Clustering is a vital step in data mining to discover the hidden pattern of an

unlabelled dataset based on mathematical measurement. This method divides the dataset into smaller subclasses that have high intra-class similarity and low inter-class similarity. The two widely used clustering algorithms are Hierarchical (Agglomerative & Divisive) [178], [179] and Partitional (K-means) [152], [180]. They explore the partition of data objects based on the number of clusters. However, the number of groups obtained from such approaches is highly sensitive to the nature of the dataset. Thus, identifying the optimal number of clusters is a significant endeavour and can be carried out using the Gap statistic method [153], [181]. K-means clustering, and the Gap statistic method are employed for optimum cluster selection in this work to cluster large and highly correlated datasets. Literature shows that this is the first work that employs the gap statistic method to develop the viseme map.

4.7.1. Linguistic approach

In Malayalam, the language component in the audio speech, i.e., phoneme, is linguistically categorized according to articulation points and manners as in Chapter 3 (section 3.2.1). The visual speech appearance depends primarily on the lip and lower jaw movements. Visibility of teeth and tongue is also a vital element. While uttering vowel phonemes, the lips are either wide open or projected outward. However, consonant phonemes are produced by touching the active articulator tongue at different places inside the mouth area whose dynamics is not visible. Thus, the extent of appearance of active articulators is the discriminating factor to characterize the consonant phonemes. This section explores the possibilities of forming a viseme set from linguistic knowledge by rendering the expertise from the linguistic peoples.

The vowel sound is generated by the airflow from the larynx to the lips with no obstruction in the mouth region. Linguistically the five short vowel phonemes are distinguished by the tongue's position as front-high, front-mid, central-low, back-high and back-low. Since these tongue positions are partially visible, the shape information is utilized to categorize the visual equivalent of phonemes. Since the visual appearance of long vowels is the same as that of the

corresponding short vowel. It is difficult to differentiate them visually. The front-high vowel ഇ /i/ exhibits a wide horizontal opening, and the central-low vowel അ /a/ exhibit a vertical opening. Front-mid vowel എ /e/ visually placed between front-high and central-low vowel. Both back-high ഉ /u/ and back-mid ഓ /o/ vowel shows a rounded lip shape with a discriminating outward posture for the back-high vowel. The quick gliding of the tongue from one vowel to another characterizes a diphthong. The diphthong ഐ /ai/ is made from the transition from അ /a/ to ഇ /i/ and ഔ /au/ is obtained from അ /a/ to ഉ /u/. The visual characterization of vowels is taken from the selected frame, and the visual signature of the diphthongs is captured from the transition of the selected frames. Due to the diversity, two diphthongs are assigned to separate viseme classes. In short, each vowel is assigned to separate viseme classes, but monophthong short and long phonemes of the same vowel are placed in the same class. The viseme set for vowels and diphthongs in Malayalam formed from linguistic understanding is given in table 4.2.








In contrast to the vowel phoneme, the consonant sound is articulated with the complete or partial closure of the vocal tract, which is visually distinguishable only by considering the lip appearance. The most visually distinctive sound element in consonant class is bilabial sound. While uttering this sound, the lips are kept closed with a slight strain in the facial muscles. The first consonant viseme classes formed from bilabial plosives expect ഫ് /ph/. വ /va/, the only true labiodentals in the Malayalam and the Bilabial - Plosive-voiceless aspirated ഫ് /ph/ which is visually different from other Bilabial phonemes placed in the next viseme class due to the feeble presence of teeth. Viseme 10 consists of dental consonants, which have the maximum teeth visibility and some traces of tongue tip. The velar consonants and the only glottal phoneme ഹ /h/ are placed in the next viseme class since their place of articulation is the back of the tongue and has the same visual appearance. The

tongue is backward in alveolar consonants, which are less visible and grouped in viseme class 12. Retroflex consonants are produced by curling the tongue backwardly and touching the front part of the hard palate, producing the same visual appearance and assigning it as a new viseme group. Viseme 14 is linguistically characterized as palatal consonants produced by touching the tongue towards the hard palate. In brief, 50 isolated Malayalam phonemes were mapped into 14 viseme classes. The viseme set for the consonant phonemes in Malayalam formed from the linguistic background is given in table 4.3.

Table 4.2 Linguistic Classification of Vowel and Diphthong Phonemes Visually.

Viseme	Viseme Class	Phoneme with IPA	Viseme in Frame
1	Front, High – Vowel	ഇ /i/, ഞു /i:/	
2	Front, Mid – Vowel	എ /e/, ഞെ /e:/	
3	Central, Low – Vowel	അ /a/, ഞാ /a:/	
4	Back, High – Vowel	ഉ /u/, ഞു /u:/	
5	Back, Mid – Vowel	ഒ /o/, ഞൊ /o:/	
6	Diphthong 1	ഐ /ai/	
7	Diphthong 2	ഔ /au/	

Table 4.3 Linguistic Classification of Consonant Phonemes Visually

Viseme	Viseme Class	Phoneme with IPA	Viseme in Frame
8	Bilabial - Plosive-voiced and voiceless unaspirated, Nasal	പ്/p/, ബ് /b/, ഭ് /b ^h /, മ്/m/	
9	Bilabial - Plosive-voiceless aspirated And Labiodental	ഫ്/ph/, വ്/v/	
10	Dental	ത്/t/, മ്/th/, ദ്/d/, ധ്/dh/, ന്/n/	
11	Velar Glottal	ക്/k/, ഖ്/kh/, ഗ്/g/, ഘ്/gh/, ങ്/ŋ/, ഹ്/h/	
12	Alveolar	റ്/r/, ന്/n/, സ്/s/, ര്/r/, റ്/r/, ല്/l/	
13	Retroflex	ട്/t̠/, ള്/l̠h/, ഡ്/d̠/, ണ്/ɖ̠h/, ണ്/ɳ̠/, ഷ്/ʂ̠/, ഴ്/ʐ̠/, ഴ്/z̠/	
14	Palatal	ച്/c/, ച്/ch/, ജ്/j/, ഴ്/jh/, ണ്/n/, ശ്/f/, യ്/y/	

4.7.2. Data-driven Approach

In data-driven approaches, visual features are extracted from the mouth region of talking faces and viseme formed by clustering in the feature space. Both shape-based features and appearance-based features are used as visual cues in the Malayalam language. Shape-based features use information from the speaker's lip contour. The geometric feature is the shape-based feature used in this work. An appearance-based feature deals with pixel information in the ROI (Region of Interest), thereby offering high computational complexity and

is weak in capturing geometric variations compared to shape-based features. However, in a real-time application, appearance-based features show dominance over shape-based features, which have complexities related to the accurate extraction of the lip contour. Discrete Cosine Transform (DCT) is the appearance-based feature used in this work. Taking both methods together help in judging their reliability in the problem under study. K-means clustering with the Gap statistic method is used to find the viseme set by clustering in the feature space.

4.7.2.1. Geometric Visual Features

Geometric features used in this study consist of outer lip width, outer lip height, inner lip width, inner lip height, outer lip area, inner lip area and teeth area. These are extracted from the landmark points of the lip contour as discussed in section 4.6.

The outer lip width and height are taken from the difference of x-coordinate of landmark points 1 and 11 and y-coordinate of landmark points 6 and 16, respectively, as in fig. 4.10 (a). Similarly, the inner lip width and height obtained from the difference of x-coordinate of cardinal points 21 and 29 and y-coordinate of cardinal points 25 and 33, respectively, as in fig. 4.10 (b). The outer lip area is the total number of pixel points enclosed within the outer lip boundary, as in fig. 4.10 (c). The inner lip area is the oral cavity region, measured by the total number of pixel points within the inner lip boundary as in fig. 4.10 (d). The presence, absence, and area of teeth are direct indicators to distinguish many phonemes. The teeth area inside the convex hull of inner lip landmark points is computed after converting the pixels into the HSV colour space [112]. The thresholding process segments teeth pixels to the pixels inside the inner lip, as in fig. 4.10 (e). The thresholding condition is done with the Hue channel, whose intensity value ranges from 0.45 to 0.6.

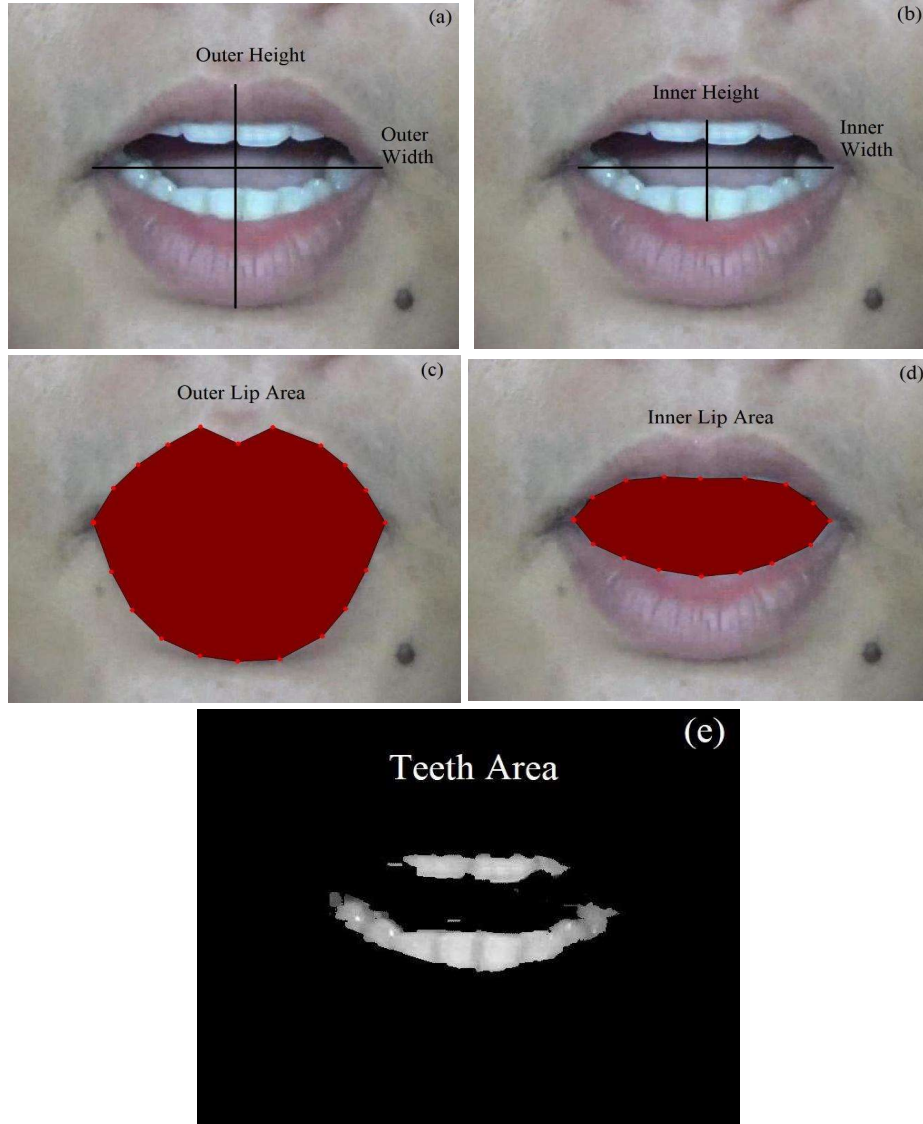


Fig. 4.10 Extraction of Seven Geometric Features from a Frame

4.7.2.2. Discrete Cosine Transform (DCT) Visual Features

DCT is one of the oldest and still popular appearance-based visual feature extraction technique in the literature. A two-dimensional DCT of an M -by- N image is represented as

$$D(i, j) = \sum_{i=1}^M \sum_{j=1}^N I(i, j) \cos\left(\frac{(2i+1)\pi i}{2M}\right) \cos\left(\frac{(2j+1)\pi j}{2N}\right) \quad (4.7)$$

Where $I(i, j)$ is the grey-scale image of the ROI. The DCT return a 2-dimensional matrix having $M*N$ coefficients. Most of the visually important information and energy is concentrated in a few coefficients of DCT, which represent the low-frequency aspect of an image. The DCT coefficient extraction process is shown in fig. 4.11. Initially, the extracted lip region of size 600 x 500 is converted into a grayscale value. The ROI is resized to 64 x 64 (as in fig. 4.11 (a)) for better implementation of the DCT algorithm, and the corresponding histogram plot is shown in fig. 4.11 (b). Histogram equalisation is used to expand the dynamical range of the intensity values, resulting in better contrast than before. Histogram equalized image and the histogram plot is shown in fig. 4.11 (c) and 4.11 (d), respectively. The ROI in the time domain is then converted to the frequency domain using Eq. 4.8. A 3-Dimensional representation of DCT coefficients and its 2-D representation in colour map is shown in fig. 4.11 (e) and 4.11 (f) respectively. To avoid the curse of dimensionality, first, 20 coefficients per frame are selected from 64 x 64 DCT coefficients in a zig-zag manner [182] starting from the DC component (D (1, 1)) as shown in fig. 4.11 (g).

4.7.2.3. Viseme set Formation by Clustering in the Parametric Space

Viseme set is formed by clustering in the feature vector space. Shape-based and appearance-based visual feature vectors are analysed separately for 50 whole phonemes and 38 consonant phonemes. The geometric feature comprises seven numerical values per frame, and DCT features comprise 20 numerical values per frame. Thus, each phoneme is represented by a 35-dimensional geometric feature vector and 100-dimensional DCT feature vector, respectively. An aggregate feature vector is created by concatenating the feature vector of 23 speakers. This feature vector was standardised for further analysis.

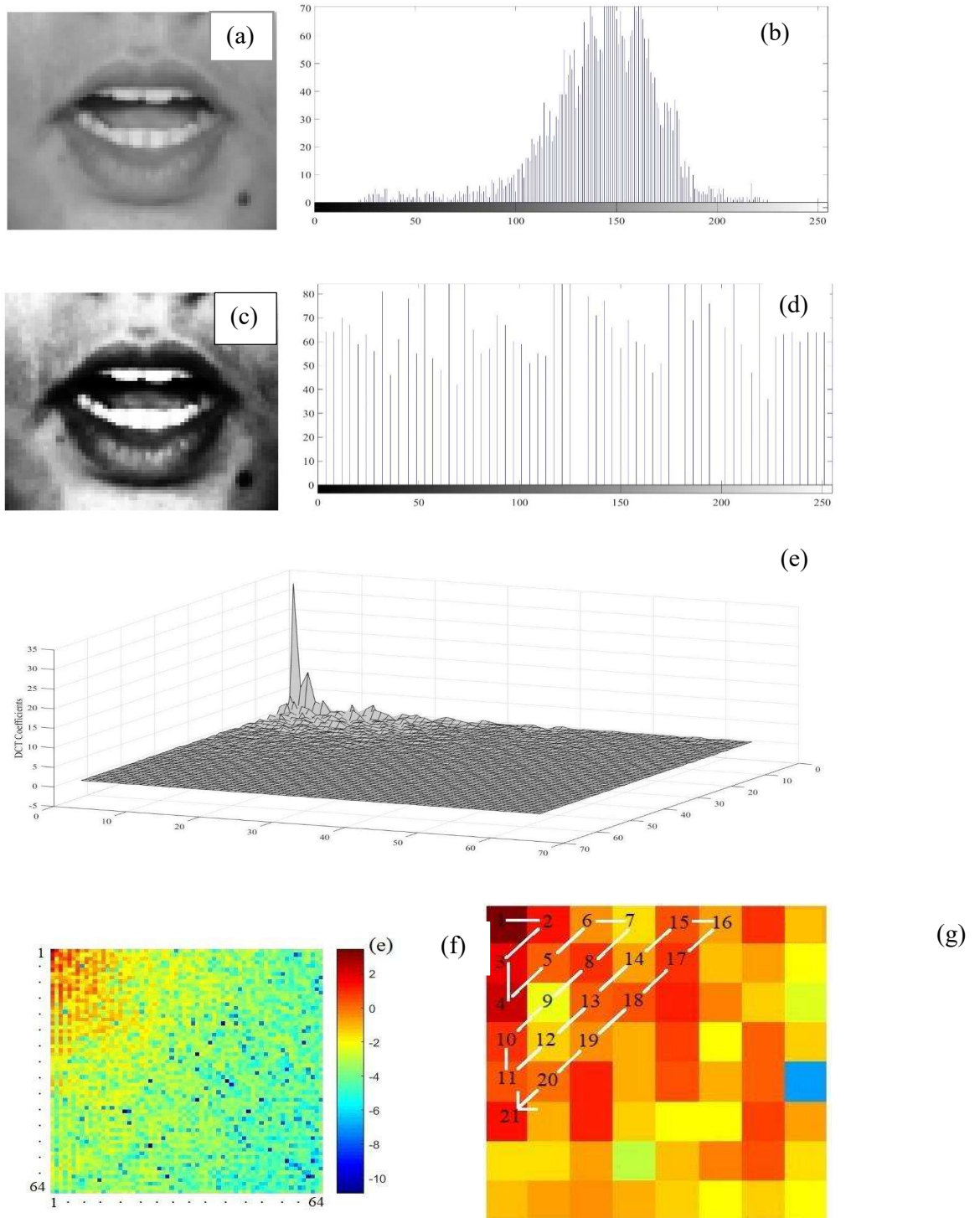


Fig. 4.11 DCT Coefficient Feature Extraction

The final feature vector is fed into the Gap statistic method for determining optimum cluster number by using the K-means algorithm for clustering purposes. K-means algorithm is one of the simplest unsupervised learning algorithms, classifying the data set into a pre-defined number of clusters based on the centroid. The algorithm inputs are the dataset containing ‘ n ’ objects and pre-defined cluster number ‘ k ’. The algorithm of K-means clustering is given below.

1. The algorithms start with initial estimates for the K centroids, either randomly generated or randomly selected from the data set.
2. Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance.
3. In this step, the centroids are recomputed by taking the mean of all data points assigned to that centroid’s cluster.
4. The algorithm iterates between steps two and three until a stopping criterion is met (i.e., no data points change clusters, the sum of the distances minimized, or some maximum number of iterations reached).

Due to the high correlation of mouth parameters for acoustically different phonemes, the clustering algorithm alone fails to estimate an optimum cluster value. The gap statistic method has proven its strength in identifying optimum cluster number in a highly correlated dataset. The gap statistic method compares the intra-class dispersion obtained from the given data with an appropriate reference distribution. The methodology of gap statistic work (using the notation from Tibshirani,2001 [181]) is as follow,

Consider a dataset $\{x_{ij}\}$ with $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, p$, consists of p features measured on n independent observations, clustered into k clusters C_1, C_2, \dots, C_k , where C_r denotes the indexes of samples in cluster r , and $n_r = |C_r|$. Let $d_{ii'}$ denotes the squared Euclidean distance between the observation i

and i' ($d_{ii'} = \sum_j (x_{ij} - x_{i'j})^2$). The sum of the pairwise distance D_r for all points in cluster r is

$$D_r = \sum_{i,i' \in C_r} d_{ii'} \quad (4.8)$$

Let W_k be the within-cluster sum of squared distances from the cluster means as

$$W_k = \sum_{r=1}^k \frac{1}{2n_r} D_r \quad (4.9)$$

W_k decreases monotonically as the number of clusters k increases. For calculation, the Gap function, Tibshirani et al.,2001 proposed to use the difference of the expected value of $\log(W_k^*)$ of an appropriate null reference and the $\log(W_k)$ of the dataset,

$$Gap_n(k) = E_n^* \log(W_k^*) - \log(W_k) \quad (4.10)$$

Where E_n^* denotes the expectation of under a sample of size n from the reference distribution. Then the proper number of clusters for the given data is the smallest k such that

$$Gap_n(k) \geq Gap_n(k+1) - s_{k+1} \quad (4.11)$$

Let s_k is the simulation error calculated from the standard deviation $sd(k)$ of B Monte Carlo replicates $\log(W_k^*)$ according to the equation $s_k = \sqrt{1 + 1/B} sd(k)$, which is represented by a vertical bar in Gap curve.

In short, a range of cluster groups is estimated using the k-means algorithm (or any other clustering algorithm), and the logarithm of within-cluster variance is compared with the exact measurement of an appropriate reference distribution of the data. The difference between these quantities (gap curve or gap function) provides the corresponding gap value for each cluster group, and these clusters show a fall at the point where the gap is maximum. Fig. 4.12 reveals different Gap statistic steps of data using K-means clustering. For a well-separated dataset, the gap function exhibits a non-monotone

behaviour. However, for a highly correlated dataset, the gap function exhibits a monotone behaviour, which directly educates us to inspect the whole gap curve instead of simply finding the cluster number with the maximum gap. The performance of the Gap statistic profoundly depends on the nature of the dataset/feature vector used. By considering the problem under study, features should be selected to display a high discriminating power between high correlated observables. This work removes the redundancy of selected features by considering a few feature coefficients that may deal with the issue under study. The optimum number of features needed to deal with the issue under study seriously remains an open research field.

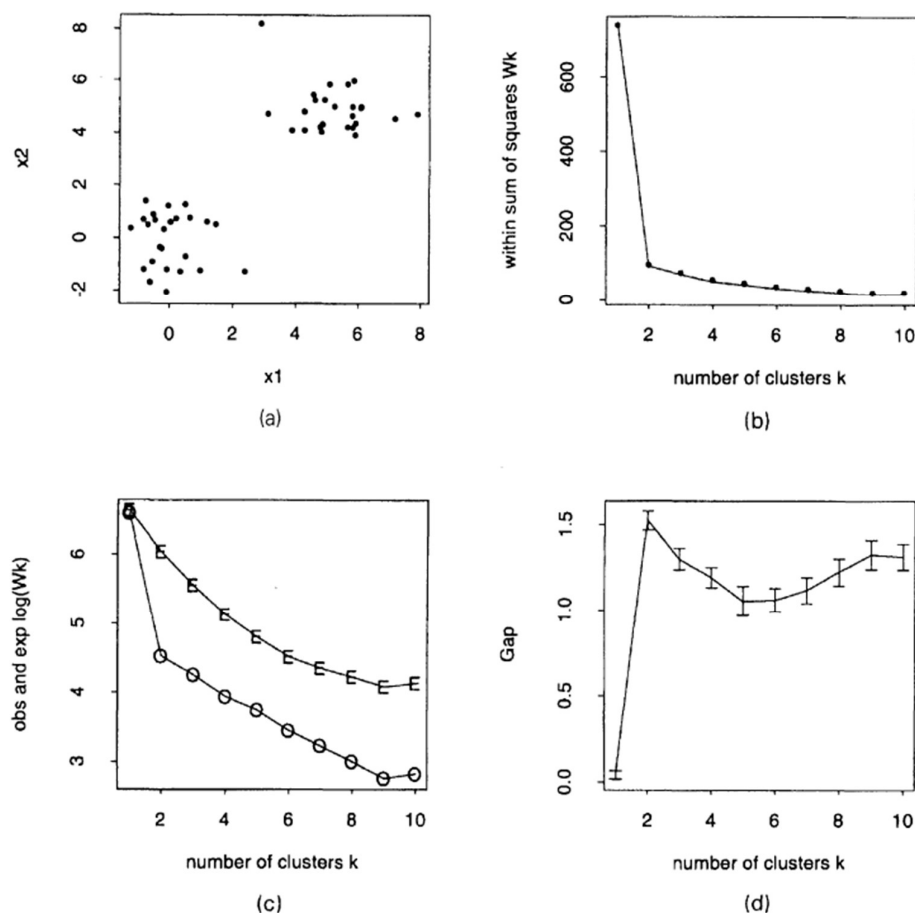


Fig. 4.12 A two cluster example: (a) data; (b) within sum of square function W_k ; (c) functions $\log(W_k)$ (O) and $E_n^*\{\log(W_k)\}$ (E); (d) gap curve (Tibshirani et al., 2001)

Depending upon the property of the dataset and underlying problem, it is better to study a reasonable range in the gap curve. Since analysing the whole gap curve is tiresome and time-consuming in estimating the optimum cluster number, especially for highly correlated data. For a highly correlated dataset and phoneme-to-viseme conversion problem, it is better to examine the gap curve between clusters 10 to 20. Though the amount of viseme is language-dependent, the majority of the published works have underlined this range in various languages. In this work, 50 Malayalam phonemes are linguistically mapped to 14 viseme classes. Besides, for a straightforward interpretation, a minimum of 2 phonemes can occupy a single viseme class due to high correlation in the visual appearance of phonemes, thereby creating a maximum of 25 viseme sets. The same methodology is carried out in the rest of this work. The gap curve is analysed for clustering the feature vectors of 50 phonemes, and the optimum cluster number is identified with maximum gap value in the cluster range 10 to 20, as shown in fig. 4.13. For 50 Malayalam phonemes, based on geometric feature vector and DCT feature vector using static frame alone is shown in table 4.4 and 4.5 respectively. The estimated viseme set using geometric and DCT features is 16.

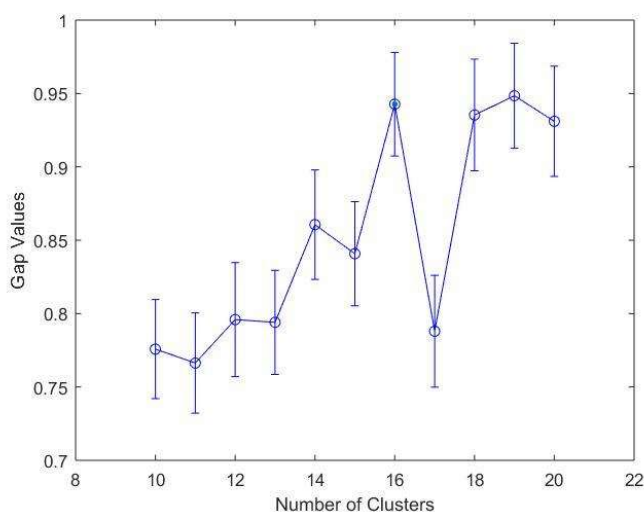


Fig. 4.13 Gap curve

Table 4.4. Phoneme-to-viseme mapping using one frame based on the Geometric features

Viseme	Phonemes with IPA
1	ച്/-c/, ചർ/-c ^h /, ജ്/-j/, ത്വ്/-j ^h /, ത്ത്/-j ⁿ /, ത്/-t/, മ്/-t ^h /, ദ്/-d/, യ്/-d ^h /, ന്/-n/, ശ്/-ʃ/, ഷ്/-ʃ/, സ്/-s/, ള്/-ɻ/
2	ബ്/-b/, മ്/-m/
3	ഹ്/-h/, ള്/-l/
4	യ്/-y/, ര്/-r/, ല്/-l/, ഴ്/-z/, റ്/-r/, ന്/-n/
5	അ/-a/, ആ/-a:/
6	ഉ/-u/, ഊ/-u:/, ഒ/-o/, ഓ/-o:/
7	ഭ്/-b ^h /
8	പ്/-p/
9	എ/-e/, ഐ/-e:/
10	ഫ്/-p ^h /, വ്/-v/
11	ഇ/-i/, ഊ/-i:/
12	ഠ്/-t ^h /, ഡ്/-d ^h /
13	ഔ/-au/
14	ഐ/-ai/
15	ട്/-t/, ഡ്/-d/, ണ്/-n/
16	ക്/-k/, ക്/-k ^h /, ഗ്/-g/, ഘ്/-g ^h /, ണ്/-ŋ/

Table 4.5. Phoneme-to-viseme mapping using one frame based on the DCT features

Viseme	Phonemes with IPA
1	ഉ-/u/, ഊ-/u:/
2	ല്/-l/
3	ഒ-/o/, ഓ-/o:/
4	ച് -/c/, ച്ച് -/c ^h /, ജ് -/j/, ത് -/t ^h /, ത്ത -/t/
5	ഔ-/au/
6	ക് -/k/, ക്ക് -/k ^h /, ഗ് -/g/, ഘ് -/g ^h /, ങ് -/ŋ/, ന് -/n/
7	ള് -/l/, റ് -/r/
8	എ -/e/, ഏ -/e:/, ഐ -/ai/
9	റ് -/r/
10	ത് -/t/, ത് -/t ^h /, ദ് -/d/, ധ് -/d ^h /, ന് -/n/, റ് -/r ^h /, ള് -/ɖ/, ഴ് -/ɖ ^h /, റ് -/p ^h /, വ് -/v/
11	ട് -/t/, യ് -/y/, റ് -/r/, ഴ് -/z/, ന് -/n/
12	പ് -/p/, ബ് -/b/, ഭ് -/b ^h /, മ് -/m/
13	സ് -/s/
14	അ -/a/, ഓ -/a:/, ഹ് -/h/
15	ശ് -/ʃ/, ഷ് -/ʃ/
16	ഇ -/i/, ഊ -/i:/

One of the essential inferences from table 4.4 and 4.5 is that the visual appearance of the vowel phoneme is almost embedded in a single frame. In addition, the velar consonant phonemes (ക് -/k/, ക്ക് -/k^h/, ഗ് -/g/, ഘ് -/g^h/, ങ് -/ŋ/) were distinguished from other consonant groups, just like in the linguistic picture. All other consonant groups were randomly distributed while analyzing viseme using a single frame. The viseme map using three frames is shown in Tables 4.6 and 4.7.

Table 4.6. Phoneme-to-viseme mapping using three frames based on the Geometric features

Viseme	Phonemes with IPA
1	ക്-/k/, ഖ്-/k ^h /, ഗ്-/g/, ഘ്-/g ^h /, ങ്-/ŋ/
2	ഓ-/o:/
3	പ്-/p/, ബ്-/b/, ഭ്-/b ^h /, മ്-/m/
4	ട്-/t/, റ്-/t ^h /, ഡ്-/d/, ള്-/d ^h /, ണ്-/ɳ/
5	ഉ-/u/, ഊ-/u:/
6	ഫ്-/p ^h /
7	യ്-/y/, ര്-/r/, ല്-/l/, ഹ്-/h/, ള്-/l/, ഴ്-/z/, റ്-/r/, ന്-/n/
8	ച്-/c/, ണ്-/ɲ/
9	എ-/e/, ഏ-/e:/, ഐ-/ai/
10	ഔ-/au/
11	ഇ-/i/, ഊ-/i:/, ത്-/t/, മ്-/t ^h /, ദ്-/d/, ധ്-/d ^h /, ന്-/ɳ/
12	ശ്-/ʃ/, ഷ്-/ʃ/, സ്-/s/, റ്-/r/
13	വ്-/v/
14	ഛ്-/c ^h /, ജ്-/j/, ത്വ്-/t ^h /
15	അ-/a/, ആ-/a:/
16	ഒ-/o/

Table 4.7. Phoneme-to-viseme mapping using three frames based on the DCT features

Viseme	Phonemes with IPA
1	പ്/-p/, ബ്/-b/, ഭ്/-b ^h /, മ്/-m/
2	എ/-e/, ഏ/-e:/
3	ഐ/-ai/
4	ഇ/-i/, ഊ/-i:/
5	ശ്/-ʃ/, ഷ്/-ʃ/, സ്/-s/, റ്/-r/
6	ച്/-c ^h /, ജ്/-j/, ത്/-t ^h /
7	ഒ/-o/, ഉ/-u/, ഊ/-u:/
8	അ/-a/, ആ/-a:/, ഹ്/-h/
9	ഫ്/-p ^h /, വ്/-v/
10	ച്/-c/, ണ്/-ɲ/
11	ര്/-r/, ല്/-l/, ള്/-l/, റ്/-r/, ഴ്/-z/, ന്/-n/
12	ട്/-t/, റ്/-t ^h /, ഡ്/-d/, ഡ്/-d ^h /, ണ്/-ɲ/ ത്/-t/, മ്/-t ^h /, ദ്/-d/, ധ്/-d ^h /, ന്/-ɲ /
13	ഓ/-o:/, ഔ/-au/
14	യ്/-y/
15	ക്/-k/, ക്/-k ^h /, ഗ്/-g/, ഘ്/-g ^h /, ണ്/-ɲ/

While taking the preceding and following frames majority of the consonant phonemes were exactly matches with table 4.3. Along with the velar consonant phonemes, bilabial consonant phonemes (പ്/-p/, ബ്/-b/, ഭ്/-b^h/, മ്/-m/) and labiodental consonant phoneme (ഫ്/-p^h/, വ്/-v/) were separated into distinct viseme classes in both features. While comparing tables 4.5 and 4.7, the DCT feature vector has shown a close resemblance to the linguistic map than the geometric features. Viseme representation using five frames is shown in tables 4.8 and 4.9.

Table 4.8. Phoneme-to-viseme mapping using 5 frames based on the Geometric features

Viseme	Phonemes with IPA
1	റു- <i>r</i> /
2	ഉ- <i>u</i> /, ഊ- <i>u:</i> /, ഒ- <i>o</i> /, ഓ- <i>o:</i> /
3	അ- <i>a</i> /, ആ- <i>a:</i> /, ഹ- <i>h</i> /
4	പ- <i>p</i> /, ബ- <i>b</i> /, ഭ- <i>b^h</i> /, മ- <i>m</i> /
5	ല- <i>l</i> /, ള- <i>l</i> /
6	ഫ- <i>p^h</i> /
7	ത- <i>t</i> /, മ- <i>t^h</i> /, ദ- <i>d</i> /, ധ- <i>d^h</i> /, ന- <i>n</i> /
8	യ- <i>y</i> /, ര- <i>r</i> /, റ്- <i>n</i> /
9	ച- <i>c</i> /, ഛ- <i>c^h</i> /, ജ- <i>ʒ</i> /, ഝ- <i>ʒ^h</i> /, ഞ- <i>ɲ</i> /
10	റു- <i>r</i> /
11	ക- <i>k</i> /, ഖ- <i>k^h</i> /, ഗ- <i>g</i> /, ഘ- <i>g^h</i> /, ങ- <i>ŋ</i> /
12	വ- <i>v</i> /
13	ശ- <i>ʃ</i> /, ഷ- <i>ʃ^h</i> /, ഡ- <i>d</i> /, റ്- <i>d^h</i> /, ണ- <i>ɳ</i> /, ഡ- <i>ʃ</i> /, ണ- <i>ʃ</i> /, സ- <i>s</i> /, ഴ- <i>z</i> /
14	ഔ- <i>au</i> /
15	ഇ- <i>i</i> /, ഊ- <i>i:</i> /, എ- <i>e</i> /, ഏ- <i>e:</i> /, ഐ- <i>ai</i> /

Table 4.9. Phoneme-to-viseme mapping using 5 frames based on the DCT features

Viseme	Phonemes with IPA
1	ഒ- <i>o</i> /, ഓ- <i>o:</i> /
2	ക- <i>k</i> /, ഖ- <i>k^h</i> /, ഗ- <i>g</i> /, ഘ- <i>g^h</i> /, ങ- <i>ŋ</i> /
3	ഉ- <i>u</i> /, ഊ- <i>u:</i> /
4	പ- <i>p</i> /, ബ- <i>b</i> /, ഭ- <i>b^h</i> /, മ- <i>m</i> /
5	അ- <i>a</i> /, ആ- <i>a:</i> /, ഹ- <i>h</i> /
6	ച- <i>c</i> /, ഛ- <i>c^h</i> /, ജ- <i>ʒ</i> /, ഝ- <i>ʒ^h</i> /, ഞ- <i>ɲ</i> /
7	എ- <i>e</i> /, ഏ- <i>e:</i> /, ഐ- <i>ai</i> /
8	ഇ- <i>i</i> /, ഊ- <i>i:</i> /
9	ശ- <i>ʃ</i> /, ഷ- <i>ʃ^h</i> /, ഡ- <i>d</i> /, റ്- <i>d^h</i> /, ണ- <i>ɳ</i> /
10	റു- <i>r</i> /, ന- <i>n</i> /
11	ഔ- <i>au</i> /
12	ശ- <i>ʃ</i> /, ണ- <i>ʃ</i> /, സ- <i>s</i> /
13	യ- <i>y</i> /, ര- <i>r</i> /, ല- <i>l</i> /
14	ത- <i>t</i> /, മ- <i>t^h</i> /, ദ- <i>d</i> /, ധ- <i>d^h</i> /, ന- <i>n</i> /
15	ള- <i>l</i> /, ഴ- <i>z</i> /, റു- <i>r</i> /
16	ഫ- <i>p^h</i> /, വ- <i>v</i> /

Based on geometric features, the diphthong vowel phonemes (ഓൗ-/au/), central vowel phonemes (അ-/a/, ആ-/a:/), bilabial consonant phonemes (പ്-/P/, ബ്-/b/, ഭ്-/bh/, മ്-/m/), labiodental consonant phonemes (ഫ്-/p^h/, വ്-/v/) and dental consonant phonemes (ത്-/t/, മ്-/t^h/, ദ്-/d/, ധ്-/d^h/, ന്-/n/) velar consonant phonemes (ക്-/k/, ച്-/k^h/, ഗ്-/g/, ഘ്-/g^h/, ങ്-/ŋ/) and most of the palatal consonant phonemes (ച്-/c/, ഛ്-/c^h/, ജ്-/j/, ഴ്-/j^h/, ഴ്-/j/) are grouped exactly in the same manner as in linguistic approach (table 4.2 and 4.3). Due to lip contour similarity between front high vowel (ഇ-/i/, ഊ-/i:/) and front-mid vowel (എ-/e/, ഏ-/e:/) and back high vowel (ഉ-/u/, ഊ-/u:/) and back mid vowel (ഒ-/o/, ഓ-/o:/) are selected to the same viseme class rather than different classes as in linguistic approach. The remaining consonant phonemes are distributed so that it follows some traces of linguistic point of consonant phoneme cluster.

Phoneme-to-viseme mapping was also studied for the DCT feature vector with an estimated optimum cluster number equal to 16. The vowel phonemes are distributed precisely in the same fashion of linguistic approach but with overlapping of a diphthong phoneme (ഐ-/ai/) into the front-mid vowel phoneme class (എ-/e/, ഏ-/e:/). Almost all consonant phonemes were precisely clustered as in the linguistic approach. The remaining consonant phonemes are randomly distributed, just like in geometric feature map.

In addition, a viseme map using seven frames is also studied; however, no noticeable variation is observed from the viseme map using five frames. This is because all vowel phonemes can be represented with a single frame, whereas consonant phonemes require nearly five frames to express their visual appearance. It may be because the visual appearance of the articulators in the

pronunciation of consonant phonemes is less than that of vowel phonemes, which primarily rely on lip shape information.

Analyzing the visual speech in terms of duration also highlights this finding. From fig. 3.28, the duration of vowel phonemes ranges between 0.50 s to 0.70 s, which corresponds to 12 and 17 frames, respectively (frame rate of the video is 25 Hz). The duration of consonant phonemes varies between 0.10 s to 0.30 s, corresponding to two and seven frames, respectively. The average duration of Malayalam vowel and consonant phonemes from the visual speech is 0.606 s and 0.200 s, corresponding to 15 and 5 frames, respectively. Hence it is better to represent the visual equivalent of a phoneme with the time evolution of a linguistically identified frame of duration 0.20 s, more precisely, preceding and proceeding two frames from the selected frame. Thus, from the viseme map and the durational analysis, a relevant frame with the preceding two and following two frames (total five frames) is necessary to model the Malayalam visual speech element, especially the visual equivalent of the consonant phonemes. The following section consolidates the work on the Malayalam visual speech analysis.

4.7.3 Phoneme-to-Viseme Mapping - An Overview

A linguistic involved data-driven approach for the viseme set identification problem. Automatic identification of visemes (frames) for corresponding phonemes from visual speech is still an open research area. From a computational point of view, a linguistic approach alone cannot have the ability to process massive visual speech information. In this work, viseme is defined as the time evolution of a linguistically selected frame for the underlying phoneme. For this, three phoneme-to-viseme mappings were developed using a linguistically selected frame, selected frame along with the previous and following frame and selected frame with two previous and two following frames. Besides, the strength of visual speech features (Geometric-

based and Image-based) was also evaluated by comparing the corresponding viseme map with the linguistic map.

Even when the viseme is represented by a single frame, as in table 4.4 and 4.5, the vowel phonemes alone have shown almost comparable categorisation. However, the classification becomes worse for viseme mapping based on geometric features, as in table 4.8. The viseme map based on DCT characteristics, on the other hand, has demonstrated its ability to categorise vowel phonemes in all three viseme formats. Hence the visual appearance of the vowel phonemes can be modelled even with a single frame based on both features. However, Geometric features drain out its strength when viseme is represented by the time evolution of the selected frame. For consonant phoneme grouping, viseme maps using five frames (Table 4.9) were almost closely resembling the linguistic map as in table 4.3. Hence the visual appearance of the consonant phoneme can be modelled only by considering the time evolution of the linguistically identified frame. In terms of features, DCT features are always dominating over the Geometric features in each viseme mapping. Besides, the complexity of extracting exact lip contour for the Geometric features (in this work) makes the image-based features (here DCT) more prior to the visual speech extraction process.

Even though DCT is effective in visual speech analysis in this study, the processes for extracting DCT coefficients essentially require landmark points along the lip contour. This section mainly focuses on the extraction of ROI without using landmark points on the lip contour. For this work, the Hue channel information in the HSV colour model is utilized further. Even though the Hue channels fails to extract the accurate lip contour, it almost isolates the lip region from its background. Histogram equalisation is used to improve the contrast of the Hue channel information, as seen in fig. 4.14. (a). The equalized image is converted into binary by applying the thresholding condition discussed

in section 4.5.1 (fig. 4.14 (b)). As in fig. 4.14 (c), small white pixel areas were converted into a black pixel to isolate the lip region from the background.

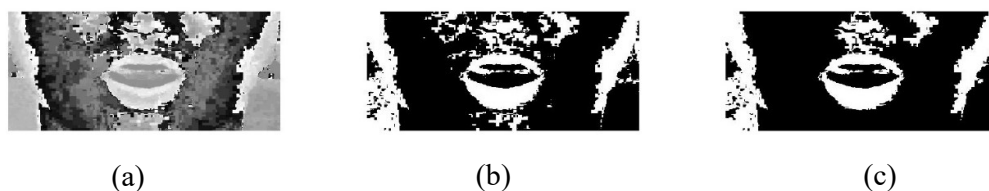


Fig. 4.14 Thresholded Lip Region

From the thresholded lip region, the centroid of all connected white pixels was estimated by summing them and divided by the number of white pixels in it. The centroid of all connected white regions (red spot) was shown in fig. 4.15.



Fig. 4.15 Centroid of Connected White Pixel Regions

Centroid points with the shortest Euclidean distance from the image's midpoint are evaluated to find the centroid spot in the lip region. Then a rectangle frame having the length of 600 pixels and breadth of 500 pixels is selected. The centre of the rectangular frame is aligned with the centroid spot in the lip contour. This rectangular frame is then embedded in the original image as in fig. 4.16.

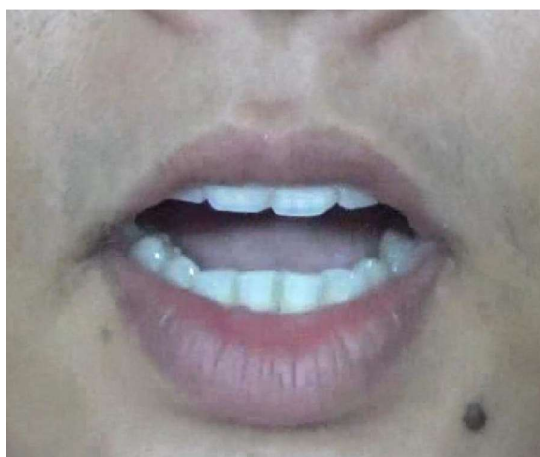


Fig. 4.16 Extracted Lip Region from HSV Colour Model

The extracted lip region in figures 4.9 and 4.16 was nearly identical. Thus, this approach bypasses the time-consuming process of extracting the lip region. However, making the extracted lip region rotation invariant, manual marking at the corners of the lip (points 1 and 11 as in fig. 4.3) is needed. Thus, the original image (fig. 4.3) and the coordinate points of the lip corners were needed to study the visual speech analysis. This method is made use for the extraction of ROI for Allophone-to-Viseme mapping.

4.8 Allophone-to-Viseme Mapping / Many-to-Many Mapping

While addressing the visual language, the phoneme-to-viseme (many-to-one mapping) must be needed to exhibit the high correlation between phonemes and the visual look. However, this mapping admits viseme as a static visual speech unit by eliminating the coarticulation effects of visual signal. The visual coarticulation effect creates a visibly different appearance for the same phoneme in a different context, creating a further subdivision in phoneme-to-viseme mapping. To put it differently, the visual appearance of a phoneme intensely depends not only on its articulation properties but also on the presence and nature of its neighbouring phoneme in the word or sentences. The contextual variation in Malayalam phonemes is modelled using allophonic

characterisation. For accurately describing the visual speech information in a different context, a comprehensive phoneme-to-viseme mapping is needed, which ought to be an allophone-to-viseme (many-to-many mapping). Only a few works were reported the coarticulation effects of visual speech by assembling the many-to-many mapping. Hilder et al., have described a novel method of segmenting the visual speech by capturing the patterns of behaviour of the articulators and clustered the behaviour that appears similar into a set of visemes, thereby obtaining a different viseme label for different allophones of a phoneme [183]. Taylor et al., have modelled the coarticulation effects of visual speech by considering the motion of visual speech articulators rather than static mouth representation and animated a talking head [145]. Mattheyses et al., have introduced a many-to-many phoneme-to-viseme mapping using tree-based and k-means clustering approaches, which ensure a more accurate description of visual speech when compared to phoneme based and many-to-one viseme based speech labels [104]. Katsaggelos et al., discussed the challenges associated with audio-visual fusion, especially the need for audio-visual synchrony since the range and directionality of coarticulation patterns differ across languages [184].

As seen in tables 3.3 and 3.4, there are 106 allophones, with 28 vowel allophones, three diphthong allophones, and 75 consonant allophones. Since the linguistic knowledge about mapping contextual variations of phonemes to viseme is still undiscovered, a data-driven approach must be used to build many-to-many visual speech mapping. Therefore, DCT visual features are captured from every allophone and clustered using the K-mean algorithm and Gap statistic as in phoneme-to-viseme mapping. For clustering the DCT features of 106 allophones, the gap curve is analysed between the range 20 and 40, and the estimated cluster group is 33 and shown in table 4.10.

Table 4.10 Allophone-to-viseme mapping based on the DCT feature vector

Viseme	Allophones with IPA
1	ച്2-[c], ച്3-[c], ച്4-[c]
2	ഉ4-[ə], ഉ5-[əʰ], ക്5-[t], ത്4-[d], മ്1-[tʰ], ദ്1-[d], ദ്2-[d], ന്1-[n], ന്2-[n]
3	ക്2-[kj], റ്2-[Tʰ]
4	ആ2-[a], ഹ്2-[h]
5	ഉ6-[uʷ], ട്3-[t], ഴ്1-[ʃ], ഷ്1-[ʃ]
6	ട്1-[d]
7	ഞ്1-[ɲ]
8	ഏ3-[E], വ്1-[kʰ], ല്1-[l], വ്1-[w]
9	ട്4-[T], റ്1-[tʰ], റ്2-[d], റ്3-[t]
10	ഐ2-[ei], ന്2-[n]
11	ഇ1-[i], ജ്1-[j]
12	ആ1-[a:], ത്3-[ð]
13	ഒ1-[ʷO], ഓ1-[ʷO:], ഓ2-[O:], ഒ്4-[ɳ>]
14	ബ്1-[B]
15	അ2-[A], ഏ1-[ʷe:], ഏ2-[e:], ഏ3-[e:]
16	റ്1-[r], ഹ്1-[H]
17	ട്2-[t], ഴ്1-[dʰ]
18	ഫ്1-[pʰ], മ്3-[m], വ്2-[v]
19	ഈ2-[i:]
20	യ്1-[y], റ്1-[r], ഉ്1-[l], ഴ്1-[z]
21	ഇ2-[ʷi], ഇ3-[yʰ], ഈ1-[ʷi:], ഏ1-[ʷe], ഏ2-[eʷ], സ്1-[s], ന്1-[nʰ]
22	ഉ1-[ʷu], ഉ2-[uʷ], ഉ7-[U]
23	ഔ1-[au], ഘ്1-[gʰ]
24	അ1-[ʌ], ക്4-[g], ക്6-[K], വ്3-[Kʰ], ഒ്1-[ɳ], ഒ്2-[ɳj]
25	ഐ1-[ai]
26	ഉ3-[u]
27	മ്2-[M]
28	പ്1-[p], പ്2-[β], പ്3-[b], പ്4-[P], വ്2-[b], ഭ്1-[bʰ], മ്1-[mʰ], മ്4-[m]
29	ക്3-[ɣ], ഗ്1-[G], ഒ്3-[ɳ<], ഒ്5-[ɳʰ]
30	ഊ1-[ʷu:], ഊ2-[u:], ഒ2-[O]
31	ച്1-[c], ച്2-[cʰ], ച്3-[Cʰ], ജ്2-[j]
32	ഡ്1-[d], ണ്1-[ɳ], ത്2-[tʰ]
33	ക്1-[k], വ്2-[Kʰ], ഗ്2-[g], ത്1-[tʰ], യ്1-[dʰ]

The allophones of the very same phonemes are grouped randomly into 33 viseme groups. Some of the vowel allophones are grouped with other vowel allophone classes and consonant allophone groups. However, it is fascinating that bilabial consonant allophones dispersed themselves without a crossover with other allophones, which shows its distinctive individuality in visual speech classification. After assessing the data-driven approach, it is necessary to study vowel allophone-to-viseme mapping and consonant allophone-to-viseme mapping since both allophones are randomly connected, as in table 4.10.

For grouping the DCT features of 31 vowel allophones, the gap function is analysed between the range 5 and 15, and the estimated cluster group is 11, as shown in table 4.11.

Table 4.11 Vowel Allophone-to-viseme mapping based on the DCT feature vector

Viseme	Allophone
1	ഇ1-[i], ഇ2-[yɪ]
2	എ1-[ye:], ഐ1-[ai], ഐ2-[ei]
3	അ2-[A], എ3-[E], ഉ7-[U]
4	ഇ3-[y ⁱ], എ2-[e ^y]
5	ഓ2-[O]
6	ഉ3-[u]
7	അ1-[Λ], ആ1-[a:], ആ2-[a]
8	എ3-[e:]
9	ഉ1-[^w u], ഉ2-[u ^w], ഉ6-[u ^v], ഊ1-[^w u:], ഊ2-[u], ഓ1-[^w O], ഓ2-[O], ഓ3-[^w O:], ഓ1-[au]
10	ഈ1-[yi:], ഈ2-[i:], എ1-[ye], എ2-[e ^r :]
11	ഉ4-[ə], ഉ5-[ə*]

The vowel allophones of the very same phonemes are grouped into different classes, thereby producing an overlapped many-to-many mapping. Not even a single phoneme has shown an individual clustering result compared to the phoneme-to-viseme mapping (fig. 4.9), which directly highlights the

sophistication of many-to-many mapping. The back high vowel $\text{ഉ}/\text{u}/$ and back mid vowel $\text{ഒ}/\text{o}/$ are highly correlated when it appears in a word. The number of elements in a viseme group varies from one (minimum) to nine (maximum).

The complexity of consonant allophone-to-viseme mapping is studied by clustering the DCT visual features and assessing the gap curve from 20 to 35. Table 4.12 shows the clustering of 75 consonant allophones into 25 viseme groups.

Table 4.12 Consonant Allophone-to-viseme mapping based on DCT feature vector.

Viseme	Allophones with IPA
1	ഫ്1-[p ^h]
2	ക്5-[t], ച്2-[c], ച്3-[c], ച്4-[c], ഉർ2-[C ^h], ജ്1-[J], ഷ്1-[s], ള്2-[t]
3	ണ്1-[n], ഴ്1-[z]
4	ഹ്2-[h]
5	ക്1-[k], വ്2-[K ^h], ട്3-[t], യ്1-[y], ര്1-[r], റ്1-[r]
6	ദ്1-[d], ഴ്1-[d ^h]
7	വ്3-[K ^h]
8	ഡ്1-[d], ന്1-[n ^h], ന്2-[n]
9	ഘ്1-[g ^h], ന്1-[n], ല്1-[l]
10	ത്2-[t ^h], ത്4-[d], മ്1-[t ^h], ദ്2-[d], ന്2-[n]
11	ണ്4-[n>]
12	ഉർ1-[c ^h], ജ്2-[j], ത്1-[t ^h], ശ്1-[ʃ]
13	പ്1-[p], പ്2-[β], പ്3-[b], പ്4-[P], ബ്1-[B], ബ്2-[b], ഭ്1-[b ^h], മ്2-[M], മ്4-[m]
14	ച്1-[c], ണ്1-[n], ത്1-[t], സ്1-[s]
15	ഖ്1-[k ^h], ട്2-[t], ഴ്1-[d ^h]
16	ട്1-[d]
17	ക്6-[K], ണ്2-[nj], ണ്3-[n<], ണ്5-[ŋ ^h]
18	ഗ്2-[g], ണ്1-[ŋ]
19	ക്3-[γ], ക്4-[g], ഗ്1-[G], ട്4-[T], റ്1-[t ^h], ള്1-[l], ഴ്1-[d]
20	മ്1-[m ^h]
21	ഹ്1-[H]
22	വ്1-[w]
23	മ്3-[m], വ്2-[v]
24	ക്2-[kj], റ്2-[T ^h], ത്3-[ð]

The bilabial consonant allophones have revealed their distinguishing power nearly in the same fashion as allophone-to-viseme mapping. The contextual variation of all other phonemes is randomly distributed. While comparing tables 4.10, 4.11 and 4.12, the common behaviour found is that allophones are unevenly distributed, i.e., there are a large number of smaller groups and a small number of larger groups. In addition, visual speech for allophones are analysed with five consecutive frames, which is suitable for phoneme level analysis. However, it may need either less than five or a dynamic number of frames for analysing the allophones because the time duration of the allophone may vary with the presence of neighbourhood phonemes in a word. To overcome this issue, a reference tool must be developed, just like in phonemes. Thus, to investigate further in visual speech, especially at the word level, the foremost work is to create an allophone-to-viseme mapping based on linguistic knowledge, which is still an unaddressed research problem. Hence, the remaining portion of this work deals with speech processing at the phoneme level only.

4.9 Conclusions

The core objective is to decode the visual information from the lips, which can be obtained by developing the viseme set in the Malayalam language. For viseme mapping a linguistically involved data-driven approach is used, which can take advantage of individual perception modelling from a linguistic perspective and the computational ease of a data-driven approach. In the first phase of the study, phoneme-to-viseme or many-to-one mapping is made according to the linguistic and data-driven approach. In the linguistic approach, 50 phonemes grouped into 14 viseme classes based on linguistic knowledge. In the data-driven approach, the viseme set is created by clustering the numerical representation of phonemes using the k-means algorithm and gap statistics. Geometric and DCT visual features are used in the data-driven

approach. Three viseme mappings (static frame alone, three frames, and five frames) were compared with the linguistic picture and visual speech duration, thereby identifying the best representation of visual speech in terms of frames. The above statement is verified by analyzing the visual duration of all Malayalam phonemes. The average durational range of vowel phonemes is 0.50 s to 0.70 s, which corresponds to 12 frames to 17 frames, respectively. Since the vowel phonemes can be visually modelled using a single relevant frame, additional frames may increase the computational cost. The average durational range of consonant phonemes is 0.10 s and 0.30 s, which correspond to 2 frames to 7 frames. Thus, for the Malayalam language, consecutive five frames are needed to represent the visual speech element, especially for the consonant phoneme. After analyzing three viseme sets, DCT features based viseme map has shown a close resemblance with the linguistic map by classifying 50 phonemes into 16 viseme classes. In addition, only two landmark points are enough to extract a rotation-invariant ROI from the original image, which bypasses the time-consuming process of manual lip marking for extracting the lip region.

To examine the coarticulation effect in visual speech, an allophone-to-viseme many-to-many mapping is created. The data-driven approach-based allophone-to-viseme mapping is created in the second phase. DCT visual features clustered the 106 allophones into 33 viseme classes. In addition, vowel allophone-to-viseme mapping and consonant allophone-to-viseme mapping are investigated to provide a visual clustering result of vowel and consonant allophones. Due to the unavailability of linguistic classification of the visual part of allophones, the rest of this work is confined to phoneme level audio-visual speech processing.

CHAPTER 5

EFFECT OF INTENSE BACKGROUND NOISE ON ACOUSTICAL SPEECH PARAMETERS

5.1 Introduction

The previous chapters discussed the audio-visual speech database named “MOZHI”, its related attributes; simulated noisy speech and audio-visual asynchrony, and visual speech analysis. *Speech* is a very complex feedback process that involves production, perception and information processing in the brain. Various sources of corruption, such as background noise, reverberation, channel distortion, and so on, frequently occur in audio speech signals. In the development of speech recognition systems, background noise is a major factor to be considered. As a result, making speech processing robust to this type of real-world corruption is the primary challenge of any speech-based application.

Speech has a set of very distinct traits which capture human’s physiological features, language, background noise, etc. So, converting these traits to machine-readable format is an important task in speech processing termed feature extraction. The primary purpose of feature extraction is to provide a perceptually relevant representation of a digitalized waveform while removing redundant information, making computation easier. Since a single method is not enough to retrieve all relevant information from the speech signal, a comprehensive method is needed which explore the potentials of both time and frequency domain information. In addition, such a method must have an acceptable accuracy even when there is ambient noise.

There are various techniques to extract the relevant acoustical speech parameters like fundamental frequency, formant frequencies, MFCCs (Mel Frequency Cepstral Coefficients) [185]. This chapter discusses their performance in different noisy conditions and proposes modifications in these algorithms to make them noise-robust. The time-domain speech signal is transformed into the autocorrelation domain in which the effect of noise is suppressed to achieve noise robustness.

Pitch is a vital aspect of human speech. The fundamental frequency (F0) [186], or the lowest harmonic frequency, is frequently confused with pitch. The first is a physical characteristic of the underlying audio stream, whereas the second is a perceptual characteristic. Pitch is defined acoustically as the fundamental frequency (F0) of a quasi-periodic waveform. It represents the vibrational frequency of the vocal cords during the production of voiced speech. The popular pitch extraction methods (Praat [187], YIN [188], RAPT [189], and PEFAC [190]) are compared with the proposed method, especially in noisy conditions. The human vocal tract is a muscular tube with varying cross-sectional area which is excited by chopped air from the vocal cords (at one end of the tube) or turbulent air at constriction (a point along the tube). The frequencies in the speech that transfer most of the acoustical energy from the excitation source to the lip (output) are called resonance frequencies or formant frequencies. The first two formant frequencies are estimated using the proposed method and evaluated in different noisy conditions. Modification in the MFCC, ACR-MFCC (Autocorrelation MFCC), is also reported and compared its performance with respective standard versions in an acoustically corrupted speech signal.

The content of this chapter is arranged as follows. Section 5.2 discusses different pre-processing steps in speech processing followed by a noise reduction method based on a higher lag autocorrelation function. Section 5.3

explains the proposed fundamental frequency estimation method and its performance in different noisy conditions. Section 5.4 deals with a noise-robust formant frequency estimation method. Section 5.5 displays the ACR-MFCC and its performance in different noisy conditions. Finally, section 5.6 concludes the work.

5.2 Pre-processing

It is necessary to make the segmented speech signal (Eq. 3.4) tuned for further analysis collectively termed as Pre-Processing. The performance of a speech recognition system depends significantly on different operations in the pre-processing stage. Amplitude Normalization, Pre-emphasis Filtering, Framing and Windowing are the commonly used pre-processing steps. In this work, the key step in all the acoustic speech feature extraction processes is transforming the pre-processed time-domain signal to the autocorrelation domain.

5.2.1 Amplitude Normalization

Speech signals must be normalised in speech recognition systems to successfully compare unfamiliar spoken data with stored patterns or models. While recording a speech signal, the energy levels often vary due to speaker, speaking style, channel, background noise and microphone distance. As a result, differences in amplitude (or energy) among different speakers' utterances or the same speaker at various times must be avoided or minimised. Amplitude Normalisation eliminates the variable energy levels between signals, improving performance in energy-related metrics. Divide each signal sample by its greatest absolute value to normalise the amplitude; thus, the dynamic range of amplitude is limited to $[-1, +1]$. The segmented speech signal is shown in Fig. 5.1 (a), and the normalised speech signal is shown in Fig. 5.1 (b).

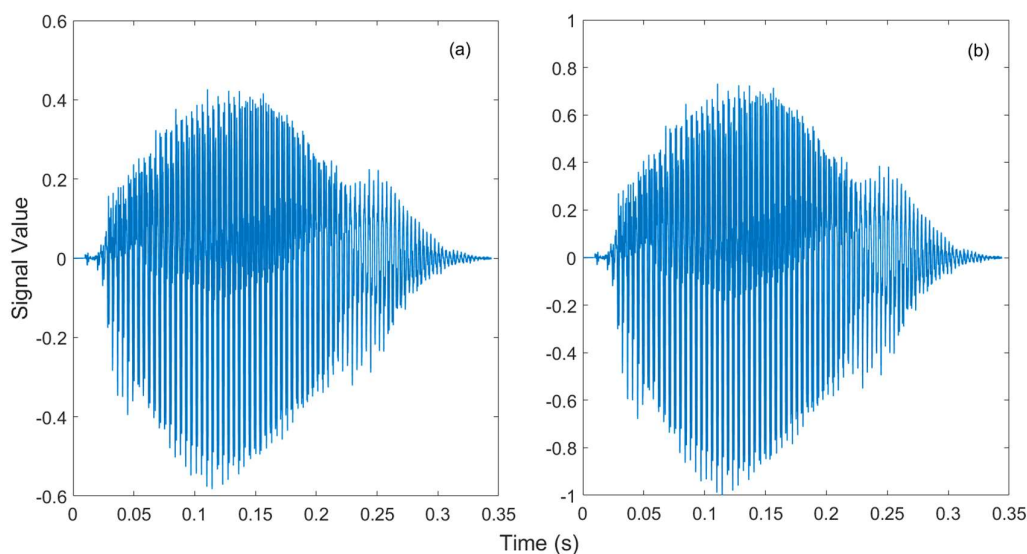


Fig. 5.1 Normalized of Speech Signal

5.2.2 Pre-emphasis Filtering

Because higher frequencies in the recorded speech signal contain less energy, it is necessary to emphasise energy at higher frequencies before processing the signal. For this the signal's value are re-evaluated using the formula:

$$x(n) = x(m) - 0.95 \times x(m-1) \quad (5.1)$$

The operation corresponds to a high-pass filter of the first order. Low-frequency signals sampled at a high sampling frequency (f_s) produce samples with identical numerical values. Because low frequency fundamentally indicates slow time variation, the numerical values of a low-frequency signal vary slowly or smoothly from one sample to the next. By excluding the parts of the samples that did not change with their neighbouring samples, the part of the signal that varies rapidly, i.e. its high-frequency components, is left. Thus, a pre-emphasis filter is used before using feature extraction algorithms to enhance the relative energy of the high-frequency spectrum. Fig. 5.2 (a) and fig. 5.2 (b) shows the normalized speech signal and its corresponding frequency

domain representation. Fig. 5.2 (c) shows the pre-emphasized speech signal, and the corresponding frequency response is shown in fig. 5.2 (d). This method must be chosen wisely in frequency domain analysis because the amplitude of the low-frequency region is reduced, so it is necessary to bypass this method for the fundamental frequency (F0) estimation task. However, it is suitable for the extraction of formant frequencies where higher frequency components are important.

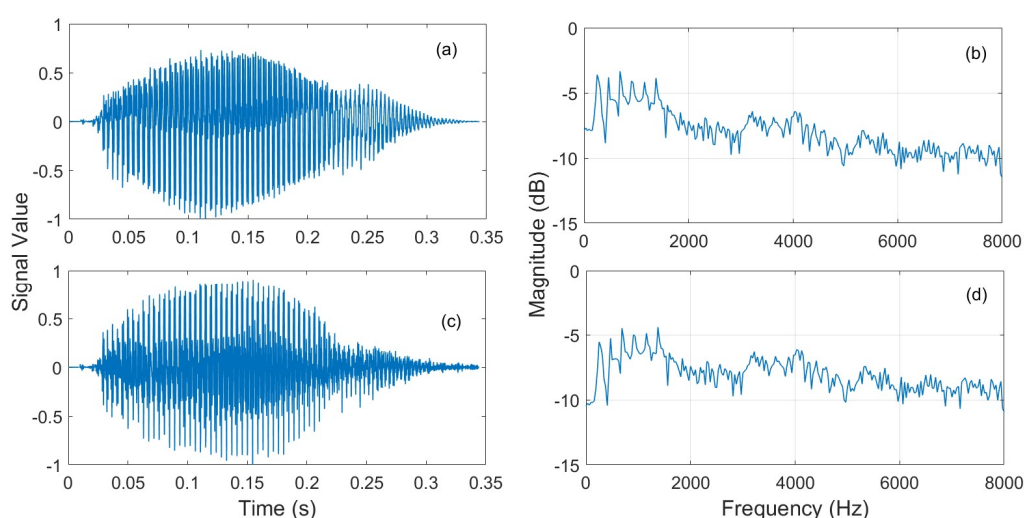


Fig. 5.2 Pre-emphasized Speech Signal

5.2.3 Framing

A time-varying vocal tract system with time-varying excitation produces speech. Thus, the speech signal becomes quasi-stationary. Usually, signal processing principles deal with a time-invariant signal. As a result, signal processing methods are not immediately applicable to speech processing. The speech signal may appear stationary for an interval of 10-35 ms, therefore the speech signal must segment before applying different signal processing methods. The speech signal is divided into several frames based on a certain predefined number of samples. As a result, by extracting the essence from the

speech part-by-part rather than the entire signal, framing the speech signal will help improve the recognition part's accuracy and stability.

Two parameters are required to frame a speech signal properly: window length (f_{length}) and overlapping length ($f_{overlap}$). The window length is the number of samples in each frame of the speech signal. Overlapping length is the number of samples in the overlapped frames of the speech signal. It is the same as that of window length, but the difference comes to play in the way where the overlapping frame is located. While analysing the speech signal only with the individual frame, some of the information at the beginning and end of each frame is lost. By using an overlapping frame, the lost information is reincorporated back into the extracted features. Usually, overlapping frames begin from the centre of the previous frame and ends in the centre of the next frame. The window length chosen is 25 ms, which capture enough information and, it will be relatively stationary, and the overlapping length is 10ms. Thus, the speech signal is divided into N frames, each of length 400 samples ($f_s = 16$ kHz). The first frame starts from 0 ms to 25 ms, then the second frame starts from 10 ms to 35 ms, as in fig. 5.3. The total number of frames in a speech signal $x[n]$ is

$$\text{Number of Frames} = \frac{\text{length} \{x(n)\} - f_{length}}{f_{overlap}} + 1 \quad (5.2)$$

Sometimes, the last frame may contain fewer samples than that frame is extended to the original size of the previous one by padding zeros in the end. Fig. 5.3 shows the framing of speech signals.

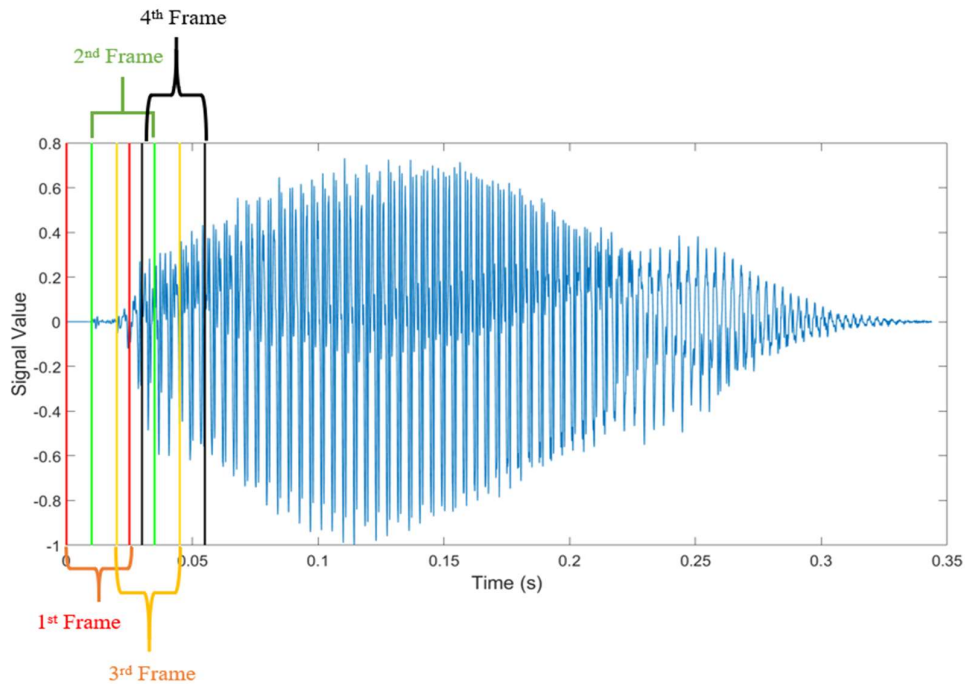


Fig. 5.3 Framing of Speech Signal

5.2.4 Windowing

Framing is the process of selecting a portion of the signal in which the nature of the signal is almost stationary. A windowing function is applied before processing stage to avoid the sudden discontinuity in each frame and distortion in the frequency domain. For given frame size, the nature of spectral information varies slightly with the window function. There are many types of windows such as Rectangular, Hamming, Hanning, Gaussian windows etc. Window function maintains its property within the frame and drops to zero outside it. The rectangular window function maintains the signals within the frame, but its spectral features change abruptly at the beginning and ending points. As a result, a window function that gradually reduces the points near the beginning and end of each frame to zero must be used. In speech recognition, the most commonly used window function is Hamming window. Hamming window is defined by

$$W_{\text{hamm}}(n) = 0.54 - 0.46 \cos\left(2\pi \frac{n}{N-1}\right), \quad 0 \leq n \leq N-1 \quad (5.3)$$

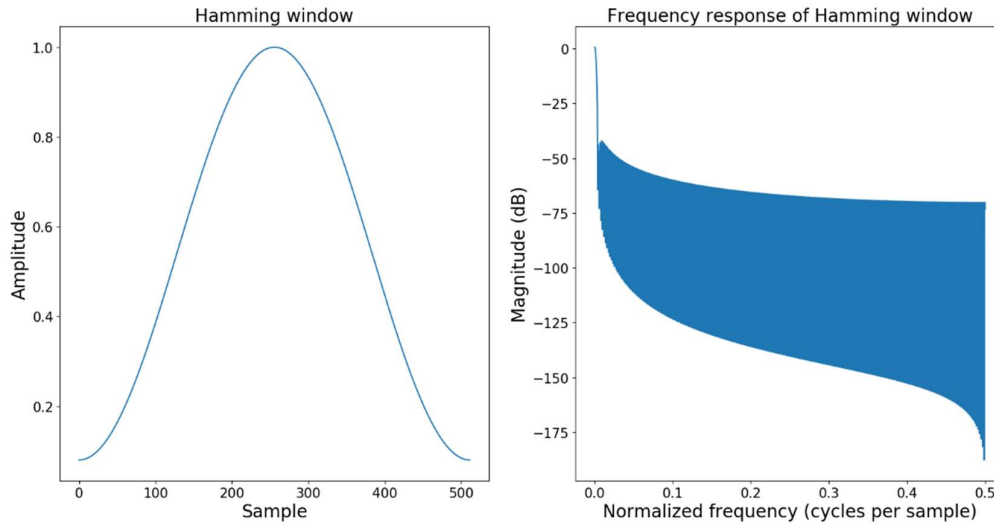


Fig. 5.4 Hamming Window

The framed speech signal $x(n)$, $n=0, 1, 2, \dots, N-1$, is multiplied with the Hamming window then, the resulting windowed speech signal is defined as and shown in fig. 5.5

$$x_{\text{hamm}}(n) = x(n) W_{\text{hamm}}(n) \quad (5.4)$$

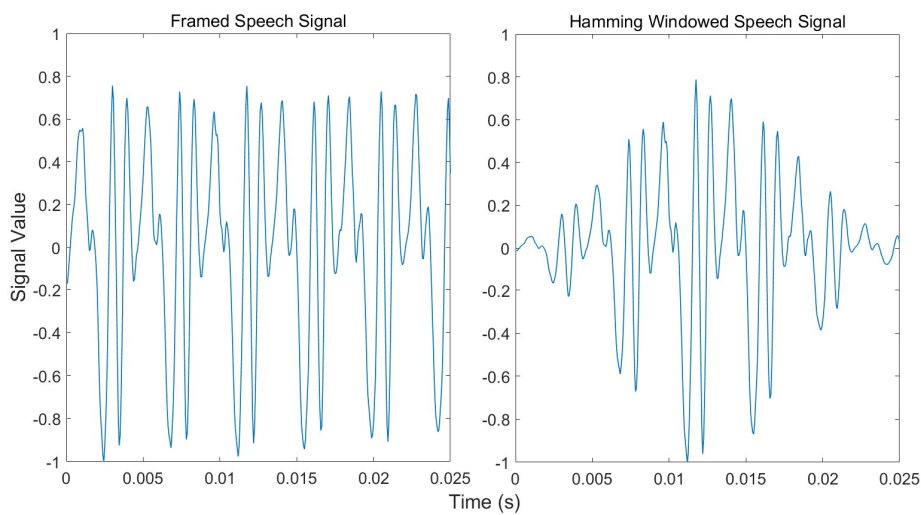


Fig. 5.5 Hamming Windowed Speech Signal

5.2.5 Autocorrelation

Autocorrelation is a measure of similarity between a signal and its time delayed version samples. It is estimated as the sum of products of the corresponding samples of a signal and its time-shifted version. It is considered as cross-correlation of a signal with itself. It is an even function with length $2n-1$ (as in fig.5.6 (b)), where n is the signal's length. So, correlation coefficients estimated for negative lags are discarded. Such correlation function is called one-sided autocorrelation function and is estimated as

$$r(i) = \frac{1}{N} \sum_{n=0}^{N-1-i} x_{\text{hamm}}(n)x_{\text{hamm}}(n - i) \quad (5.5)$$

Each frame of the windowed speech signal is then autocorrelated. The autocorrelation function of the speech signal is the convolution of the autocorrelation function of the source (glottal impulse) and system (vocal tract) components. It is periodic with smooth spectral envelop information repeated periodically over the entire lag. Fig. 5.6 shows the power spectrum and autocorrelation function of the Hamming windowed speech signal of vowel phoneme /a/.

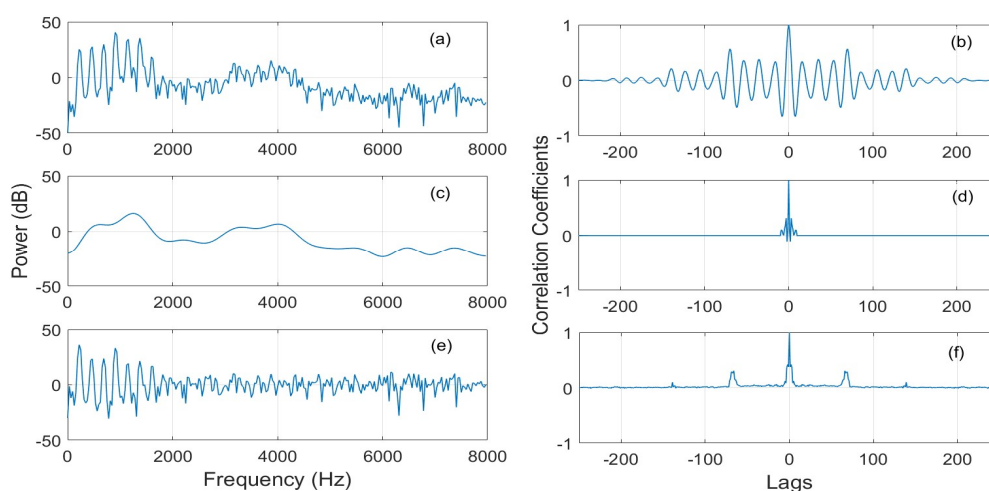


Fig. 5.6 Power Spectrum and Autocorrelation function of Hamming windowed speech signal.

Fig. 5.6 (a) shows the power spectrum of the speech signal, and its associated autocorrelation function is shown in fig. 5.6 (b). The system (vocal tract) information is represented by a smooth spectral envelope (fig. 5.6 (c)) which is computed from the first 20 cepstral coefficients of the speech spectrum (discussed in section 5.4). Fig. 5.6 (d) shows the autocorrelation of the Inverse Fast Fourier transform (IFFT) of the smooth spectral envelope. The source (glottal impulse) information is represented by the excitation power spectrum, which is obtained by subtracting the smooth spectral envelope from the speech power spectrum and is shown in fig. 5.6 (e). Its associated autocorrelation function is obtained by subtracting the IFFT of the smooth spectral envelope from the speech spectrum and is shown in fig. 5.6 (f). The system information's autocorrelation function is limited to lower lags alone. On the other hand, the source information provides the voiced speech signal's periodicity and the unvoiced speech signal's non-periodicity. Convolution of the autocorrelation functions of system and source information along with its corresponding phase information (which is avoided while computing power spectrum) will get the autocorrelation function of speech signal as in fig. 5.6 (b). The speech information is provided as a function of lag (or time shift) in the autocorrelation domain, while the power spectrum is presented as a function of frequency. From fig. 5.6 (d), the autocorrelation peaks represent the formant frequencies in lags which corresponds to the spectral peaks in fig. 5.6 (c). The separation of spectral peaks in fig. 5.6 (e) gives the vocal cord vibration information, also retained as autocorrelation peaks in fig. 5.6 (f). Thus, analyzing the speech signal in the autocorrelation domain is enough since it contains the same information as the signal's power spectrum.

Because this work is being done on noisy speech signals, it is worth examining the power of autocorrelation to suppress the noise. When the speech signal $x(m)$ is corrupted with additive noisy signal $n(m)$, the autocorrelation function of the noisy speech signal $y(m)$ is given by

$$\begin{aligned} r_{yy}(i) &= \frac{1}{N} \sum_{n=0}^{N-1-i} [x(m) + n(m)][x(m-i) + n(m-i)] \\ &= r_{xx}(i) + r_{xn}(i) + r_{nx}(i) + r_{nn}(i) \end{aligned} \quad (5.6)$$

Since the speech signal and noisy signal are uncorrelated, the cross-correlation terms $r_{xn}(i)$ and $r_{nx}(i)$ are relatively small. As a result, Eq. 5.6 contains the autocorrelation function of the speech signal and noisy signal only. Due to the random nature of the noisy signal, the autocorrelation function $r_{nn}(i)$ will have a peak at zero lag and is expected to decay rapidly to zero for higher lag values. Consequently, only $r_{xx}(i)$ is expected to have large peaks at $i \geq 0$. Thus, taking the autocorrelation function of the noisy speech signal will retain most of the speech information even at ambient noise level.

In the rest of this thesis, the autocorrelation function is denoted by ACR. Fig. 5.7 shows a comparison of different noises at 10 dB noise level both in the time domain and autocorrelation domain. It is evident from fig. 5.7(a) that the periodic nature of speech signal (as in fig. 5.5) is almost ruined in time-domain while adding white Gaussian noise at 10 dB noise level. However, the periodic nature is retained in its corresponding autocorrelation domain (fig. 5.7(b)), making ACR a better domain for representing noisy speech signals for extracting the source information. Fig. 5.7 also exhibits the effect of different noises at the same noise level on a speech signal. Compared to the similar noise levels of pink and red noise, the periodic nature and audible quality of speech signals are greatly affected by white gaussian noise. The noisy signals (fig. 5.7(c), (g), and (k)) were extracted by subtracting the noisy speech signal (fig. 5.7(a), (e), and (i)) from the clean speech signal fig. 5.5 respectively. Noisy effect in autocorrelation domain (fig. 5.7(d), (h) and (l)) is less prominent when compared to time-domain representation (fig. 5.7(c), (g) and (k)). The noisy signal's autocorrelation coefficients are concentrated around the lower lag, whereas the autocorrelation coefficients near the higher lag are very small.

Thus, in addition to retaining the periodic nature, ACR is also robust for analysing speech signals in intense background noise. This work uses the ACR for robustly extracting different features, especially in intense background noise.

Repeating pattern in the ACR is utilised to extract the fundamental frequency (F0) from the speech signal. It is achieved by estimating the difference between two consecutive local maxima in the ACR. The performance is evaluated in different noisy speech signals at different noisy levels, discussed in section 5.3. Frequency domain representation of speech signal is efficient for characterizing the vocal tract features especially, formant frequencies (F1, F2 and F3). Formant frequencies are reflected as spectral peaks in the frequency domain. MFCC (Mel Frequency Cepstral Coefficients) is the widely utilized acoustic feature in speech recognition systems. MFCC performs well for speech recognition in a clean environment, but its performance deteriorates in noisy speech processing. Background noise has a significant impact on power spectrum estimates in MFCC, lowering the recognition rate. Since ACR is robust in noisy conditions, it is used as an initial step in the MFCC algorithm, which enhances the efficiency of MFCC features in noisy speech. Since power spectrum estimate is the foundation of formant frequencies and MFCC estimation, it is crucial to analyse the spectrum behaviour of ACR of the speech signal, which is shown in fig. 5.8. Fig. 5.8 (a) shows the power spectrum of the Hamming windowed speech signal

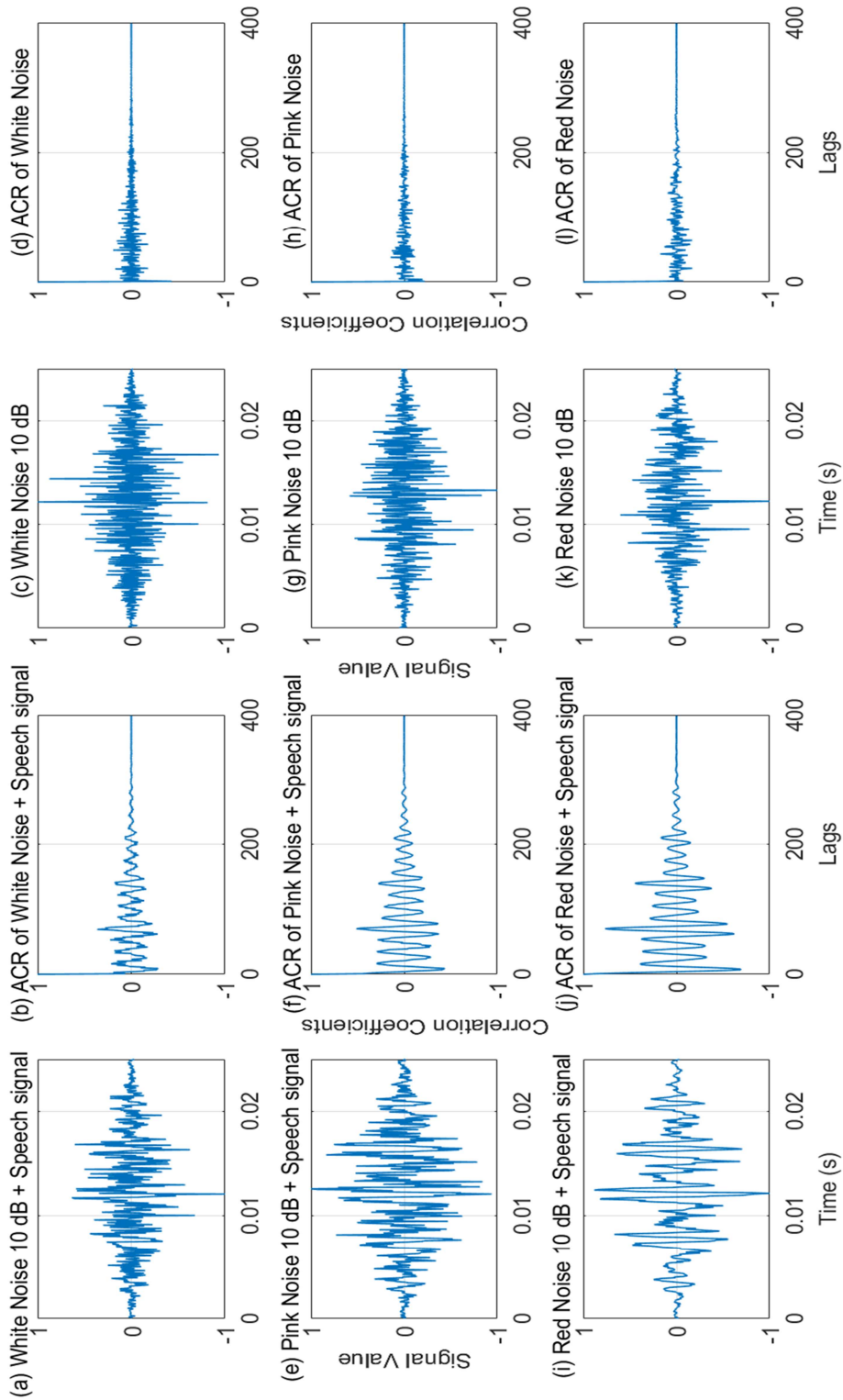


Fig. 5.7 Comparison of different Noises at 10 dB level in Time domain and Autocorrelation Domain

(dotted line), and fig. 5.8 (b) displays the power spectrum of ACR of the Hamming windowed speech signal. At low frequencies, the power spectrum of autocorrelation collects harmonic information, but it completely misses harmonics with magnitudes below -5 dB. In other words, the autocorrelation power spectrum's dynamic range is limited to around 43 dB (between 38 dB and -5 dB). It is because of the dynamic range of the Hamming window, where the the maximum side lobe's magnitude is around 43 dB lower than the main lobe's, as illustrated in fig. 5.4.

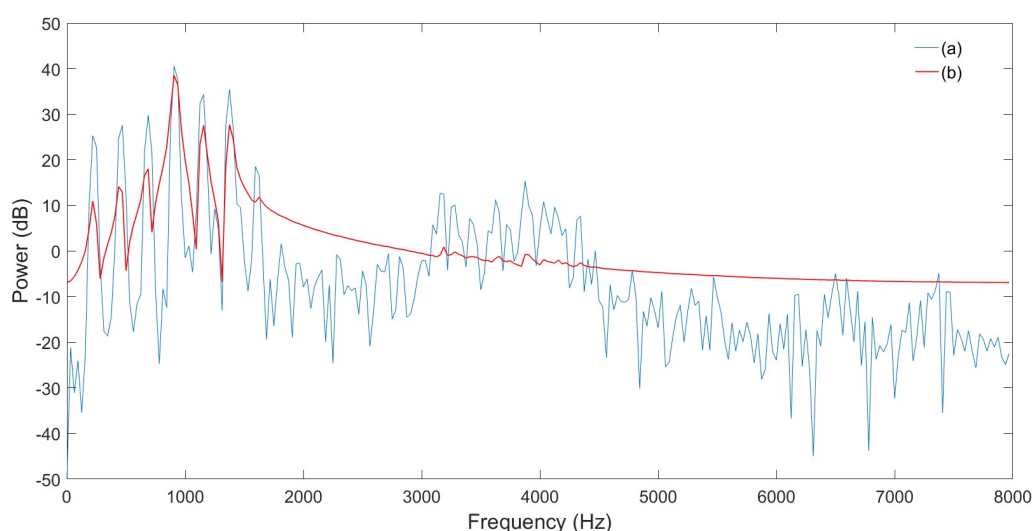


Fig. 5.8 Power Spectrum of Speech Signal and corresponding ACR.

To overcome this issue while retaining the noise robustness of ACR, a new window function is applied before spectral analysis. The most common window function used in speech recognition tasks is Hamming window. However, the dynamic range of a Hamming window is approximately 43 dB (fig. 5.4). To produce a power spectrum of ACR equivalent to that of the power spectrum of the speech signal, a window function with double dynamic range must be used. A Hamming window with a double dynamic range is designed by creating a Hamming window of size $N/2$. Then, its two-sided autocorrelation sequence of length $N-1$ ($2*(N/2)-1$) is computed with maximum value at the

zeroth lag. Finally, a zero is padded at the end of the autocorrelation sequence to obtain a window of length N. This window function is referred as Double Dynamic Range (DDR) Hamming window [134]. DDR Hamming window function and its frequency response is shown in fig. 5.9.

Thus, before estimating the power spectrum of ACR, $ACR(R(n))$ of the speech signal is multiplied with the DDR Hamming window ($W_{DDRhamm}(n)$) to get the windowed ACR.

$$R_W(n) = R(n) \cdot W_{DDRhamm}(n) \tag{5.7}$$

where $n=0, 1, \dots, N-1$. The windowed ACR along with ACR is shown in fig. 5.10.

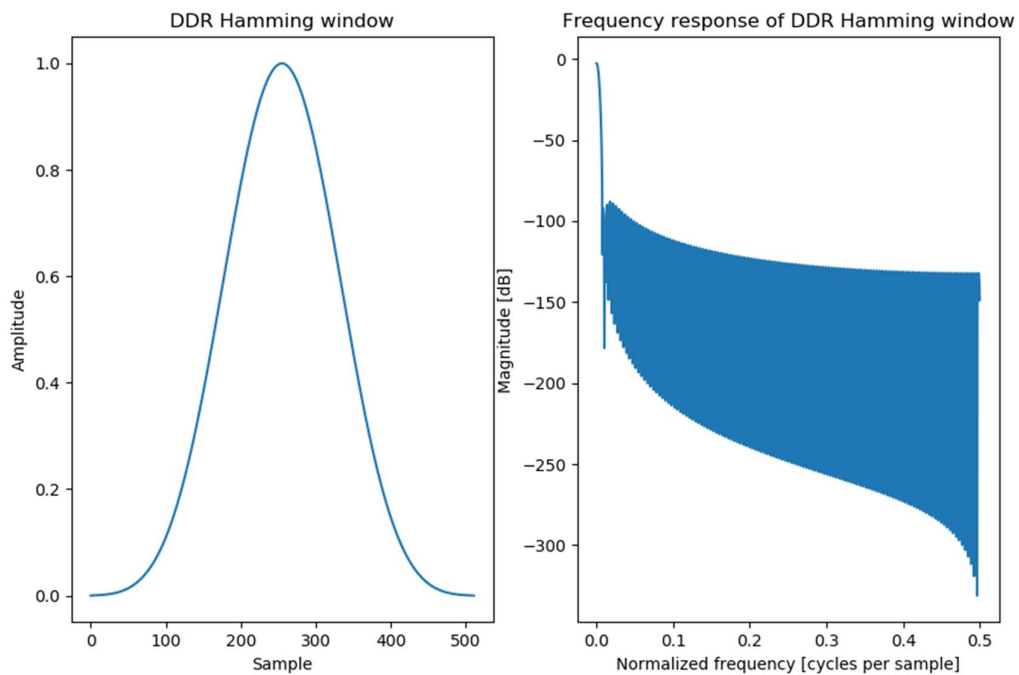


Fig. 5.9 DDR Hamming Window

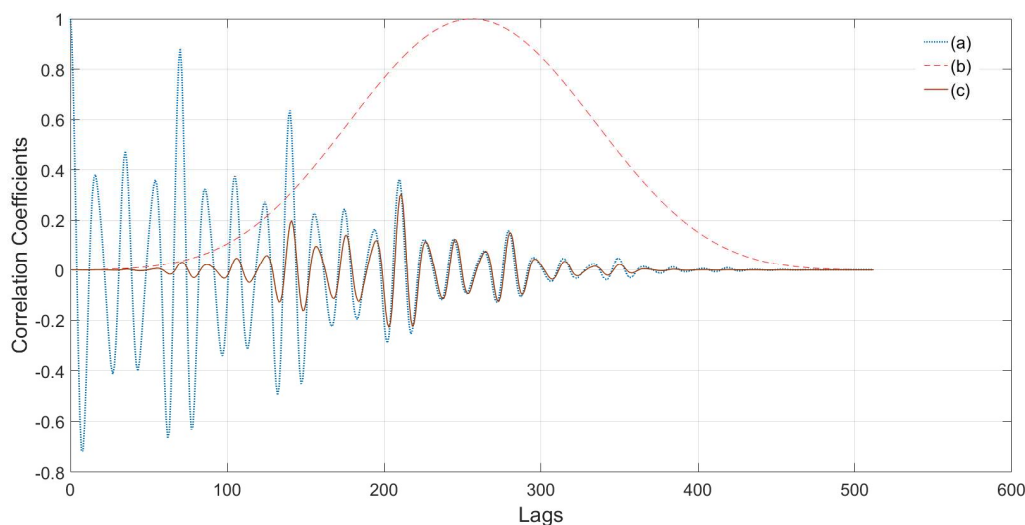


Fig. 5.10 Illustrations of DDR Hamming window on One-sided ACR. (a) One-sided ACR. (b) DDR Hamming window. (c) Windowed ACR.

The power spectrum of DDR Hamming windowed ACR and the Hamming windowed speech signal is shown in fig. 5.11 (c) and 5.11 (a) respectively. The dynamic range problem is solved using the DDR Hamming window function by retaining the entire spectrum structure marginally lowered compared with the speech spectrum. It also neglects small fluctuations near the harmonics by creating slightly wider harmonics which does not affect the performance of the speech recognition system since the DDR Hamming window captures the formant structure (fig. 5.11 (d)) similar as in Hamming windowed speech signal (fig. 5.11 (b)).

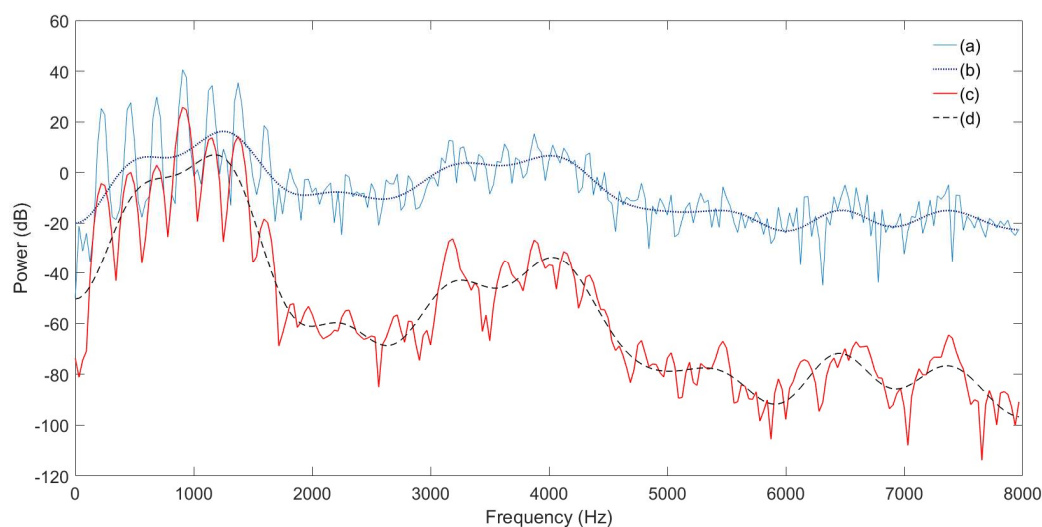


Fig. 5.11 Comparison of Power Spectrum. (a) Hamming windowed Speech signal. (c) DDR Hamming windowed ACR. Smooth spectral envelope of speech signal (b) and DDR Hamming windowed ACR (d).

The spectral behaviour ACR is to be investigated for its noise robustness in different noisy conditions at different noise levels. Fig. 5.12 shows the power spectrum of DDR Hamming windowed ACR (right) and its corresponding speech power spectrum (left) along with formant structure is also displayed. The speech signal corrupted with noisy signals at 10 dB noise level is used for this illustration. Fig 5.12 (a) and (b) shows the power spectrum of clean speech signal and its corresponding ACR power spectrum. Fig 5.12 (c) and (d) displays the power spectrum of white gaussian noise added speech signal and its corresponding ACR power spectrum respectively. It is quite evident from this illustration that white gaussian noise added speech spectrum displays harmonics similar to the speech spectrum contrary to its behaviour in time domain and autocorrelation domain representation as in fig. 5.7 (a) and (b), respectively, when compared with other noisy speech signals. Fig 5.12 (e) and (g) displays the power spectrum of pink noise added speech signal and red noise added speech signal respectively. Both noisy signals speech spectrum completely losses its prominent harmonics at the lower frequency range which

directly affect the formant frequency estimation. Fig 5.12 (f) and (h) shows the ACR power spectrum of pink noise added speech signal and red noise added speech signal respectively. Due to the presence of wide harmonics relatively better formant structure is obtained when compared to that of in fig. 5.12 (e) and (g). The performance of pink and red noise in frequency representation is quite different from its autocorrection representation (fig. 5.7 (f) and (j) respectively) where it is very similar to speech signal ACR (fig 5.6 (b)). Thus, in autocorrelation domain, pink noise added speech signal exhibit less error in fundamental frequency estimation when compared to other background noises. In contrary, white gaussian noise added speech signal displays better result especially in the first formant estimation (F1) when compared to other noises.

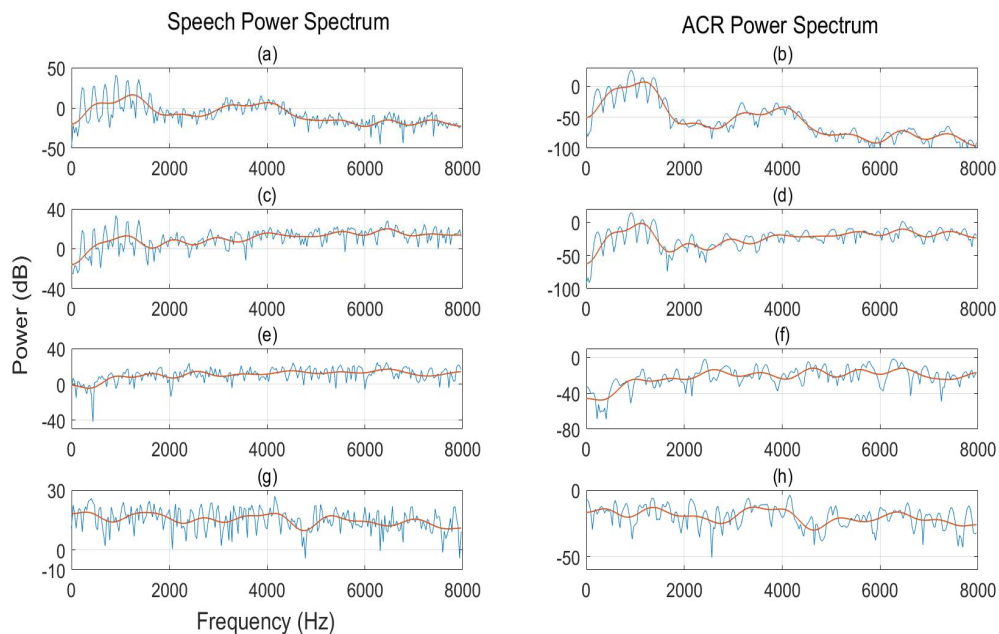


Fig. 5.12 Comparison of Speech Power Spectrum and ACR Power Spectrum in different Noisy Coinditions

The steps involved the modified pre-processing stage for extracting the noise robust acoustical speech features is summarized in fig. 5.13.

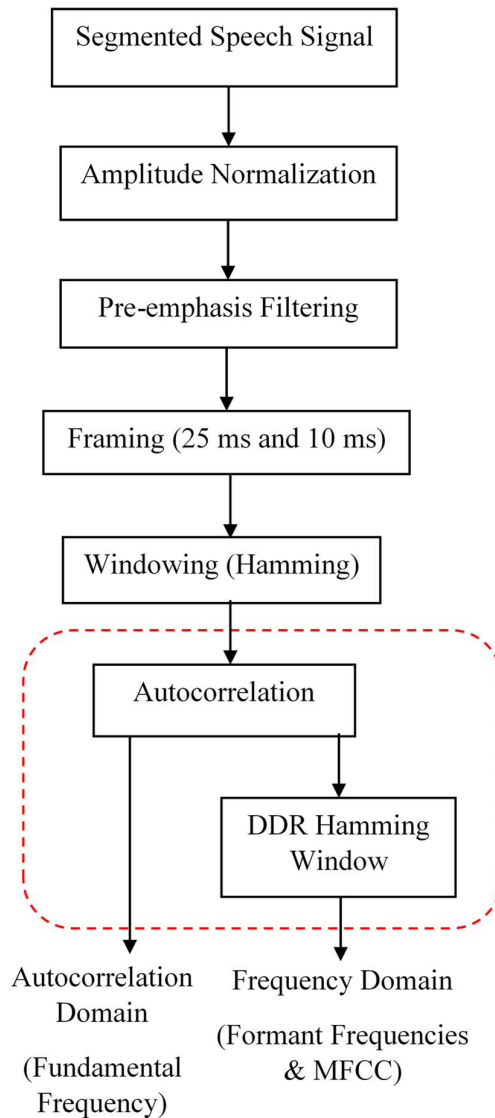


Fig. 5.13 Pre-processing Steps

5.3 Acoustical Speech Parameter- Fundamental Frequency (F0)

Fundamental frequency (F0) is the smallest frequency in a voiced speech signal, which is estimated from the frequency of quasi-periodic nature of the source sounds (vocal cords). Source sounds are complex and harmonic in nature. It consists of different frequencies, which are almost integral multiples

of the fundamental frequency. The spectral representation of source sound is characterized by the relative strength of fundamental frequency and its harmonics. Fig. 5.16 (e) shows the spectral representation of source sound, which is influenced by different factors like vocal cord's stress, F0, and phonation type. In domains like time-domain (fig. 5.5), autocorrelation domain (fig. 5.6 (b)) and spectral-domain (fig. 5.6(e)), the fundamental frequency is visible as peaks or differences between consecutive peaks. Typically, F0 in speech varies roughly between 70 to 400 Hz. The F0 of an individual speaker primarily depends on the elasticity of the vocal cords and is also related to the physiological features, language and ethnicity.

Automatic estimation of the fundamental frequency (F0), a widely used speech feature, is a fundamental problem in most speech and music processing applications. A variety of noise-robust F0 estimation algorithms were proposed in the past, but most of them focus only on its accuracy while few aims to improve the computation speed. Therefore, it is essential to develop an F0 estimation method that is accurate and computationally fast to suit real-time applications. The quasi-stationary behaviour of speech, the effect of vocal tract resonances, abrupt articulator change, and degradations owing to environmental factors such as noise and reverberation are all key elements that influence the F0 estimate.

In time-domain methods, F0 is computed from the repeating patterns of the speech signal or its transformed version. The most explored domain for F0 estimation is the autocorrelation domain, in which the stronger correlation peaks carry the source sound information. Methods such as Robust Algorithm for Pitch Tracking (RAPT) [189], Yet Another Algorithm for Pitch Tracking (YAAPT) [191], Praat estimate F0 by extracting local maxima of the autocorrelation or cross-correlation function. Various modifications to autocorrelation-based algorithms were made to minimize errors in the

estimated F0 as in the YIN method. In the frequency domain, the harmonic structure exhibits rich information about the pitch. Sub-harmonics to harmonics ratio (SHRP) [192], summation of residual harmonics [193], Sawtooth Waveform Inspired Pitch Estimator (SWIPE) [194], and others [195], [196] are examples of this type. The speech signal is split into numerous frequency bands in time-frequency domain pitch extraction algorithms, and each sub-band signal is subjected to time-domain procedures. The auditory-model correlogram based technique is a prominent time-frequency domain method [197]. An auditory filter bank is used to decompose the signal, followed by autocorrelation computation on each sub-band signal. Data-driven approaches are used in some methods [198], [199] to learn how noise influences the magnitude and location of peaks in the speech spectrum. Methods in [200]–[203] use statistical approaches to improve F0 estimation. The exact value of F0 of the speech signal is unknown as it is quasi-stationary. So different F0 estimation methods reported their accuracy by comparing with the F0 value estimated from the laryngograph (parallelly recorded glottal activity) [204]–[206] or Praat (a famous speech analysis tool).

Compared to formant frequencies estimate and other acoustical properties, F0 estimation is deemed as the most researched and active research topic in speech processing. Not a single conference or journal can pass without discussing a novel method or a modification in a popular algorithm for the F0 estimation in each year, which tells it is still a hot research topic. The main reason for this is the time-varying nature of speech signals. Speech features are vulnerable to speaker variations, languages and different parameters and steps in the feature extraction algorithms. Measurement parameters include minimum F0, maximum F0, window size, window type and other thresholding variables. Thus, the ‘true’ or ‘reference’ value (for both F0 and formant frequencies) is still unknown. So, researchers are forced to check the performance of different methods on synthesized signals where frequency

components are known even though this approach is easy to pick the most accurate method but fails to accommodate the dynamic nature of speech signals.

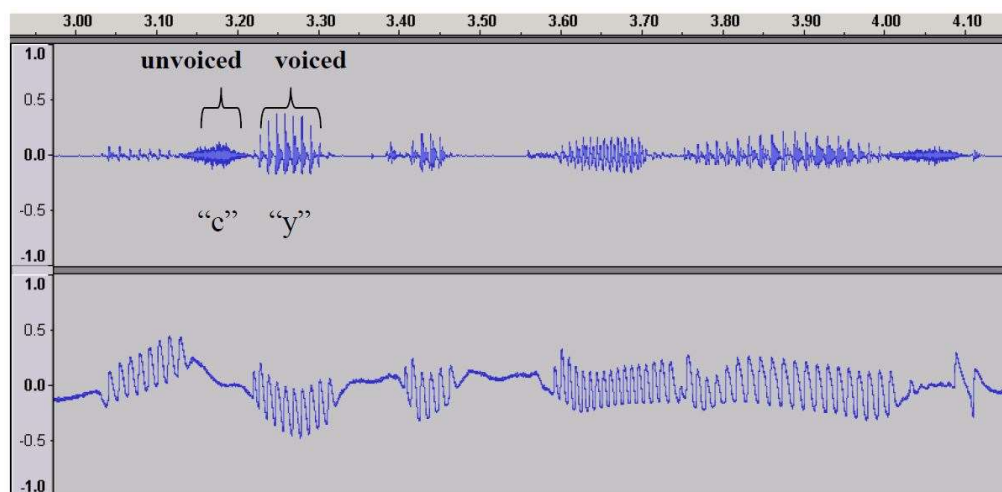


Fig. 5.14 The top figure shows the recorded microphone waveform over time, whereas the bottom waveform represents the laryngograph signal of the word “encyclopedias”. Speech includes voiced and unvoiced parts [205].

Another approach is to capture laryngograph signal (as discussed above) simultaneously with speech signal where the effect of vocal tract resonance is somehow diminished. However, only very few databases were created in this manner; most of the database in the literature captures only the speech signal using microphones. Furthermore, issues with asynchrony between two recording devices must be addressed, and variations in the laryngograph signal (fig. 5.14 (bottom)) must be normalised before feature extraction.

A performance evaluation strategy suitable for the database used for the study is to be adopted to select the outperforming F0 estimation method. This work uses an audio speech database recorded with a microphone. The performance of the F0 estimation algorithm is evaluated in this work as follows. The measured value of F0 of each frame of the clean audio speech signal of the isolated phoneme is treated as the true F0 value of that frame. The F0 of each

is measured for the speech signal with added noise. To evaluate the performance of the presented method, an estimate of the number of frames (n) having variances less than 5% of the true F0 value is utilised. The accuracy is estimated as

$$\text{Accuracy} = \frac{n}{N} 100 \% \quad (5.8)$$

Where N is the total number of frames.

An autocorrelation domain algorithm is used for F0 estimation, which is robust to different noises at low SNR levels. In this algorithm, the consecutive autocorrelation peaks are identified in each ‘clean’ speech frame, and the difference between their corresponding lags is used to estimate the ‘true’ fundamental frequency. The F0 value in Hz is the ratio of the sampling frequency (f_s) to the lag difference. The zero-lag coefficient gives the first correlation peak. The next highest peak is chosen as the second peak. However, this algorithm performs poorly in noisy conditions because of numerous peaks (depends on the noise nature) in between the prominent peaks. To overcome this issue, peaks corresponding to frequencies higher than the maximum expected fundamental frequency (400 Hz) are skipped. Thus, after picking the zeroth lag correlation peak, the first 40 lags ($f_s/F0_{\max}$), which contain frequency components higher than the F0 range, are omitted during the second peak selection. Since most noisy information is concentrated near the lower lag correlation coefficients, this simple algorithm will reduce the noise components to an extent. The block diagram of the proposed F0 estimation algorithm is shown in fig. 5.15.

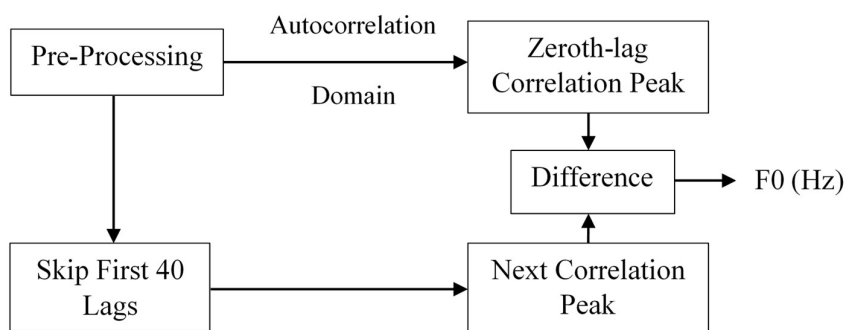


Fig. 5.15 Block Diagram of proposed ACR F0 Estimation Algorithm

Hundreds of fundamental frequency estimation methods have been proposed in the past. Most of them are innovative in their methodology and sophisticated in their algorithm. Some of them are dedicated to specific acoustic signals, and only very few were reliable in extreme background noise. The success rate of the proposed method in capturing the periodic nature of speech signals in low SNR conditions is evaluated in this work. The performance of the proposed ACR method is compared with that of the four prominent algorithms: Praat, YIN, PEFAC and RAPT. These were selected because of their wide acceptability and other potential features. *Praat* is a standard speech analyzing software in broad research domains like linguistics, acoustics, engineering, etc. In Praat, the F0 estimation is computed with autocorrelation function in each 10 ms speech segment with a 75-500 Hz pitch search range. Another flavour of the autocorrelation function is YIN. Starting from the autocorrelation function, YIN introduced different modifications to avoid estimation errors. It has no upper limit for the F0 search range, which also makes it useful for music signals. PEFAC estimates the fundamental frequency of each frame by convolving its power spectrum in the log-frequency domain with a filter that sums the energy of the pitch harmonics. It estimates the pitch reliably even at negative SNR. PEFAC includes several algorithm parameters whose values were determined empirically from the training data. RAPT

capture peaks in the normalized autocorrelation function and uses dynamic programming for pitch selection within the F0 search range 60-400 Hz. Although RAPT is an old method, it is included in this study as it uses the value of F0 obtained from the laryngogram as the true F0 value.

For conducting this experiment, an audio database (which is not mentioned in chapter 3) consist of a single speaker uttering a vowel phoneme 20 times. It is utilized to compare the performance of different fundamental and formant frequency estimation algorithms with the proposed methods. Table 5.1 shows the variation of F0 estimation accuracy with SNR of differed noise types based on Praat, YIN, PEFAC, RAPT and proposed method (ACR). The reference F0 value estimated from each method is represented as mean \pm standard deviation. The performance is evaluated on three noisy signals with SNRs 20 dB, 10dB, 0dB, -10 dB and -20dB. This comparison is carried out without changing the default parameter values of the reported methods.

Table 5.1 Variation of F0 estimation accuracy with SNR of differed noise types based on Praat, YIN, PEFAC, RAPT and ACR

Method	F0 (Hz) Clean Speech	Noise type	Frames within 5% deviation (%)					Frames
			20 dB	10 dB	0 dB	-10 dB	-20 dB	
Praat	252 \pm 31	White	97	87	53	0	0	30
		Pink	97	83	63	0	0	
		Red	100	100	93	53	0	
YIN	293 \pm 19	White	100	94	50	0	0	18
		Pink	100	94	39	0	0	
		Red	100	100	56	0	0	
PEFAC	255 \pm 38	White	96	96	0	0	0	26
		Pink	96	96	92	0	0	
		Red	100	96	88	62	19	
RAPT	244 \pm 45	White	97	81	29	0	0	31
		Pink	94	77	52	0	0	
		Red	97	90	71	32	0	
ACR	254 \pm 33	White	97	88	72	38	22	32
		Pink	97	88	75	44	25	
		Red	97	91	81	41	9	

This section summarises the performance of five algorithms in different noisy conditions. Examining the accuracy of these methods, it is found that the proposed method ACR outperforms other well-known methods. All methods perform well in red noise added speech signal, which is evident from fig. 5.7 (j) where speech information is retained even at extreme negative SNR. Thus, the best method is decided by the performance of these methods in the other two noisy signals, especially in white Gaussian noise, where the speech structure is intensively affected compared to pink noise. Praat shows comparable results up to 0 dB; after that, it fails to capture the periodic nature, especially in white Gaussian noise added speech signal. Another noticeable finding is; YIN is inferior to others in terms of estimation of true F0 value. However, YIN performs well in white Gaussian noise added speech signal when compared to PEFAC and RAPT. PEFAC has emerged as a robust method in extremely noisy conditions, but in this work, it fails to exhibit its robustness, especially in white Gaussian noise added speech signal. However, it is the only method that outperforms other reported methods and the proposed method in red noise added speech signal. RAPT is the only method that does not achieve a significant result in all three noisy speech signals. The proposed method (ACR) exhibit outstanding performance mainly in white Gaussian noise added speech signal. However, it fails to maintain that performance in red noise added speech signal where it is below the Praat and PEFAC method. The ACR method is the only one that can extract the periodic information in more than one speech frame, even in extremely noisy conditions in all three cases. Fig. 5.16, 5.17 and 5.18 shows the performance of different methods in white Gaussian noise, pink noise and red noise added speech signal, respectively.

By analysing the performance of the five methods, it is noticeable that the proposed method (ACR) performs comparatively well in all three noisy speech signals in extremely noisy conditions. However, despite this conclusion, some literature suggests YIN and PEFAC is superior for noise robustness in

other databases (language). Thus, it is evident that the language under study highly influences the estimation of the fundamental frequency. In addition, it is also noticeable that each method does not act in the same way to the type of noise and SNR level. Thus, a combination of different methods (based on its performance) should be even more noise-robust when compared to its performance. The following section deals with the formant frequencies estimation in the noisy condition, which contain the same accuracy estimating policy that is not discussed there.

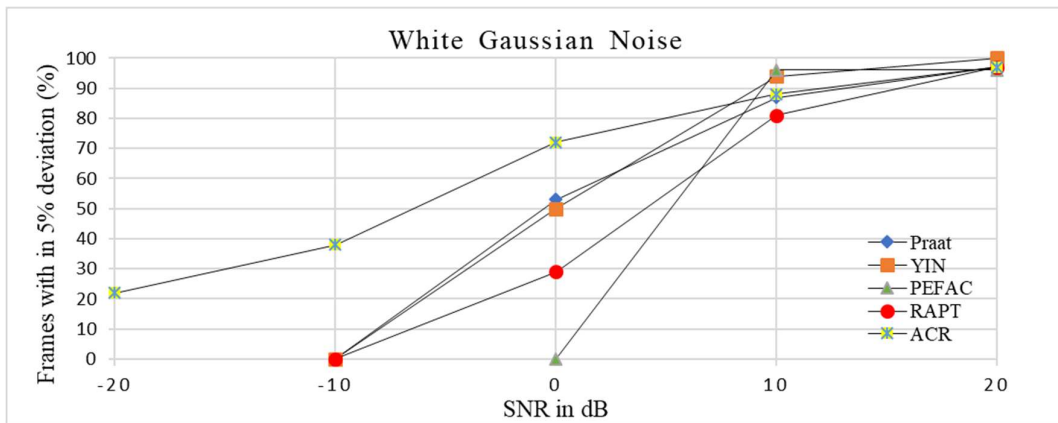


Fig. 5.16 Performance of Different Pitch Estimation Methods in White Gaussian Noise added Speech signal.

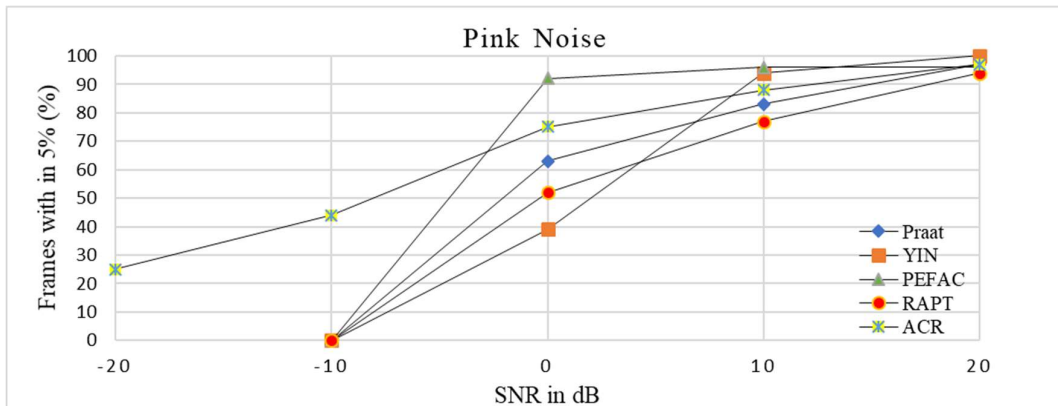


Fig. 5.17 Performance of Different Pitch Estimation Methods in Pink Noise added Speech signal.

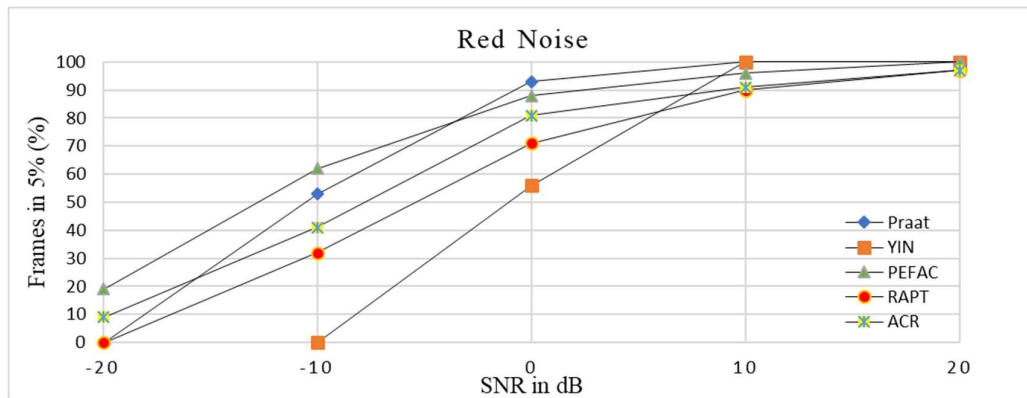


Fig. 5.18 Performance of Different Pitch Estimation Methods in Red Noise added Speech signal.

5.4 Acoustical Speech Parameter- Format Frequencies

The speech signal is composed of various frequency components which characterise the unique features of source sound (vocal cords) and system sound (vocal tract) in the speech production mechanism. Some frequency components are prominent in deciding the content of the spoken utterance. As the fundamental frequency of a speech sound is directly related to the characteristics of the vocal cords, formant frequencies are related to the vocal tract structure. The vocal tract is an acoustic space that shapes the frequency spectrum of the sound from the vocal folds as it passes through it. The frequency response of the vocal tract is characterised by its shape and size, which is influenced by the position of active articulators, mainly the tongue. However, a similar concept that frequently appears in the literature is resonant frequencies. Therefore, some authors blended the two concepts as identical, and some authors treated them as distinct. Resonant frequencies are the acoustical property of the vocal tract, while formant frequencies are that of the speech signal radiated from the lips. In this work, vocal tract characteristics is extracted from the speech signal followed by a filtering process. Thus, the corresponding frequency response of the vocal tract is termed formant frequencies. It is a concentration of acoustic energy represented by spectral peaks in the spectrum

or dark horizontal bands on the spectrogram, which is shown in fig. 5.19 and 5.20, respectively.

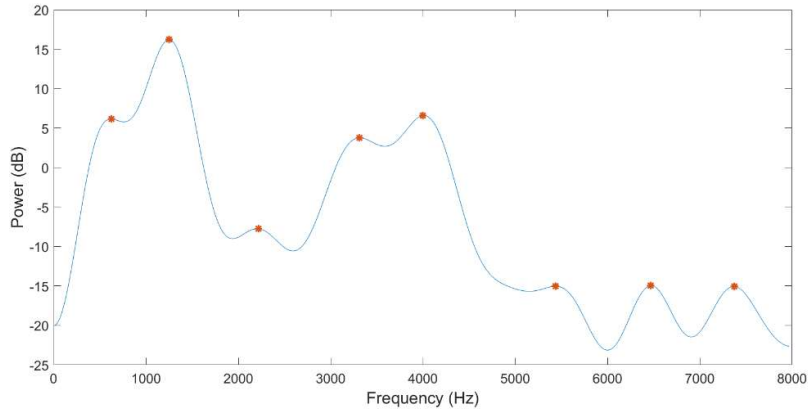


Fig. 5.19 Smooth Spectral envelope of vowel phoneme.

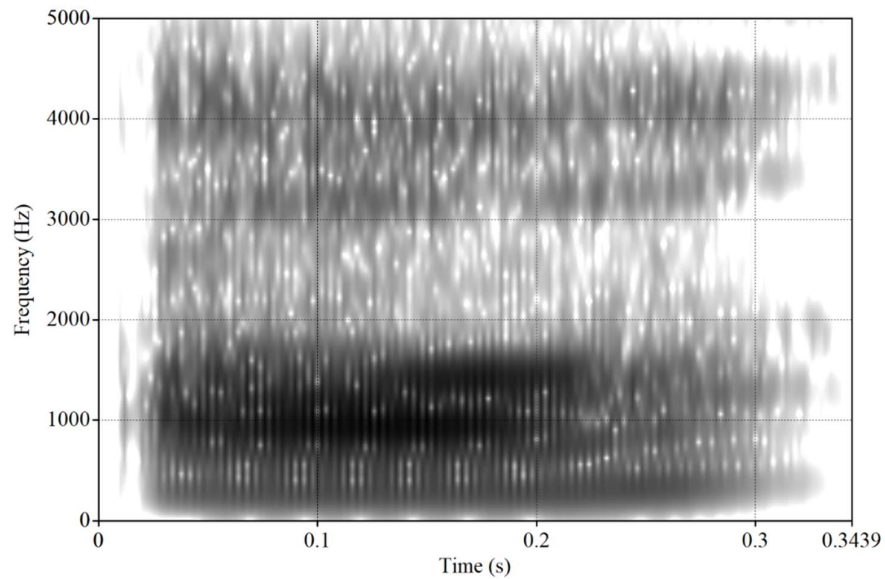


Fig. 5.20 Spectrogram of vowel Phoneme /a/.

In the speech, the voice spans a range of frequencies 70 Hz to 5000 Hz. Theoretically, there are one formant in each 1000 Hz band; thus, there will be five formants in a speech signal. The lowest formant frequency is F1, and the following higher frequencies were named F2, F3, F4 and F5, respectively. However, in the speech recognition perspective, F1 and F2 are enough to

capture speech content [207]. Thus, first two formant frequencies were used in the comparison process. From the speech production perspective, sudden change of vocal tract configuration is the primary source of formant variations. Even though the relationship between articulator configuration and formants are complex, the first formant F1 is correlated with the tongue height, i.e., the higher the tongue height, the lower F1. The second formant, F2, is related to the tongue's position, where F2 is higher for front vowels than back vowels. This relation is shown in fig. 5.21, where ഇ /i/ and ഉ /u/ are high vowels that have lower F1 when compared to others. Similarly, ഇ /i/ and എ /e/ are front vowels that have higher F2 when compared to others.

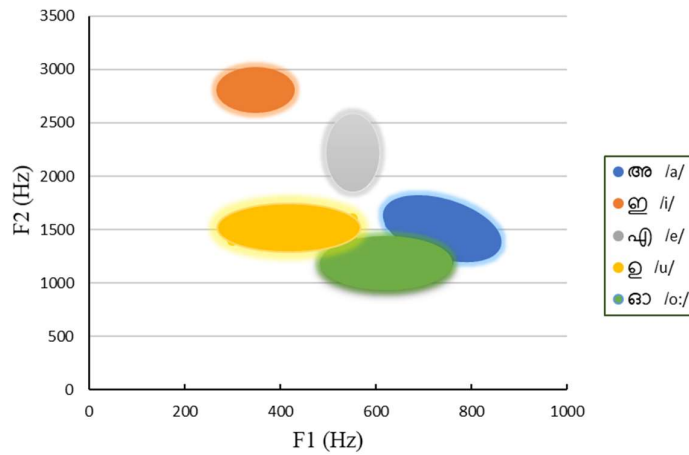


Fig. 5.21 F1 vs F2 plot of short Malayalam vowel phonemes

In addition, different people uttering the same vowel will have different formant structures due to the different vocal tract lengths among men, women, and children. The most error-prone area in the estimation of formant frequencies is the computational perspective. Since the speech signal is produced by convolving the source signal and system signal and the inherent limitations of the frequency domain restrict a broad, innovative outcome in the formant estimation methods compared to F0 estimation methods. The most

common formant frequencies estimation methods are a spectrum [208], spectrogram [209] and LPC (Linear Predictive Coding) or LP analysis [210].

Spectrum is one of the most straightforward ways to examine the frequency component of a signal. Since speech is a convolved signal, it is pretty challenging to pick the formants (from fig. 5.6 (a)) because their location is inferred from the relative amplitude of the harmonics. Another way is to deconvolve the speech signal and extract the spectral envelope of the system signal alone, as shown in fig. 5.19. This homomorphic decomposition is carried out in the cepstral domain. The red marks indicate the position of formant frequencies. Intense care is needed in the smoothing process since it may affect the shape and position of the formants. The harmonics become visible when the smoothing is not enough, and closer formant peaks will be lost when the smoothing is too high. The spectrogram is the visual representation of frequencies of a signal with time. The horizontal axis represents the time information, and the vertical axis represents the frequencies. The colour scale (usually greyscale) represents the amplitude of the observed frequency at a time. Darker shades of grey indicate higher amplitudes. Attempts to locate the position of the formants is quite problematic, especially for higher formants where energies are relatively low compared to the first few formants, as in fig. 5.20. Even though spectrogram is referred to as speech fingerprint, measurement of formant frequencies from spectrogram requires some degree of interpretation and linguistic knowledge. Linear Prediction analysis is one of the frequently used formant frequency estimation methods. In Linear prediction, the individual samples in a speech signal are predicted from previous samples' weighted combinations. These weights are termed coefficients which characterizes a digital filter that has the frequency response of the formants. LP analysis extracts the formant frequencies either by generating the LPC spectrum from the filter's frequency response or by finding the roots of the prediction polynomial. One of the critical parameters associated

with LPC is LPC order which determines the number of coefficients in the linear prediction. This parameter has the same problem as the smoothness of the spectrum when picking peaks. LPC is a simplified speech production model that does not account for the interaction between the vocal tract and source information and may perform poorly when the model does not sufficiently represent the speech signal. Since each method has its own merits and demerits, the most desirable approach is to consider these issues while proposing a method. The proposed method extracts the formant structure from the cepstral domain by fusing the noise robustness nature of the autocorrelation domain. The proposed method is termed ACR Cepstrum, whose block diagram is shown in fig. 5.22.

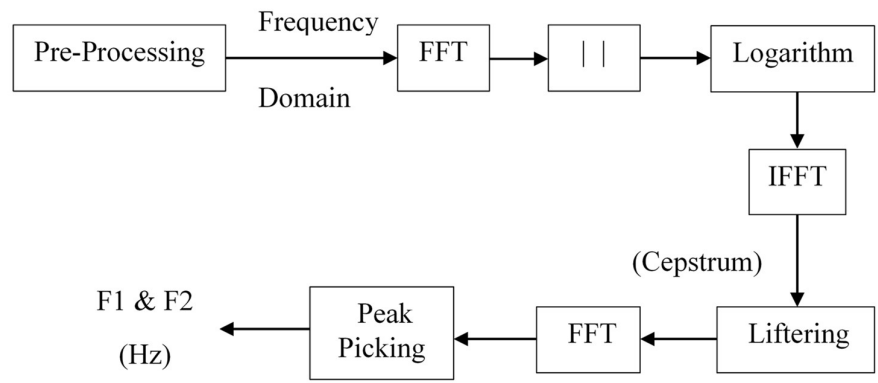


Fig. 5.22 Block Diagram of proposed ACR Cepstrum Algorithm.

The cepstrum of a signal is defined as the Inverse Fourier Transform (IFFT) of the logarithm of the signal spectrum [211]. Speech is a convolved signal in the time domain as well as in the frequency domain. To extract the vocal tract information $V(n)$ from the source information $S(n)$, the signal must be converted to a special domain where convolution becomes a summation problem. This domain is called the cepstral domain, and the process is termed homomorphic decomposition.

$$R_w(n) = S(n) * V(n) \quad (5.9)$$

Where $R_w(n)$ is the DDR Hamming windowed ACR and $*$ denotes the convolution operation. The windowed ACR is transformed to the frequency domain by taking the magnitude of the power spectrum on both sides of the above equation.

$$|R_w(n)|^2 = |S(n)|^2 \times |V(n)|^2 \quad (5.10)$$

Since convolution in the time domain is equivalent to multiplication in the frequency domain, the convolution operation is changed to product problem. Taking logarithm on both sides of Eq. 5.10.

$$\log |R_w(n)|^2 = \log |S(n)|^2 + \log |V(n)|^2 \quad (5.11)$$

The logarithmic spectrum helps visualize the spectral content by uniformly distributing the magnitude values throughout the spectrum, as shown in fig. 5.6 (a). Even though the convolution problem is transformed to a summation problem, $S(n)$ and $V(n)$ are still fused. To separate the two components of the speech signal, Inverse Fourier Transform is applied on the log spectrum as shown in Eq. 5.12.

$$|\mathcal{F}^{-1}\{\log |R_w(n)|^2\}|^2 = |\mathcal{F}^{-1}\{\log |S(n)|^2\}|^2 + |\mathcal{F}^{-1}\{\log |V(n)|^2\}|^2 \quad (5.12)$$

Where \mathcal{F}^{-1} represent the Inverse Fourier Transform. Even though the cepstrum involves two frequency transforms, the cepstral domain is not the same as that of the time domain of the original signal. The x-axis of a cepstrum is termed as quefrequency axis, which is expressed in the unit seconds. In cepstrum, the low quefrequencies contain the slow varying feature of the logarithmic spectrum, giving the formants structure information. The harmonic feature of the logarithmic spectrum is visible as comb structure in the high quefrequency range, as shown in fig. 5.23.

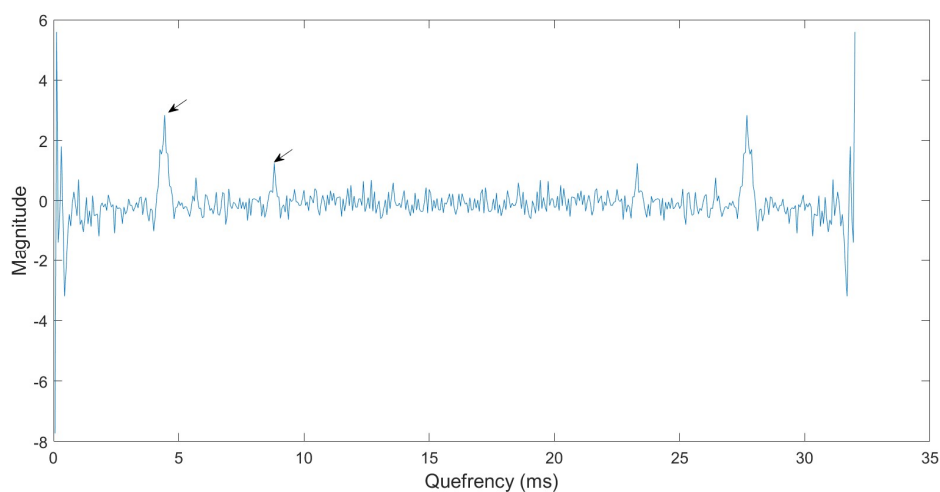


Fig. 5.23 Cepstrum of Speech Segment.

The arrow indicates the high quefreny region from which the fundamental frequency is estimated by taking the reciprocal of corresponding time information. Since the formant information is distributed over several cepstral coefficients, the information cannot be extracted as simple as F0 estimation. To remove the harmonic information from the cepstrum, a rectangular window is applied, termed Liftering (like quefreny). The number of cepstral coefficients determines the smoothness of the spectral envelope, which is obtained by taking the frequency response of the filtered cepstrum. This work chose the first 20 cepstral coefficients based on the trial and error method to extract the smooth spectrum, as shown in fig. 5.19. After extracting the smooth envelope, the spectral maxima were picked, corresponding to formant frequencies in the speech signal. The first formant is localized from the frequency band, which ranges from F0 to 1200 Hz. The F0 is obtained from the comb structure in the cepstrum. This wide range is used to include the variation due to speech production perspective as discussed above. The second formant is localized in the frequency range from F1 to 3000 Hz.

The proposed work is compared with the Praat method, which applies a Gaussian-like window (25 ms) and computes the LPC coefficients with the

algorithm by Burg. The performance of the proposed method is compared with the Praat method without changing its default settings on a database used in the F0 estimation stage. Table 5.2 shows the variation of F1 and F2 estimation accuracy with SNR of differed noise types based on Praat and ACR Cepstrum method. The most noticeable element is that Praat is more precise than ACR Cepstrum but fails to maintain robustness. The precision of measurement is more prominent for the first formant than the second formant in both methods. Both methods drop their performance mainly in pink noise added speech signal than the other two. While concerning tables 5.2 and 5.1, the performance of the proposed ACR method in white Gaussian noise added speech is contrary in the autocorrelation domain and frequency domain.

Table 5.2 Variation of F1 & F2 estimation accuracy with SNR of differed noise types based on Praat and ACR Cepstrum

Method	Clean Speech (Hz)	Noise type	Frames within 5% deviation (%)					Frames
			20 dB	10 dB	0 dB	-10 dB	-20 dB	
Praat	F1 900±100	White	93	81	65	44	28	48
		Pink	89	75	53	33	11	
		Red	90	76	60	39	20	
	F2 1400±150	White	88	73	55	33	22	
		Pink	82	65	46	26	8	
		Red	84	67	50	30	14	
ACR Cepstrum	F1 700±150	White	93	84	77	55	33	32
		Pink	90	77	59	43	19	
		Red	90	80	63	47	27	
	F2 1200±220	White	88	76	50	41	28	
		Pink	83	69	51	33	15	
		Red	84	72	55	37	19	

5.5 Acoustical Speech Parameter- ACR MFCC

Mel-frequency cepstral coefficients (MFCCs) are the most widely used acoustic feature in speech recognition systems [212]. In speech recognition, knowing how we hear is more important than speaking in feature extraction. Humans' auditory perception is non-uniform, linear at low frequency (1000 Hz) and nonlinear after 1000 Hz. Thus, the recorded speech signal must be nonlinearly mapped into the perceived scale. Mel scale maps the measured frequency exactly like human auditory perception. From the computational perspective, a group of overlapped triangular bandpass filters that simulate the characteristics of the Mel scale is used and is shown in fig. 5.24. In MFCC, these filter banks are applied on the speech spectrum to get the Mel scale power spectrum. Then the log operator is used to get the log-filter-bank output. Finally, the discrete cosine transforms (DCT) is used to generate 12 MFCCs. The 13th parameter is the log energy computed from each speech segment. To capture the dynamic information of the speech lost during frame-by-frame analysis, temporal derivatives were captured as Δ MFCCs and $\Delta\Delta$ MFCCs. So, the final feature vector in each speech segment contains 13 MFCCs, 13 Δ MFCCs and 13 $\Delta\Delta$ MFCCs.

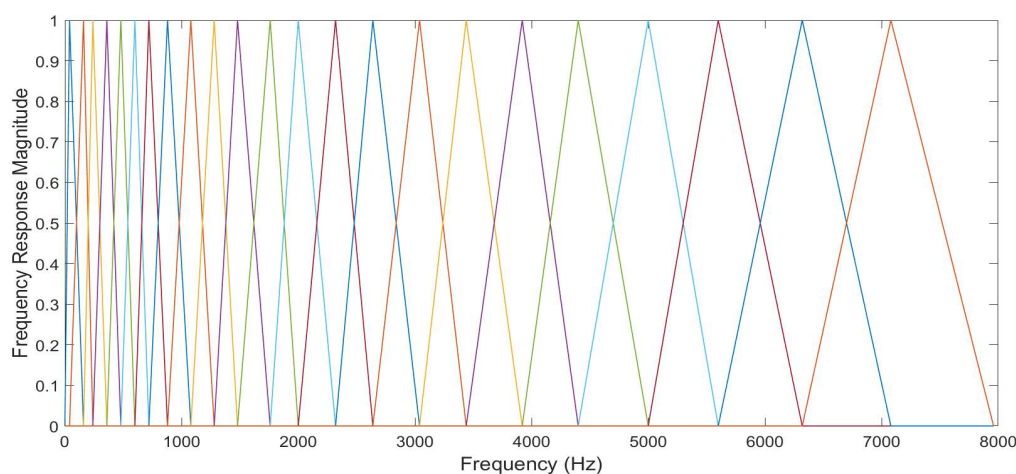


Fig. 5.24 Mel-scale Filter Bank

The MFCC features are good at recognising clean speech but not so good at recognising noisy speech. This is because the additive background noise significantly impacts the power spectral estimate utilised in MFCC computation, thereby lowering the recognition for noisy speech. This work proposes a noise-robust MFCC extraction method, termed ACR-MFCC, as shown in fig. 5.25.

To compare the performance of the proposed method (ACR-MFCC) with the MFCC Support Vector Machine (SVM) classifier [213] is used, which is discussed in Chapter 6. The database utilised for this comparison is five short vowels with 40 speech samples each which is also corrupted with three noisy signals. Table 5.3 shows the performance of ACR-MFCC and MFCC with SNR of different noise types. As discussed above, the proposed method and MFCC perform similarly in clean speech, but MFCC degrades its performance in noisy speech. MFCC perform similarly as the proposed one in three noisy speech up to 0 dB. Both methods are affected drastically in pink noise added speech signal than white Gaussian noise, but less in red noise. It is because the structure of the ACR spectrum is distracted heavily by pink noise, as in fig.5.12 (f). Thus, ACR MFCC is a valuable method for robust feature extraction when compared with MFCC.

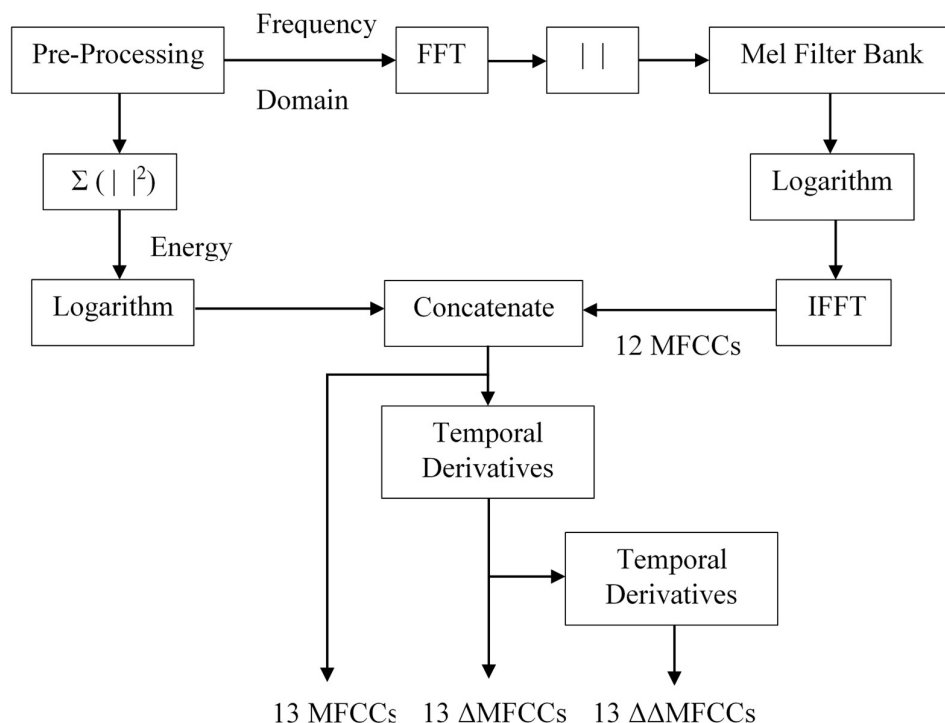


Fig. 5.25 Block Diagram of ACR-MFCC Feature Extraction Algorithm

Table 5.3 Performance of ACR MFCC and MFCC with SNR of differed noise type for Short Vowel Phoneme

Method	Noise type	Average Accuracy (%)					Clean Speech
		20 dB	10 dB	0 dB	-10 dB	-20 dB	
ACR MFCC	White	94	90	83	71	56	95
	Pink	93	88	83	70	63	
	Red	94	91	90	87	73	
MFCC	White	88	85	80	60	48	92
	Pink	88	83	78	63	52	
	Red	94	90	9	80	63	

The main element in extracting different speech features is the autocorrelation function (ACR). ACR has proven its noise robustness property in each feature extraction when compared with other prominent methods. Thus,

the three feature extraction methods can be viewed in a unified framework, as shown in fig. 5.26.

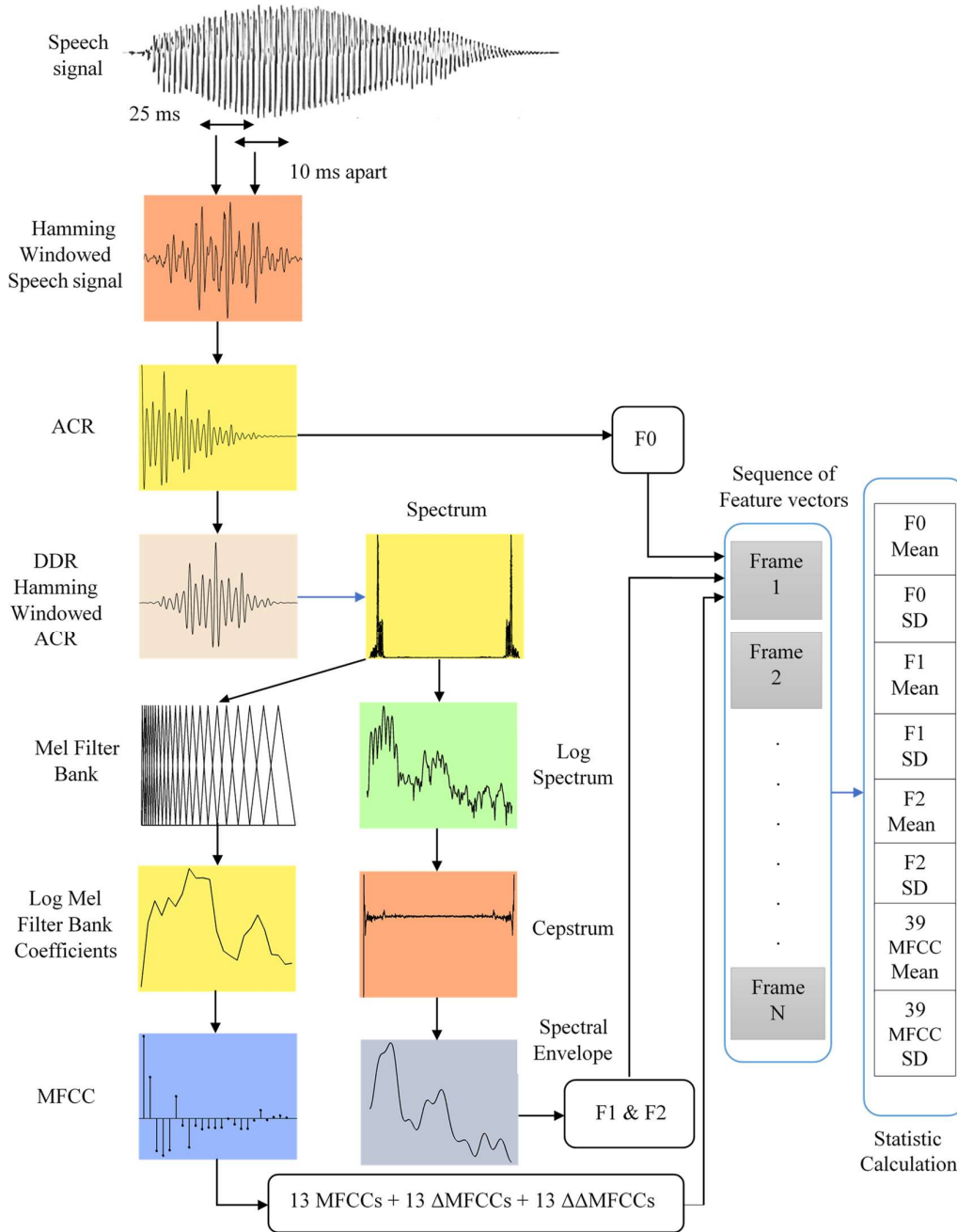


Fig. 5.26 Unified Frame work of Acoustical Feature Extraction

5.6 Conclusions

An Autocorrelation (ACR) based acoustical feature extraction process is presented. Features extracted are fundamental frequency (F0), formant frequencies (F1 and F2) and Mel-Frequency Cepstral Coefficients (MFCCs). This work treats the frequency estimated from each segmented frame of a “clean” speech signal as the true values for F0 and formants. The method’s performance is evaluated by analyzing the variation of corresponding true values from each frame of a noisy speech signal. The performance of each method has been evaluated without changing its default parameter. Fundamental frequency of the speech signal is estimated from the autocorrelation sequence and it outperforms the prominent methods like Praat, YIN, RAPT and PEFAC in all noisy conditions. The proposed method (ACR) performs outstanding, especially in white Gaussian noisy speech signals. To obtain the spectral information of speech signal in autocorrelation domain, Hamming window with double dynamic range is applied on ACR. The first two formant frequencies were extracted from the cepstral domain by incorporating the noise robustness nature of ACR. The proposed method (ACR Cepstrum) is compared with the Praat method. Even though ACR Cepstrum is inferior in terms of standard deviation, it is superior for noise robustness compared to Praat. The performance of both methods is affected mainly by pink noisy speech signal. A noise-robust MFCC method (ACR MFCC) is introduced, and an SVM classifier evaluates its performance. ACR MFCC surpasses MFCC in all noisy conditions but performs similarly in clean speech.

CHAPTER 6

AUDIO-VISUAL SPEECH RECOGNITION USING SUPPORT VECTOR MACHINE CLASSIFIER

6.1 Introduction

The human brain makes bimodal judgements based on circumstances, especially in a noisy background. Motivated from this, intelligent speech-based systems respond to a stimulus by analyzing and interpreting the information captured from microphones and cameras. The key component in achieving human-like decision making in such a system is the selection of audio and visual speech information integration strategy and decision making. Intense caring is needed to fuse the modalities and make a decision which requires linguistic knowledge and technical knowledge for the fine-tuning of parameters. This area has witnessed a vast technical outbreak to cope up with the wide usage of speech-based systems in different domains of life.

In this work, chapter 3 deals with creating an audio-visual speech database for the Malayalam language named “MOZHI”. It also addressed the issues related to database creation like audio and video speech segmentation and labelling process, corrupting the audio speech with different noise at different dB levels and audio-visual asynchrony. Chapter 4 displayed a detailed work on visual speech alone based on phoneme-to-viseme mapping and allophone-to-viseme mapping. Based on phoneme-to-viseme, visemes are represented by five consecutive frames. The DCT feature (appearance-based feature) has shown dominance over Geometric features (shape-based features) in visual speech analysis. Due to the random distribution of allophones in the allophone-to-viseme mapping, the visual equivalent of allophone (viseme) requires a detailed study in the linguistic domain. After that, chapter 5 analysis

the audio speech alone and extracts the acoustical speech parameters like fundamental frequency (F0), formant frequencies (F1 and F2) and ACR MFCC. It also discussed the performance of the proposed method in corrupted audio speech, which surpasses the traditional methods. A well-established linguistic foundation and computational update are at the need of any speech-based system. So, in the present scenario (in the rest of the work), phonemes and visemes are the centres of discussion. This will be the first effort to develop audio-visual speech recognition in Malayalam.

Even though the central problem in audio-visual speech recognition is an intelligent way of combining the audio and visual features, which is addressed in section 6.3.5. The key issue is, “which classifiers should be utilised to recognise uttered phonemes from audio and visual feature vectors?”. A *classifier* is a machine learning algorithm that maps the input data into specific classes based on certain rules to discover patterns and regularities hidden in the data. Machine learning algorithms are broadly classified into two classes: Unsupervised learning and Supervised learning. An unsupervised learning algorithm uses an unlabelled dataset to classify, revealing similarities and structures hidden in the data. Unsupervised learning is helpful for finding useful insights from the data. K-means clustering, Hierarchical Clustering are some of the popular unsupervised learning algorithms. A supervised learning algorithm uses training datasets from which they learn to classify similar unlabelled data. Naïve Bayes [214], Artificial Neural Network (ANN) [215], Decision Tree Classification [81], Support Vector Machine (SVM) [216], K-Nearest Neighbours (K-NN) [217] are some of the popular supervised learning algorithms. A classification model tries to draw some conclusions from the training data during the training mode, and it will predict the class labels for the new data during the testing mode. Classification models are grouped into two classes: Discriminative Models and Generative Models [218]. A discriminative model models the decision boundary between the data in the data space, while

a generative model explicitly models how data is distributed in the data space. The discriminative model learns the conditional probability distribution $P(Y|X)$, while the generative model learns the joint probability distribution $P(X,Y)$. Since the generative model relies on the Bayes theorem, it can handle more complex tasks than discriminative models. Hidden Markov Model (HMM) [219], Gaussian Mixture Model (GMM) [220], Bayesian Network [221] belongs to generative models. Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbours (KNN) are some of the discriminative models.

Hidden Markov Model (HMM) and its hybrid version (Coupled HMM [222], HMM/GMM [223], HMM/ANN [224], HMM/SVM [225]) are the most widely used in recent audio-visual speech recognition systems for modelling and recognizing speech. HMM's capacity to statistically model acoustic and temporal variation in speech is the primary reason for its popularity. However, the HMM-based speech recognition system suffers from performance loss due to a disparity between training and testing settings. One of the main problems related to the HMM-based speech recognition system is its implementation side compared to discriminative models like SVM for a multi-modal phoneme level task. Another powerful classifier widely used in speech recognition tasks is Artificial Neural Network (ANN). The neural network algorithms strive to reduce the difference between the desired output and the network's generated output [226]. In contrast, the training of an SVM classifier relies on maximizing the margins between the borders of classes. Unlike certain HMM modifications that only minimise the empirical risk on the training set, SVMs also minimise the structural risk, resulting in improved generalisation even with less training data. The exceptional generalisation properties of SVM is due to the maximised distance, termed as the margin. In the presence of noise, the maximum margin solution allows SVM to outperform most nonlinear classifiers, which has long been a challenge in speech recognition [227].

An audio-visual speech recognition system using support vector machine classifiers is presented in detail. Viseme is utilized to recognize the underlined phoneme in intense background noise. The different problems which arise before and after fusing the audio and visual speech information are addressed in detail. The rest of the chapter is organized as follows. Section 6.2 describes the fundamentals of the SVM classifier. Section 6.3 presents a detailed outline of the audio-visual speech information fusion process. Section 6.4 introduces the proposed audio-visual speech recognition system. Section 6.5 discusses the experimental part and section 6.6 concludes the work.

6.2 Support Vector Machine (SVM) Classifier

SVM is a supervised machine learning algorithm that has been used in many real-world tasks, especially for classification purposes. SVM is based on the statistical learning theory of Vapnik in 1974 and quadratic programming [226]. SVM has got different modifications which makes it still an active algorithm with relatively simple concepts. SVM was inherently a binary nonlinear classifier capable of distinguishing whether the input data belongs to class 1 or 2. It defines an optimum hyperplane that separates different classes with the maximum margin between the boundary points (support vectors) in a higher dimensional feature space. Decision boundaries will be of different size and shape depending upon the nature of the problem. For separable cases, it will be a line in 2-dimension, a plane in 3-dimension and a hyperplane in N-dimension. For non-separable cases, SVM uses a trick named kernel trick that transforms the non-separable n-dimensional sequence of feature vectors to linearly separable higher dimensional kernel feature space ϕ .

Empirical Risk Minimisation (ERM) is the goal of any classifier: it reduces the number of misclassifications in the training set. In machine learning, a generalised model must be chosen from a finite data set, which leads to the problem of overfitting, i.e., when the model becomes overly tuned to the

training set's characteristics and hence fails to generalise to fresh data. This problem is solved using the Structural Risk Minimisation (SRM) principle, which balances the model's complexity against its ability to match the training data. The optimal hyperplane maximises the margin while minimising the empirical risk [213], [228]. As shown in fig. 6.1, there will be hyperplanes that can accurately categorise all data points, resulting in zero empirical risk. However, H1 is preferred over H2 because H1 has a higher margin than H2 and is thus less prone to overfitting. In other words, a linear SVM can be configured to learn a hyperplane that can tolerate a limited number of non-separable data points.

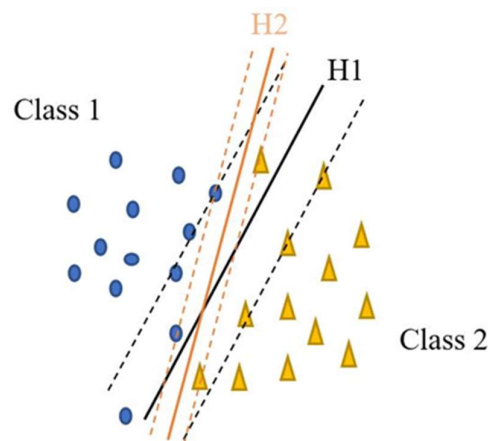


Fig. 6.1 Hyperplanes for Classifying the Non-separable Data points

The amount of non-separable data items that SVM takes into account should be limited. The decision boundaries with a very high margin will create a chance of misclassification of many data points. As a result, there should be a trade-off between margin width and misclassification error. To avoid this issue, a penalization term is added to the optimal condition, as shown below.

$$M^* = \arg \min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^p \xi_i$$

$$\text{Subject to } \min y_i(w^T x_i + w_0) - 1 + \xi_i \geq 0 \tag{6.1}$$

$$\xi_i \geq 0, i = 1, 2, \dots, p$$

where w is a vector of model parameter which defines the decision boundary, C is the penalty parameter, which represents misclassification or error term, ξ_i is the positive slack variables. The misclassification or error term informs the SVM optimization about the acceptable level of error. A smaller C value results in a small margin, while a bigger C value results in a larger margin. The parameter C establish an understanding between ERM and SRM using an iterative search process (Grid search). After the SVM has been trained with training data, the number of support vectors determines the classifier's complexity.

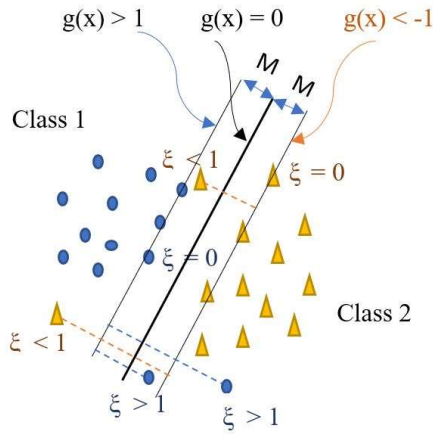


Fig. 6.2 Linear Separating Hyperplane for the Non-separable Data points

In real-world classification problems, the data points are highly overlapped, making the above-stated classifier not yield accurate classification. In other words, the decision boundary may not be a linear or nonlinear hyperplane instead a hypersurface. The input data is transferred into a higher-dimensional space to address this problem, as the data's dimension does not determine the classifier's success. Then look for a linear decision boundary to

separate the higher-dimensional data that has been transformed. A method known as the kernel trick gives a revolutionary answer to the issue mentioned above. The kernel approach conveys the data only through pairwise similarity comparisons between the original data observations x rather than explicitly applying the transformations $f(x)$ and expressing the data by these modified coordinates in the higher dimensional feature space [229]. In short, the kernel function accepts lower-dimensional inputs and provides the dot product of converted vectors in higher-dimensional space. The dot product is frequently used to compare two input vectors. Kernel function $K(x_i, x_j)$ is a real function defined on R such that there exist a function $\phi: R^m \rightarrow R^n$, where $n > m$

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \rightarrow x_i \cdot x_j \quad (6.2)$$

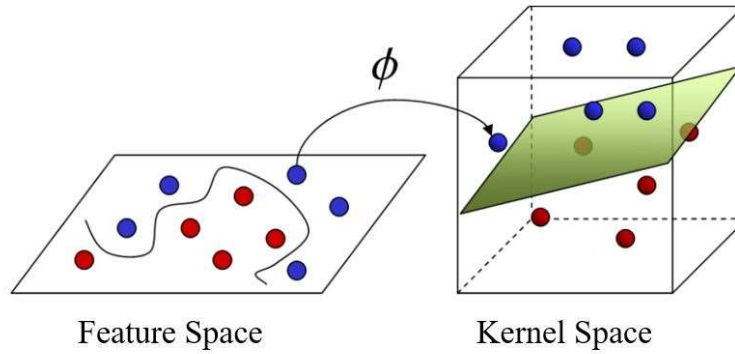


Fig. 6.3 Transformation of Non-Separable Data points in Feature Space to Separable Data points in Kernel Space

Some of the most widely used kernel functions are linear, polynomial, Gaussian radial basic and sigmoid. The kernel function which provides a better result in this work is the Gaussian radial basic kernel. The classification result using different kernels is shown in table1. The database used for this is task is the MFCC feature vector extracted from a clean speech by considering all attributes discussed in section 6.3.

$$\text{Gaussian Radial Basic Kernel: } K_G(x_i, x_j) = \exp\left(-\frac{|x_i - x_j|^2}{2\sigma^2}\right) \quad (6.3)$$

$$\text{Gamma} = \frac{1}{2\sigma^2}$$

Table 6.1 Performance of SVM Classifier for Different Kernel Functions

Kernel Type	Accuracy (%)
Linear	89
Polynomial (degree=3)	91
Gaussian Radial Basic	96
Sigmoid	90

6.3 Problems and Strategies: Before and After Audio-Visual Integration

This work has discussed a brief background of the SVM classifier so far. Identifying optimal hyperplane with the maximum margin guarantees better generalization makes SVM a very promising tool for the speech recognition perspective. In addition, its structural risk minimization (SRM) ability will remarkably improve the robustness of speech-based systems, especially in a noisy environment. However, the presented SVM still needs fine-tuning for the recognition task while considering the fusion of different modalities. The following session addresses different factors that need prime attention during audio and visual feature integration. By analyzing all aspects, a proposed work is presented in session 6.4.

6.3.1 Multi-class Problems

SVM discussed deals with the classification of binary classes, but a speech recognition system is a many class problem. The multi-class classification problem is handled by combining many binary SVM classifiers. It can be achieved by two approaches: One-vs-all and one-vs-one classification.

In one-vs-all or one-vs-rest classification, each class is compared with all rest of the classes. The number of binary classifiers needed is the same as that of the number of class labels in the data set. Data from one class is positive for each binary classifier, and all other classes have a negative attribute. Thus, each model predicts a probability-like score. The class index with the most significant score is then used to predict a class.

In one-vs-one classification, each class is confronted with each other classes separately. The number of classifier models needed in this method is $n(n-1)/2$, where n is the number of classes in the problem. For implementing this approach, datasets are split into one binary classification dataset for each pair of classes. In this work, the one-vs-all classification approach is used, as shown in fig. 6.4.

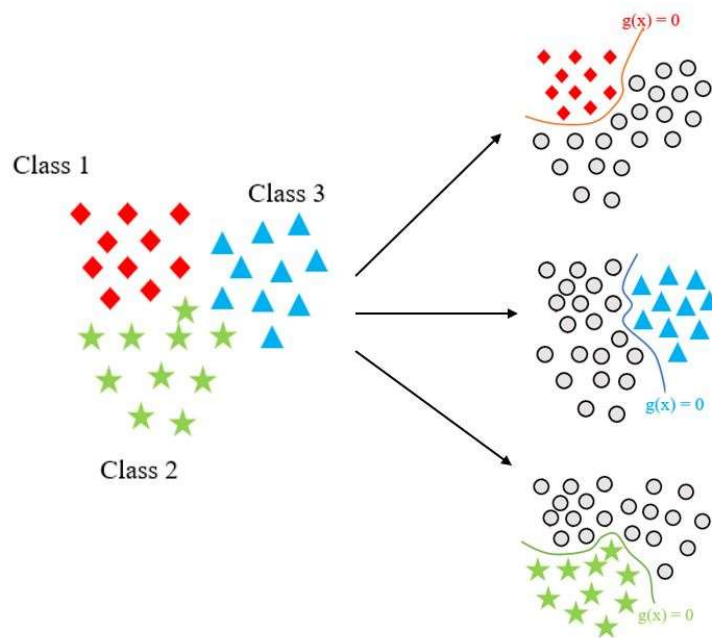


Fig. 6.4 One-vs-All SVM Classifier

6.3.2 Audio-Visual Integration Approach

The task of fusing audio and visual speech information is one of the major components in determining the overall performance of the speech recognition system. Incorporating the visual speech information along with acoustical speech provides better recognition accuracy than an individual one. The way of integration profoundly depends on the problem under study. Thus, it is better to introduce the mode of integration before presenting the data structure (as discussed in 6.3.3). Before fusing the data from both domains, it is necessary to address one of the problems associated with audio-visual integration, i.e., lack of one-to-one correspondence between phonemes and visemes. In general, visemes carry comparatively less information than phonemes; thus, the number of visemes are very less. In Malayalam, 50 phonemes are mapped to 14 visemes. Thus, in this work (Phoneme recognition), data from both streams (audio and video) should be presented either as 50 or 14 classes. Presenting the visual speech information in 50 classes may degrade the overall performance of the system. Thus, a strategy that initially depends on the visual speech, thereby achieving robustness to acoustic noise, is presented.

Early integration (Feature fusion) and Late integration (Decision fusion) are the most widely used approaches in audio-visual integration process [184]. In Early Integration, audio and visual features extracted from the respective domain are combined, then used to recognise the corresponding phoneme. Since the physical characteristics of both domains are different, an intelligent way is required to combine both features where audio-visual asynchrony problem arises. In Late integration, both audio and visual domain has individual recogniser whose outputs are combined to get the overall results. In this approach, the fusion of the two domains can be controlled by weighting the domain. At high SNR (Signal-to-Noise Ratio), the audio domain is enough to

recognise the corresponding phoneme. At low SNR, the visual domain alone is adequate to provide better recognition. In this work, a hierarchical approach [230] is utilised to obtain the mentioned strategy.

The hierarchical approach computes the likelihood for a phoneme in two phases. The visual speech data is used in the first phase, with 14 groups (viseme). The most likely viseme class is determined from phase 1. In the second phase, audio and visual speech data are divided into 50 classes to determine the most likely phoneme within the viseme class determined in the first phase. When the most likely viseme class in phase 1 has only one phoneme (viseme class 6 and 7 as in table 4.2), then the second phase is skipped as it is not required. In the second phase, the phonemes within the most likely viseme class are investigated. The computational overhead involved in this phase is less. In practical view, there will be three datasets (which is discussed in section 6.3.3): one with video speech data grouped into 14 classes (Visemes) and second with audio speech data grouped into 50 classes (Phonemes) and third with visual speech data grouped into 50 classes (visual speech of 50 phonemes).

Another important component of the fusion process is the weighting of each stream (audio and video), which allows for choosing to increase or decrease one input's role based on its decision-making efficiency. Proper integration weight ensures better improvement in the performance of AVSR than audio-only and visual-only speech recognition. The sum of the stream weights must equal one. Various methods for assigning weight values to streams have been proposed in the literature. These weight values can be assigned based on the quality of test data [141], [222], train data [231], [232], or both [184]. The stream weight is manually estimated in some cases. As a result, the challenge of determining stream weight under various SNR situations remains under investigation. In this work, the weight value is estimated automatically depending on the noise level. For this, the reliability of

each stream is measured from the dispersion of the posterior probability (outputs of the classifier) [233], [234]. The output of the SVM classifier is not a probability [235]. Building a classifier to produce a posterior probability $P(class/input)$ is useful. The distance of x from the separating hyperplane for a test sample x is represented by SVMs' output, $f(x)$. While the class prediction ($y = +1$ or $y = -1$) is determined by the sign of the SVM output, the magnitude of the SVM output can represent the level of confidence in that prediction. As a result, the SVM output cannot be directly translated into a probability value used to estimate confidence. To tackle this problem, Platt [236] used a parametric model to directly fit the posterior $P(y = +1 / f(x))$ without estimating the conditional density $p(f(x) / y)$ for each label y , as

$$P(y = +1 / f(x)) = \frac{1}{1 + e^{-[Af(x)+B]}} \quad (6.4)$$

The parameters A and B of Eq. 6.4 are fitted using maximum likelihood estimation from a training set. Thus, the reliability of a stream is estimated as

$$\sigma = \frac{1}{N} \sum_{i=1}^N [(\max \log P(O/F^i) - \log P(O/F^i))] \quad (6.5)$$

where $P(O / F^i)$ is posterior probability of the phoneme/viseme (O) given the corresponding feature vector F^i and N is the number of classes. Then, the stream weight for audio stream and video stream is estimated as

$$\lambda_A = \frac{\sigma_A}{\sigma_A + \sigma_V} \quad (6.6)$$

$$\lambda_V = 1 - \lambda_A \quad (6.7)$$

where σ_A and σ_V are the dispersion of audio and video outputs probabilities respectively. After the recognition of phonemes and visemes from separate classifiers, their outputs are combined by the weighted product rule to recognize the uttered phoneme as shown in Eq. 6.7.

$$\text{Phoneme} = \arg \max_i [P(O_V / F_A^i)^{\lambda_A} \times P(O_V / F_V^i)^{\lambda_V}] \quad (6.8)$$

The proposed Hierarchical approach in this work is shown in fig. 6.5 and its corresponding equations are given as

$$\text{Phase 1: Viseme} = \arg \max_i P(O_V / F_V^i), i = 1, 2, \dots, 14 \quad (6.9)$$

$$\text{Phase 2: Phoneme} = \arg \max_j [P(O_A / F_A^j)^{\lambda_A} \times P(O_V / F_V^j)^{\lambda_V}]$$

$$j \text{ is the phonemes in the selected viseme} \quad (6.10)$$

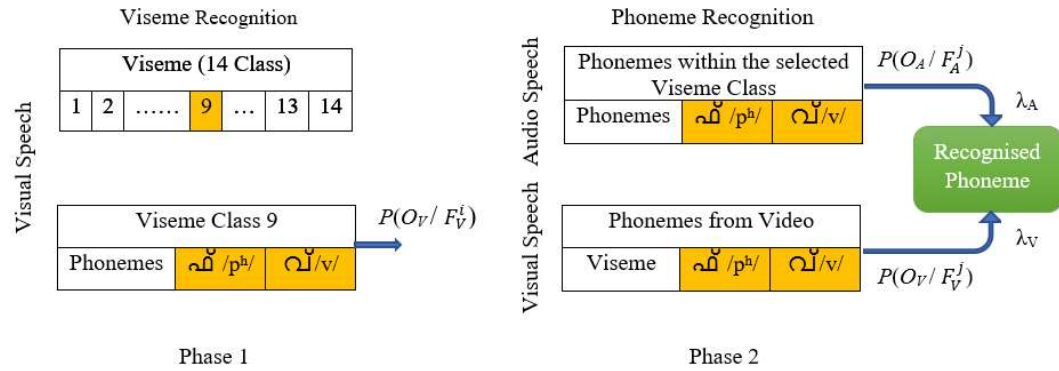


Fig. 6.5 Hierarchical Approach for Audio Visual Integration

In a real-time context, however, a minor change in the appearance of a phoneme can look almost identical to the visual appearance (viseme) of another phoneme. As a result, a modified phoneme-to-viseme mapping must be established, as shown in table 6.2, which clusters the phonemes with a broad visual speech appearance. The first three visemes in table 4.2 are front and back vowels and belong to the first broad viseme, which has a common visual attribute of mouth wideness. The back vowels belong to the second broad viseme and have a visual characteristic of lip roundness. Table 6.2 does not consider visemes 6 and 7 in table 4.2 because they have a single phoneme that is skipped in the second phase of the hierarchical approach. Visemes 8 and 9 in table 4.3 have the characteristic visual appearance of lips touching, classified

as the third broad viseme. Dental, velar, glottal, alveolar, retroflex, and palatal visemes are among the most confusing visemes of consonant classes. A barely visible tongue plays their linguistic characteristics. As shown in fig. 6.15, the visual equivalents of dental (viseme 10 in table 4.3) and velar and glottal (viseme 11 in table 4.3) were grouped as the fourth broad viseme based on their capacity to recognise better than others. The remaining consonant viseme classes were assigned to the fifth broad viseme, as shown in table 6.2.

Table 6.2 Modified Phoneme-to-Viseme Mapping

Broad Viseme	Viseme Class	Phonemes
1	Front, High – Vowel Front, Mid – Vowel Central, Low – Vowel	ഇ /i/, റു /i:/ എ /e/, ഏ /e:/ അ /a/, ആ /a:/
2	Back, High – Vowel Back, Mid – Vowel	ഉ /u/, ഊ /u:/ ഓ /o/, ഔ /o:/
3	Bilabial - Plosive-voiced and voiceless unaspirated, Nasal Bilabial - Plosive-voiceless aspirated and Labiodental	പ് /p/, ബ് /b/, ഭ് /b ^h /, മ് /m/ ഫ് /ph/, വ് /v/
4	Dental Velar, Glottal	ത് /t/, ത് /th/, ദ് /d/, ധ് /dh/, ന് /n/ ക് /k/, ക് /kh/, ഗ് /g/, ഘ് /gh/, ണ് /ŋ/, ഹ് /h/
5	Alveolar Retroflex Palatal	റ് /r/, ന് /n/, സ് /s/, ര് /r/, ര് /r/, ല് /l/ ട് /t/, ത് /th/, ഡ് /d/, ഡ് /dh/, ണ് /ŋ/, ഷ് /ʃ/, ഴ് /l/, ഴ് /z/ ച് /c/, ച് /ch/, ജ് /j/, യ് /j ^h /, ഞ് /ɲ/, ശ് /ʃ/, യ് /y/

As the phoneme-to-viseme mapping is modified to address the visual speech appearance in a real-time scenario, the corresponding hierarchical approach is also modified as fig. 6.6. In the first phase, the most likely viseme is identified from the 14 visemes as in fig. 6.5. In addition, other visemes which belong to the same broad viseme class of the most likely viseme were also selected. For instance, the most likely viseme in the first phase is viseme 9, which correspond to the visual representation of the phonemes \mathfrak{p}^h /ph/ and \mathfrak{v} /v/. However, in the broad viseme class, viseme 9 belongs to broad viseme class 3, containing viseme 8. Thus, the main aim of phase 2 is to identify the most likely phonemes of broad viseme class 3 \mathfrak{p} /p/, \mathfrak{b} /b/, \mathfrak{bh} /bh/, \mathfrak{m} /m/, \mathfrak{p}^h /ph/ and \mathfrak{v} /v/ instead of most likely phoneme of viseme 9 \mathfrak{p}^h /ph/ and \mathfrak{v} /v/ as in fig. 6.5. In the second phase, the probability of these phonemes was estimated in the audio and visual classifiers separately. Then the most likely phoneme is identified by fusing both classifiers with proper stream weight depending on the noisy acoustical condition. The corresponding equation of phase 2 in the modified hierarchical approach is given in Eq. 6.11. Since phase 1 is dedicated to identifying the most viseme from 14 viseme, its corresponding equation is Eq. 6.9.

$$\text{Phase 2:} \quad \text{Phoneme} = \arg \max_k [P(O_A / F_A^k)^{\lambda_A} \times P(O_V / F_V^k)^{\lambda_V}]$$

k is the phonemes in the selected broad viseme (6.11)

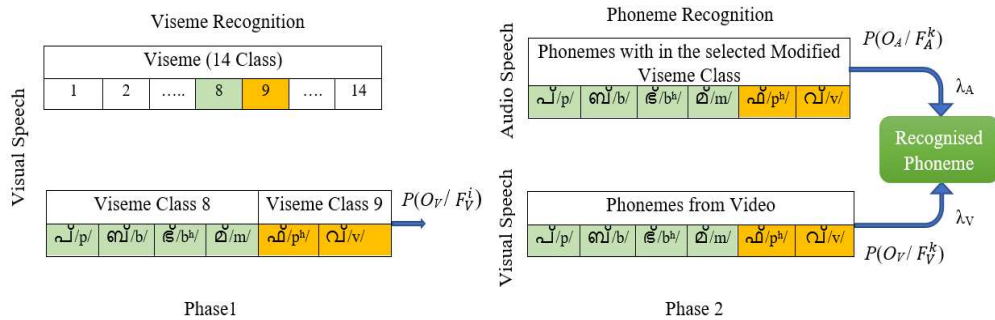


Fig. 6.6 Modified Hierarchical Approach for Audio Visual Integration

6.3.3 Dataset Preparation and Hyper Parameter Estimation

A dataset is a collection of different variables' values. Training and testing sets are created after the dataset has been prepared. The training set is inserted into the machine learning algorithm to create a predictive model. The model then predicts the labels in the testing data. The model's performance is then evaluated with precision, recall, F1-score, and accuracy.

In this work, audio speech samples were collected from 23 speakers uttering 50 utterances (vowels, diphthongs and consonant phonemes) repeating three times each. Thus, each speaker has a total of 150 utterances/observations. These audio speech samples were corrupted with three different noises with noise levels 20 dB, 10 dB, 0 dB, -10 dB and -20 dB. The visual speech was collected simultaneously with the audio speech. Since the viseme for each phoneme is represented by five consecutive frames, later, the visual representation of each phoneme is grouped based on phoneme-to-viseme mapping as in tables 4.2 and 4.3. Thus, the visual speech sample consists of 750 frames (50 phonemes x 3 repetition x 5 frames) for each speaker.

To implement the proposed modified hierarchical approach, three datasets were needed. For phase 1, the visual representation of each phoneme was grouped into 14 classes, namely visemes. Since each viseme contains an unequal number of phonemes, resulting in an unbalanced dataset with minimum and maximum of 69 and 552 observations, respectively, as shown in fig. 6.7. Random oversampling is carried out to make it a balanced dataset. In phase 1, the training and testing set consists of 14 classes. For phase 2, two datasets are needed: one having audio speech of 50 phonemes and the other having visual speech of 50 phonemes. The distribution of observations for phase 2 is shown in fig.6.8, which is a balanced dataset with 69 observations for each class. For this phase, the training set consists of 50 classes, and the testing set consists of 5 classes representing the five broad visemes.

Even though the fusion strategies and datasets were well established, specific issues associated with the classifier like selecting optimum hyperparameters, data splitting, overfitting and unbalanced dataset are to be addressed in detail. A machine learning model is a mathematical model that requires the learning of various parameters from data. Another parameter, Hyperparameters, on the other hand, cannot be learned directly from the usual training procedure. Instead, they usually are addressed before the start of the training procedure. These parameters define the model's key characteristics, such as its complexity and learning rate. Models might have many hyperparameters, thus choosing the best one is a research problem. Typically, the values for these hyperparameters are assigned at random and compare which ones produce the best results. Randomly choosing the model's parameters, on the other hand, can be tedious. Instead of choosing parameter values at random, a preferable method would be to create an algorithm that identifies the ideal parameters for a model automatically. Grid search is one of the most effective algorithms for optimising hyperparameters. The machine learning model is examined for a range of hyperparameter values using the Grid Search method. Then, it seeks the optimal combination of hyperparameters.

The penalty parameter, C and Gaussian Radial Basic Kernel parameter, γ are the hyperparameters for support vector machines [237]. The misclassification or error term, denoted by the letter C , tells the SVM optimization how much error can be tolerated. For example, a small-margin hyperplane is created by a lower value of C , while a larger-margin hyperplane is created by a higher value of C . γ decides how much curvature is needed for the decision boundary. A higher value of γ implies more curvature, and a lower value of γ means less curvature. Thus, there will be a trade-off between the misclassification term and the decision boundary. The values assigned to C and γ in grid search are $10^0, 10^1, 10^2, 10^3, 10^4$ and 10^5 and $10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$ and 10^{-5} , respectively. In this work, Nested 5-fold stratified cross-validation is used to search for the best value for tuning hyperparameter, which is discussed in the next paragraph. The grid search for

datasets of 14 visemes, visual representation of 50 phonemes and 50 phonemes are shown in table 6.3, 6.4 and 6.5, respectively. The best combination of hyperparameter combination is not chosen based on the highest score. Instead, the parameter set having the highest, most stable score value is selected as the optimum value. The highest score belongs to a parameter set with C equal to 10^4 , resulting in high computational cost. For all datasets, the optimum combination of hyperparameters is $C = 10^2$ and $\gamma = 10^{-2}$. This optimum value is chosen throughout the rest of this work.

In a machine learning process, the dataset is divided into training and test datasets; the training dataset is used to train the model, while the test dataset is used to examine the performance of a model. The training dataset typically accounts for two-thirds of the overall dataset, with the remainder used for testing. As a result, this method may neglect some of the most informative data, resulting in a more significant bias. The solution to this problem is to use k-fold cross-validation. The k-fold cross-validation procedure divides a limited dataset into k folds (a subset of the dataset). Each k fold can be used as a testing set, while the rest can be used as a training set. The method is repeated until each set has been utilised at least once for training and testing. Finally, the mean performance is presented after a total of k models are fitted and tested on the k test sets. This technique is used to select model hyperparameters to set up each model and select models that have been configured. Unfortunately, utilising the test set for both model selection and estimate overfits the test data, resulting in an optimistic bias in estimation [238]. To overcome this issue, model selection and evaluation must be done separately, which is implemented using a modified version of cross-validation termed Nested Cross-Validation.

Table 6.3 Grid search of 14 Visemes

		C													
Gamma		10 ⁰	10 ¹	10 ²	10 ³	10 ⁴	10 ⁵		10 ⁰	10 ¹	10 ²	10 ³	10 ⁴	10 ⁵	
		10 ⁰	0.170	0.188	0.188	0.188	0.188	0.188	10 ⁰	0.168	0.187	0.187	0.187	0.187	0.187
		10 ⁻¹	0.405	0.503	0.503	0.503	0.503	0.503	10 ⁻¹	0.405	0.495	0.495	0.495	0.495	0.495
		10 ⁻²	0.473	0.752	0.806	0.806	0.806	0.806	10 ⁻²	0.486	0.743	0.801	0.801	0.801	0.801
		10 ⁻³	0.359	0.539	0.744	0.885	0.912	0.912	10 ⁻³	0.368	0.544	0.748	0.877	0.925	0.925
		10 ⁻⁴	0.186	0.367	0.534	0.673	0.801	0.900	10 ⁻⁴	0.185	0.376	0.536	0.670	0.804	0.908
	10 ⁻⁵	0.159	0.188	0.369	0.535	0.645	0.726	10 ⁻⁵	0.159	0.185	0.376	0.539	0.655	0.746	
		1 st Iteration							2 nd Iteration						
	10 ⁰	0.167	0.188	0.188	0.188	0.188	0.188	10 ⁰	0.169	0.186	0.186	0.186	0.186	0.186	
	10 ⁻¹	0.398	0.493	0.493	0.493	0.493	0.493	10 ⁻¹	0.409	0.503	0.503	0.503	0.503	0.503	
	10 ⁻²	0.474	0.728	0.803	0.803	0.803	0.803	10 ⁻²	0.481	0.740	0.802	0.802	0.802	0.802	
	10 ⁻³	0.364	0.545	0.735	0.866	0.917	0.918	10 ⁻³	0.350	0.542	0.744	0.867	0.927	0.928	
	10 ⁻⁴	0.181	0.368	0.541	0.659	0.777	0.911	10 ⁻⁴	0.181	0.360	0.532	0.671	0.789	0.898	
	10 ⁻⁵	0.159	0.183	0.367	0.540	0.646	0.717	10 ⁻⁵	0.159	0.183	0.361	0.533	0.650	0.712	
		3 rd Iteration							4 th Iteration						
	10 ⁰	0.170	0.186	0.186	0.186	0.186	0.186								
	10 ⁻¹	0.408	0.507	0.507	0.507	0.507	0.507								
	10 ⁻²	0.465	0.750	0.801	0.801	0.801	0.801								
	10 ⁻³	0.356	0.538	0.728	0.874	0.927	0.927								
	10 ⁻⁴	0.187	0.361	0.531	0.672	0.774	0.898								
	10 ⁻⁵	0.159	0.188	0.362	0.529	0.648	0.718								
		5 th Iteration													

Table 6.4 Grid search of 50 Visemes

		C													
Gamma		10 ⁰	10 ¹	10 ²	10 ³	10 ⁴	10 ⁵		10 ⁰	10 ¹	10 ²	10 ³	10 ⁴	10 ⁵	
	10 ⁰	0.019	0.018	0.018	0.018	0.018	0.018	10 ⁰	0.013	0.010	0.010	0.010	0.010	0.010	0.010
	10 ⁻¹	0.029	0.030	0.030	0.030	0.030	0.030	10 ⁻¹	0.033	0.034	0.034	0.034	0.034	0.034	0.034
	10 ⁻²	0.076	0.208	0.302	0.302	0.302	0.302	10 ⁻²	0.077	0.218	0.316	0.316	0.316	0.316	0.316
	10 ⁻³	0.090	0.138	0.351	0.692	0.709	0.709	10 ⁻³	0.094	0.141	0.367	0.689	0.714	0.714	0.714
	10 ⁻⁴	0.056	0.090	0.147	0.357	0.767	0.854	10 ⁻⁴	0.043	0.095	0.150	0.360	0.775	0.862	0.862
	10 ⁻⁵	0.056	0.056	0.090	0.148	0.358	0.764	10 ⁻⁵	0.042	0.042	0.096	0.151	0.359	0.762	0.762
	1 st Iteration							2 nd Iteration							
	10 ⁰	0.018	0.014	0.014	0.014	0.014	0.014	10 ⁰	0.025	0.018	0.018	0.018	0.018	0.018	0.018
	10 ⁻¹	0.026	0.030	0.030	0.030	0.030	0.030	10 ⁻¹	0.029	0.031	0.031	0.031	0.031	0.031	0.031
10 ⁻²	0.079	0.219	0.298	0.298	0.298	0.298	10 ⁻²	0.082	0.212	0.303	0.303	0.303	0.303	0.303	
10 ⁻³	0.089	0.137	0.359	0.692	0.717	0.717	10 ⁻³	0.093	0.148	0.363	0.694	0.715	0.715	0.715	
10 ⁻⁴	0.053	0.093	0.147	0.357	0.747	0.848	10 ⁻⁴	0.065	0.092	0.158	0.357	0.763	0.854	0.854	
10 ⁻⁵	0.040	0.040	0.094	0.149	0.352	0.736	10 ⁻⁵	0.065	0.065	0.092	0.159	0.362	0.751	0.751	
3 rd Iteration							4 th Iteration								
10 ⁰	0.020	0.020	0.020	0.020	0.020	0.020	10 ⁰	0.020	0.020	0.020	0.020	0.020	0.020	0.020	
10 ⁻¹	0.031	0.033	0.033	0.033	0.033	0.033	10 ⁻¹	0.031	0.033	0.033	0.033	0.033	0.033	0.033	
10 ⁻²	0.085	0.210	0.307	0.307	0.307	0.307	10 ⁻²	0.085	0.210	0.307	0.307	0.307	0.307	0.307	
10 ⁻³	0.093	0.143	0.354	0.696	0.717	0.717	10 ⁻³	0.093	0.143	0.354	0.696	0.717	0.717	0.717	
10 ⁻⁴	0.056	0.092	0.153	0.358	0.781	0.860	10 ⁻⁴	0.056	0.092	0.153	0.358	0.781	0.860	0.860	
10 ⁻⁵	0.056	0.056	0.093	0.155	0.354	0.754	10 ⁻⁵	0.056	0.056	0.093	0.155	0.354	0.754	0.754	
5 th Iteration															

Table 6.5 Grid search of 50 Phonemes

		C													
Gamma		10 ⁰	10 ¹	10 ²	10 ³	10 ⁴	10 ⁵		10 ⁰	10 ¹	10 ²	10 ³	10 ⁴	10 ⁵	
	10 ⁰	0.156	0.188	0.188	0.188	0.188	0.188	10 ⁰	0.140	0.180	0.180	0.180	0.180	0.180	0.180
	10 ⁻¹	0.772	0.796	0.796	0.796	0.796	0.796	10 ⁻¹	0.788	0.802	0.802	0.802	0.802	0.802	0.802
	10 ⁻²	0.771	0.905	0.915	0.915	0.915	0.915	10 ⁻²	0.792	0.908	0.909	0.909	0.909	0.909	
	10 ⁻³	0.435	0.778	0.890	0.903	0.903	0.903	10 ⁻³	0.415	0.779	0.897	0.902	0.902	0.902	
	10 ⁻⁴	0.081	0.433	0.768	0.887	0.902	0.903	10 ⁻⁴	0.082	0.423	0.762	0.894	0.902	0.902	
	10 ⁻⁵	0.077	0.079	0.432	0.766	0.886	0.904	10 ⁻⁵	0.080	0.080	0.423	0.759	0.893	0.900	
1 st Iteration							2 nd Iteration								
10 ⁰	0.145	0.171	0.171	0.171	0.171	0.171	10 ⁰	0.160	0.194	0.194	0.194	0.194	0.194	0.194	
10 ⁻¹	0.766	0.788	0.788	0.788	0.788	0.788	10 ⁻¹	0.766	0.794	0.794	0.794	0.794	0.794	0.794	
10 ⁻²	0.765	0.890	0.899	0.899	0.899	0.899	10 ⁻²	0.780	0.892	0.901	0.901	0.901	0.901		
10 ⁻³	0.409	0.752	0.880	0.891	0.891	0.891	10 ⁻³	0.422	0.763	0.882	0.894	0.894	0.894		
10 ⁻⁴	0.111	0.420	0.746	0.871	0.889	0.900	10 ⁻⁴	0.095	0.422	0.753	0.878	0.893	0.893		
10 ⁻⁵	0.106	0.109	0.421	0.742	0.870	0.888	10 ⁻⁵	0.091	0.094	0.423	0.750	0.875	0.893		
3 rd Iteration							4 th Iteration								
10 ⁰	0.135	0.168	0.168	0.168	0.168	0.168	10 ⁰	0.135	0.168	0.168	0.168	0.168	0.168		
10 ⁻¹	0.783	0.801	0.801	0.801	0.801	0.801	10 ⁻¹	0.783	0.801	0.801	0.801	0.801	0.801		
10 ⁻²	0.769	0.898	0.902	0.902	0.902	0.902	10 ⁻²	0.769	0.898	0.902	0.902	0.902	0.902		
10 ⁻³	0.414	0.762	0.884	0.897	0.897	0.897	10 ⁻³	0.414	0.762	0.884	0.897	0.897	0.897		
10 ⁻⁴	0.094	0.420	0.747	0.880	0.895	0.895	10 ⁻⁴	0.094	0.420	0.747	0.880	0.895	0.895		
10 ⁻⁵	0.091	0.094	0.417	0.747	0.877	0.894	10 ⁻⁵	0.091	0.094	0.417	0.747	0.877	0.894		
5 th Iteration															

The nested k-fold cross-validation has inner loop cross-validation nested within the outer cross-validation. The process is commonly referred to as double cross-validation since it uses two cross-validation loops. Model selection/hyperparameter tuning (validation dataset) is performed by the inner loop, while performance evaluation is done by the outer loop (test dataset). As a result, there are three datasets in nest cross-validation: a training dataset, a validation dataset, and a testing dataset. Each training dataset (inner training fold) is fed into a hyperparameter-optimisation process like grid search, then tested on the validation dataset. Estimate the average metrics score for each hyperparameter configuration, then select the best hyperparameter with the highest consistent score. Then, a model is trained with the best hyperparameter on the train and validation dataset collectively termed as outer training fold. Next, evaluate its performance on the test dataset (outer testing fold) and save the score for each iteration. At last, the mean score is estimated for all iterations. Because the model selection process is done separately, it has no chance of overfitting the dataset because it is only exposed to a portion of the dataset provided by the outer cross-validation approach.

Another issue associate with nested k-fold cross-validation is to handle a significant imbalance in the distribution of the target classes. In situations where there is a significant class imbalance, a randomly generated fold might not appropriately represent the minor class. For instance, viseme class 13 in the table. 4.3 comprises eight phonemes. While splitting the dataset into k-folds, it is essential to ensure that each phoneme's relative class frequencies in viseme eight are approximately preserved in each train and validation fold. Thus, the data in each class should be rearranged so that each fold is a good representative of the whole. This process is termed stratification, and the associated cross-validation is stratified.

To implement nested stratified k-fold cross-validation, the k (folds) value should be appropriately chosen. Commonly it is selected between 3 and 10, but the most popular value is 10. A poorly chosen value for k may affect the overall performance of the classifier. A smaller k value indicates that the system is more biased, which is undesirable. A greater k value, on the other hand, is less biased but has a lot of variation. In this work, the proposed nested stratified k-fold cross-validation performance is carried out with the value of k varying from 3 to 10, as shown in fig. 6.9. The mean, median and standard deviation of the accuracy of each iteration is computed. The k value having overlapped mean and median and minor deviation is selected as the best value. Thus, an SVM classifier using nest stratified 5-fold cross-validation is used. The visualization of the proposed method in audio-only, visual-only and audio-visual scenarios is shown in fig. 6.10, 6.11 and 6.12, respectively.

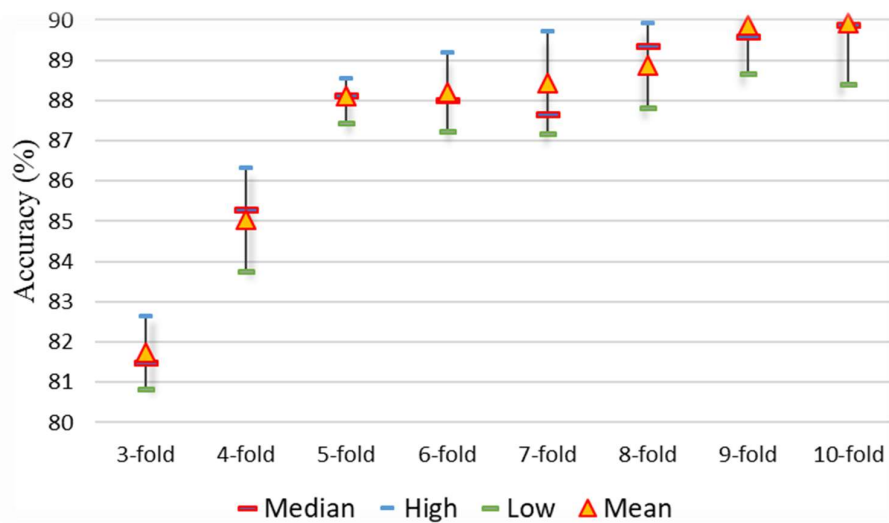


Fig. 6.9 Identification of best fold ‘k’ for Nested Stratified Cross Validation

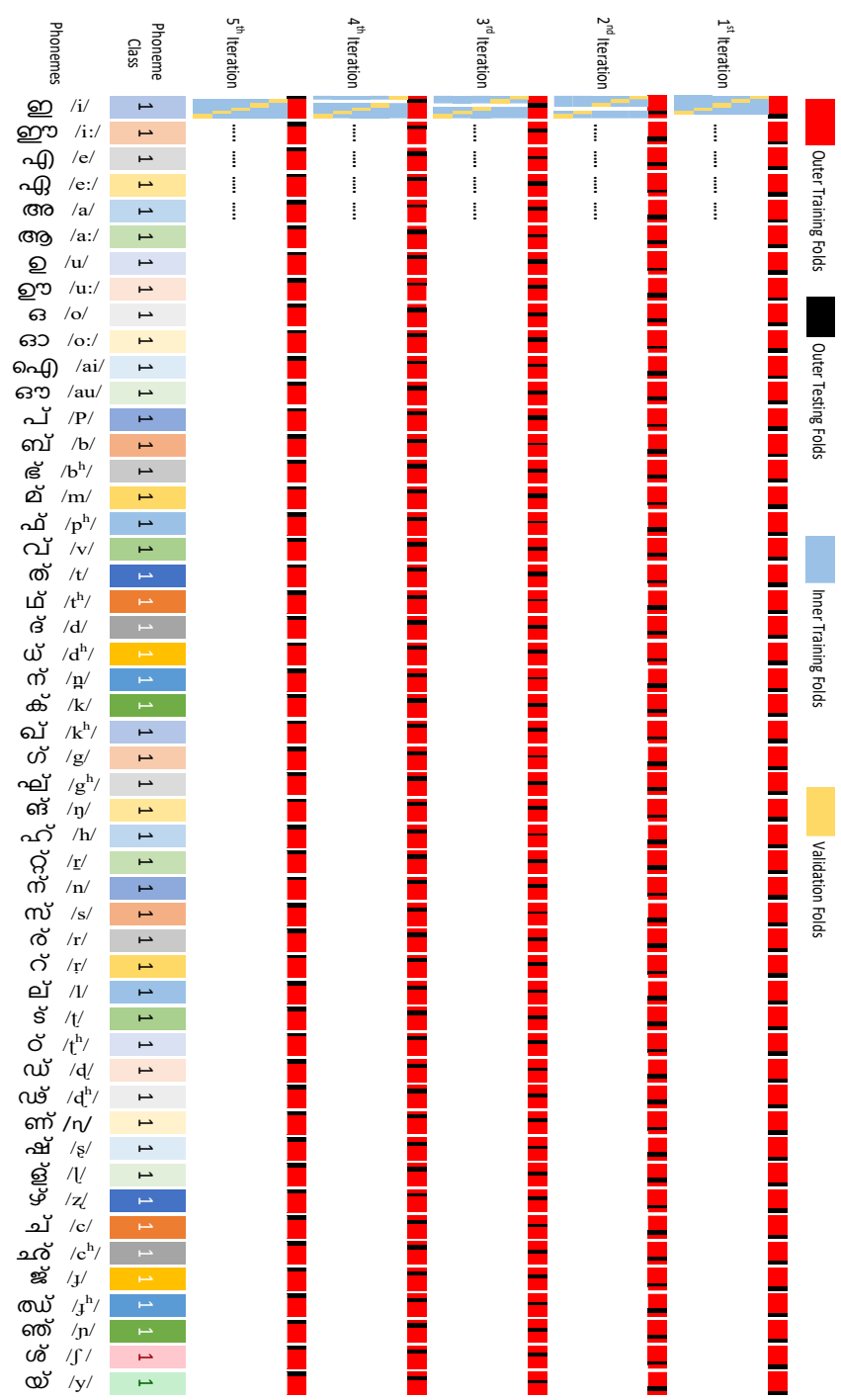


Fig. 6.10 Visualisation of Nested Stratified 5-fold Cross Validation Method for Visual-only and Audio-only Speech Recognition

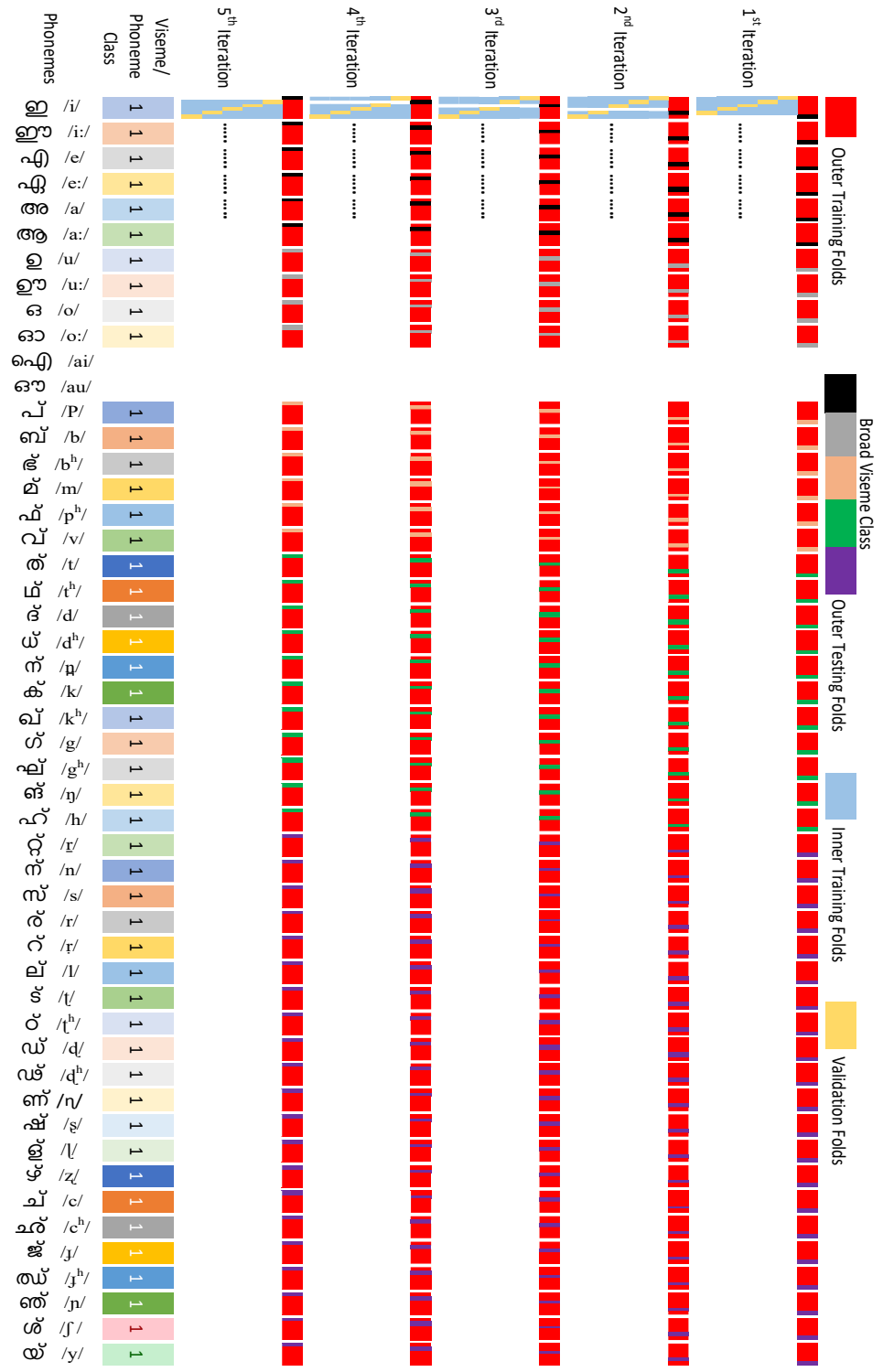


Fig. 6.12 Visualisation of Nested Stratified 5-fold Cross Validation Method for the Second Phase of Audio-Visual Speech Recognition

6.4 Proposed Audio-Visual Speech Recognition System

The schematic diagram of the audio-visual speech recognition (AVSR) system using an SVM classifier is shown in fig. 6.13. It also has audio-only speech recognition and visual-only speech recognition task. This AVSR mainly relies on the visual speech element (viseme) to recognise the underlying phoneme in intense background noise. The performance of the proposed AVSR system is discussed in the next section.

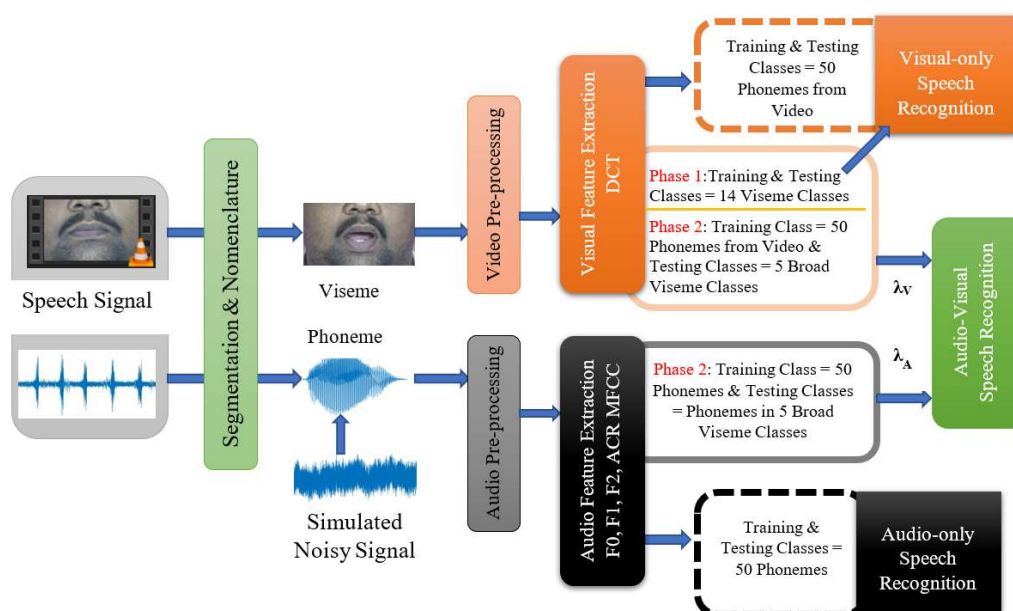


Fig. 6.13 Schematic diagram of Proposed Audio-Visual Speech recognition system

6.5 Experimental Results

The performance of the proposed AVSR system is evaluated with accuracy, precision, recall and F1-score, which are computed from four basic parameters: true positive, true negative, false positive and false negative. The role of the four parameters can be visualised using a confusion matrix, as shown in table 6.6. The class mentioned in the confusion matrix represents phoneme or viseme depending upon the problem under study.

Table 6.6 Confusion Matrix

		Predicted Class	
		Class = Yes	Class = No
Actual Class	Class = Yes	True Positive (TP)	False Negative (FN)
	Class = No	False Positive (FP)	True Negative (TN)

Accuracy is a metric that generally describes how the model performs across all classes as (Eq. 6.12). Precision attempts to answer how precise the model is out of those predicted positive, how many of them are actual positive (Eq. 6.13). Recall tells what portion of the actual class was predicted correctly (Eq. 6.14). F1-score is the Harmonic mean of the Precision and Recall (6.15).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+F} \quad (6.12)$$

$$\text{Precision} = \frac{TP}{TP+F} \quad (6.13)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6.14)$$

$$\text{F1-score} = 2 \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recal}} \quad (6.15)$$

The experimental result for each recognition task is presented in the next sections

6.5.1 Visual-only Speech Recognition

In the experimental procedure, visual speech is utilised for two purposes: visual-only speech recognition and the first phase of audio-visual speech recognition. For the first phase of the AVSR task, the visual speech is represented as 14 visemes. The confusion matrix of 14 visemes and the performance of the first phase of the AVSR task is shown in fig. 6.14 and 6.15

respectively. The experimental result shows that the viseme recognition in the first phase of the AVSR task using SVM gives a better result with an average of 90% in all metrics. Performance of the visual representation of the 50 phonemes is displayed in fig. 6.16. The poor performance of the visual-only speech recognition is quite oblivious as the visual speech carry less information. However, the exceptional performance of the visual representation of diphthongs is due to the selection of its visual signature, as discussed in section 4.7.1. Compared to all other consonants, the system fails to recognise the visual representation of plosive consonants, excluding the bilabial consonants.



Fig. 6.14 Confusion Matrix of 14 Viseme Classes

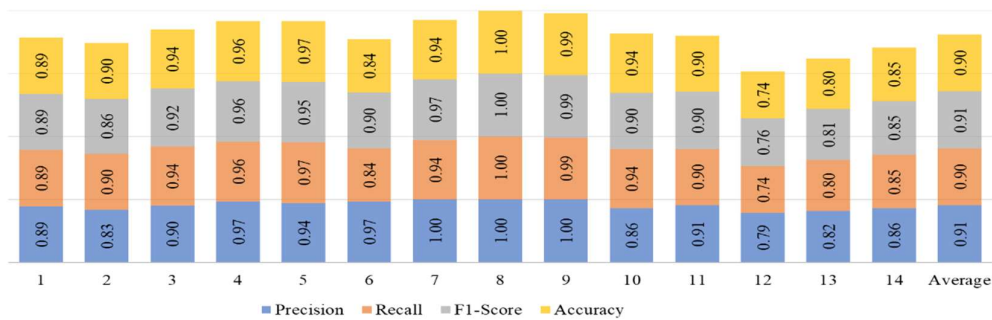


Fig. 6.15 Performance of First Phase of AV Speech Recognition (Viseme Recognition)

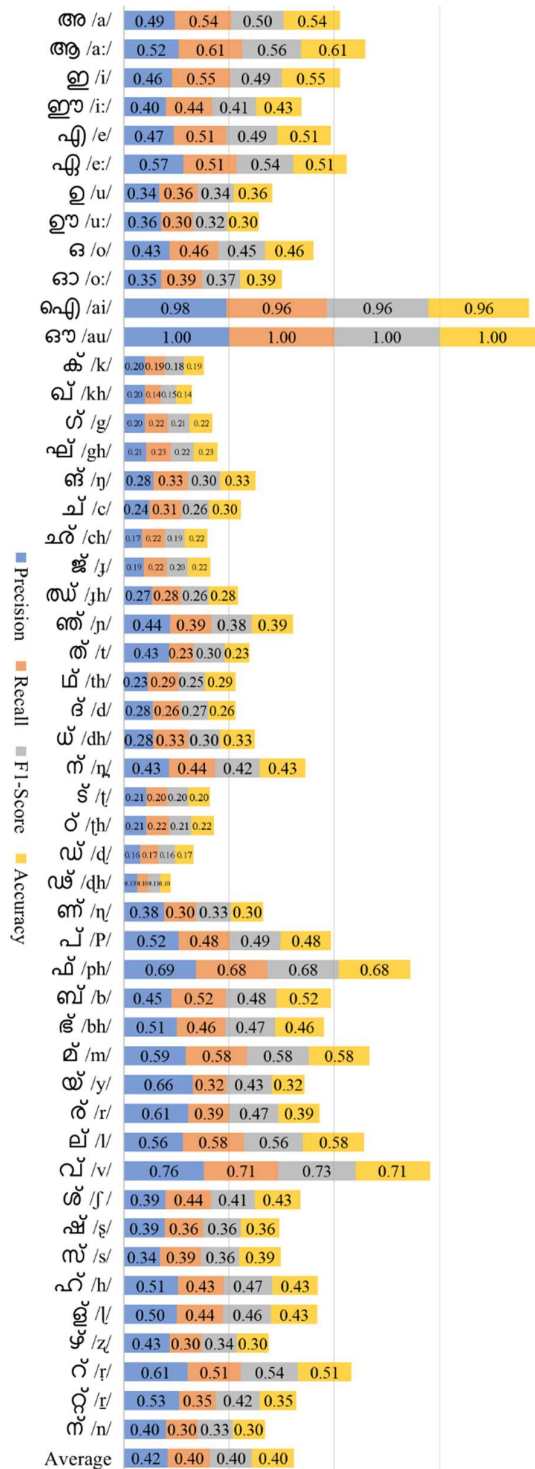


Fig. 6.16 Performance of Visual-only Speech Recognition of Visual representation of 50 Phonemes

6.5.2 Audio-only Speech Recognition

The dataset used for the audio-only speech recognition task consists of clean speech, white Gaussian noise, pink noise and red noise added speech signal. The performance of the audio-only speech recognition task in a clean environment, white Gaussian noise, pink noise and red noise are shown in fig. 6.17, 6.18, 6.19 and 6.20, respectively.

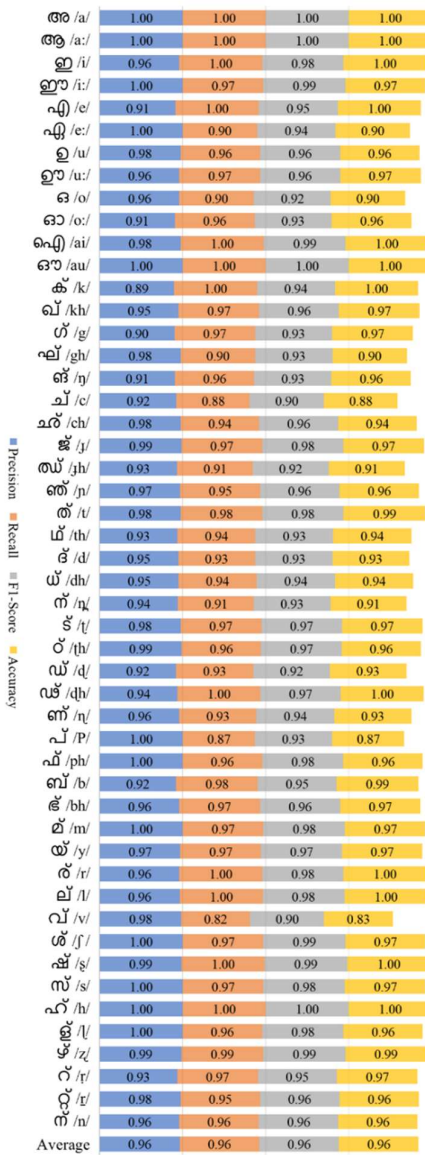


Fig. 6.17 Performance of Audio-only Speech Recognition

	20 dB	10 dB	0 dB	-10 dB	-20 dB
അ /a/	0.98 0.94 0.95 0.94	0.96 0.97 0.96 0.97	0.97 0.93 0.95 0.93	0.84 0.86 0.85 0.86	0.67 0.78 0.71 0.78
ആ /a:/	0.95 0.97 0.96 0.97	0.98 0.99 0.98 0.99	0.94 1.00 0.97 1.00	0.90 0.88 0.88 0.88	0.60 0.74 0.66 0.74
ഇ /i/	0.98 0.97 0.98 0.97	1.00 0.97 0.99 0.97	0.96 0.96 0.96 0.96	0.89 0.80 0.84 0.80	0.65 0.55 0.59 0.55
ഈ /i:/	0.99 0.99 0.99 0.99	0.97 0.99 0.98 0.99	1.00 0.96 0.98 0.96	0.80 0.96 0.87 0.96	0.65 0.71 0.67 0.71
എ /e/	0.99 1.00 0.99 1.00	1.00 1.00 1.00 1.00	0.96 1.00 0.98 1.00	0.86 0.88 0.87 0.88	0.53 0.58 0.55 0.58
ഈ /e:/	0.99 0.99 0.99 0.99	1.00 1.00 1.00 1.00	0.95 0.98 0.96 0.99	0.77 0.90 0.82 0.90	0.51 0.66 0.58 0.67
ഉ /u/	0.98 0.95 0.96 0.96	0.96 0.97 0.96 0.97	0.94 0.97 0.96 0.97	0.93 0.78 0.83 0.78	0.73 0.69 0.69 0.70
ഊ /u:/	0.96 0.97 0.96 0.97	0.96 0.97 0.96 0.97	1.00 0.99 0.99 0.99	0.84 0.85 0.85 0.86	0.70 0.71 0.70 0.71
ഒ /o/	1.00 0.99 0.99 0.99	1.00 1.00 1.00 1.00	1.00 0.93 0.96 0.93	0.94 0.85 0.89 0.86	0.37 0.45 0.50 0.45
ഓ /o:/	0.99 1.00 0.99 1.00	1.00 0.99 0.99 0.99	0.93 0.97 0.95 0.97	0.87 0.91 0.88 0.91	0.73 0.57 0.63 0.57
ഐ /ai/	0.98 1.00 0.99 1.00	1.00 1.00 1.00 1.00	0.95 1.00 0.97 1.00	0.86 0.81 0.83 0.81	0.53 0.69 0.60 0.70
ഔ /au/	1.00 1.00 1.00 1.00	1.00 1.00 1.00 1.00	1.00 0.97 0.98 0.97	0.90 0.93 0.91 0.93	0.69 0.68 0.68 0.68
ക /k/	0.87 1.00 0.92 1.00	0.87 0.99 0.92 0.99	0.86 0.93 0.89 0.93	0.76 0.91 0.81 0.91	0.72 0.94 0.81 0.94
ഖ /kh/	0.98 0.97 0.97 0.97	0.96 0.93 0.94 0.93	0.93 0.89 0.90 0.88	0.90 0.93 0.91 0.93	0.67 0.76 0.70 0.75
ഗ /g/	0.95 0.97 0.96 0.97	0.94 0.96 0.95 0.96	0.98 0.97 0.97 0.97	0.85 0.91 0.87 0.91	0.88 0.87 0.86 0.87
ഘ /gh/	0.98 0.91 0.94 0.91	0.94 0.83 0.87 0.83	0.94 0.84 0.88 0.84	0.92 0.81 0.86 0.81	0.82 0.73 0.76 0.72
ങ /ng/	0.92 0.86 0.87 0.86	0.92 0.84 0.87 0.84	0.72 0.81 0.76 0.81	0.79 0.80 0.79 0.80	0.82 0.74 0.72 0.74
ച /c/	0.94 0.88 0.91 0.88	0.96 0.84 0.89 0.84	0.94 0.88 0.91 0.88	0.88 0.78 0.82 0.78	0.84 0.77 0.79 0.77
ഛ /ch/	0.89 0.96 0.91 0.96	0.84 0.90 0.87 0.90	0.89 0.87 0.87 0.87	0.86 0.85 0.84 0.86	0.75 0.74 0.74 0.74
ജ /j/	1.00 0.97 0.99 0.97	0.99 0.97 0.98 0.97	0.93 0.93 0.92 0.93	0.89 0.85 0.86 0.86	0.79 0.72 0.75 0.72
ത /t/	0.98 0.96 0.96 0.96	0.88 0.94 0.91 0.94	0.96 0.88 0.91 0.88	0.83 0.84 0.82 0.84	0.86 0.71 0.76 0.71
ത്രി /tr/	0.96 0.94 0.95 0.94	0.91 0.88 0.89 0.88	0.86 0.88 0.87 0.88	0.79 0.85 0.81 0.86	0.71 0.74 0.72 0.74
ത /t/	0.95 1.00 0.97 1.00	0.93 0.95 0.93 0.96	0.94 0.94 0.94 0.94	0.90 0.89 0.88 0.90	0.90 0.94 0.91 0.94
ദ /d/	0.97 0.95 0.96 0.96	0.92 0.93 0.92 0.93	0.83 0.93 0.87 0.93	0.91 0.83 0.86 0.83	0.94 0.87 0.90 0.87
ദ്വ /dv/	0.92 0.88 0.89 0.88	0.90 0.83 0.85 0.83	0.84 0.77 0.79 0.77	0.84 0.80 0.81 0.80	0.89 0.80 0.83 0.80
ധ /dh/	0.97 0.93 0.95 0.93	0.88 0.91 0.89 0.91	0.93 0.87 0.90 0.87	0.86 0.81 0.83 0.81	0.91 0.75 0.81 0.75
ന /n/	0.87 0.91 0.89 0.91	0.84 0.90 0.87 0.90	0.88 0.85 0.87 0.86	0.80 0.81 0.80 0.81	0.85 0.78 0.81 0.78
ന്ദ /nd/	0.99 0.99 0.99 0.99	0.93 0.97 0.95 0.97	0.81 0.96 0.88 0.96	0.91 0.84 0.87 0.84	0.85 0.91 0.88 0.91
ന്ത /nt/	0.99 0.96 0.97 0.96	0.92 0.87 0.89 0.87	0.86 0.84 0.84 0.84	0.92 0.83 0.87 0.83	0.80 0.72 0.76 0.72
ന്ദ /nd/	0.93 0.94 0.93 0.94	0.86 0.90 0.88 0.90	0.94 0.91 0.93 0.91	0.90 0.87 0.88 0.87	0.78 0.77 0.77 0.77
ന്ത /nt/	0.92 0.98 0.95 0.99	0.92 0.98 0.95 0.99	0.85 0.97 0.90 0.97	0.93 0.96 0.94 0.96	0.71 0.87 0.78 0.87
ണ /n/	0.89 0.90 0.89 0.90	0.84 0.85 0.84 0.86	0.83 0.81 0.82 0.81	0.88 0.77 0.81 0.77	0.83 0.70 0.75 0.70
പ /p/	0.98 0.93 0.95 0.93	0.98 0.90 0.94 0.90	0.90 0.85 0.87 0.86	0.83 0.78 0.80 0.78	0.91 0.80 0.84 0.80
പ്ല /pl/	1.00 0.94 0.97 0.94	0.99 0.97 0.98 0.97	0.95 0.90 0.92 0.90	0.90 0.90 0.89 0.90	0.86 0.79 0.81 0.78
ബ /b/	0.95 0.94 0.94 0.94	0.91 0.90 0.90 0.90	0.86 0.86 0.85 0.86	0.90 0.86 0.87 0.86	0.72 0.80 0.74 0.80
ബ്ല /bl/	0.96 0.98 0.97 0.99	0.90 0.96 0.92 0.96	0.90 0.88 0.87 0.88	0.83 0.82 0.82 0.83	0.83 0.80 0.81 0.80
മ /m/	0.95 0.94 0.94 0.94	0.94 0.97 0.95 0.97	0.88 0.90 0.88 0.90	0.83 0.84 0.83 0.84	0.79 0.80 0.79 0.80
മ്ല /ml/	0.99 0.96 0.97 0.96	0.99 0.90 0.94 0.90	0.89 0.74 0.80 0.74	0.78 0.70 0.73 0.70	0.85 0.74 0.79 0.74
ര /r/	0.99 1.00 0.99 1.00	0.94 1.00 0.97 1.00	0.93 1.00 0.96 1.00	0.85 0.98 0.91 0.99	0.91 0.91 0.91 0.91
ര്ല /rl/	0.96 0.94 0.95 0.94	0.90 0.90 0.90 0.90	0.88 0.84 0.85 0.84	0.77 0.84 0.80 0.84	0.76 0.78 0.76 0.78
ല /l/	1.00 0.81 0.89 0.81	0.96 0.70 0.80 0.70	0.96 0.64 0.76 0.64	0.80 0.64 0.70 0.64	0.83 0.62 0.71 0.62
ല്ല /ll/	0.98 0.94 0.95 0.94	0.95 0.91 0.93 0.91	0.96 0.91 0.93 0.91	0.79 0.77 0.77 0.77	0.87 0.61 0.70 0.61
ഷ /sh/	0.98 0.99 0.98 0.99	0.97 1.00 0.99 1.00	0.94 1.00 0.97 1.00	0.87 0.93 0.89 0.93	0.79 0.75 0.77 0.75
സ്ല /sl/	0.97 0.99 0.98 0.99	0.94 0.87 0.90 0.87	0.91 0.81 0.85 0.81	0.86 0.72 0.77 0.72	0.68 0.61 0.63 0.61
ഹ /h/	1.00 0.97 0.98 0.97	0.99 0.97 0.98 0.97	0.93 0.97 0.95 0.97	0.84 0.86 0.84 0.86	0.60 0.75 0.66 0.75
ഹ്ല /hl/	0.97 0.96 0.96 0.96	0.94 0.91 0.92 0.91	0.89 0.85 0.87 0.86	0.87 0.84 0.85 0.84	0.86 0.78 0.80 0.78
ഝ /zh/	1.00 0.97 0.98 0.97	0.99 0.97 0.98 0.97	0.85 0.88 0.87 0.88	0.93 0.85 0.88 0.86	0.81 0.66 0.71 0.67
ഞ /j/	0.90 0.97 0.93 0.97	0.98 0.94 0.96 0.94	0.92 0.94 0.93 0.94	0.93 0.87 0.90 0.87	0.78 0.68 0.73 0.68
ട /t/	0.91 0.97 0.94 0.97	0.91 0.95 0.93 0.96	0.89 0.95 0.92 0.96	0.94 0.91 0.91 0.91	0.79 0.87 0.83 0.87
ത്വ /tr/	0.94 0.91 0.92 0.91	0.86 0.93 0.89 0.93	0.90 0.90 0.89 0.90	0.90 0.90 0.90 0.90	0.78 0.84 0.80 0.84
വ്വ /vv/	0.96 0.95 0.95 0.96	0.94 0.93 0.93 0.93	0.91 0.90 0.90 0.90	0.86 0.85 0.85 0.85	0.77 0.74 0.74 0.74

Fig. 6.18 Performance of Audio-only Speech Recognition in White Gaussian Noise

Audio-Visual Speech Recognition Using Support Vector Machine Classifier

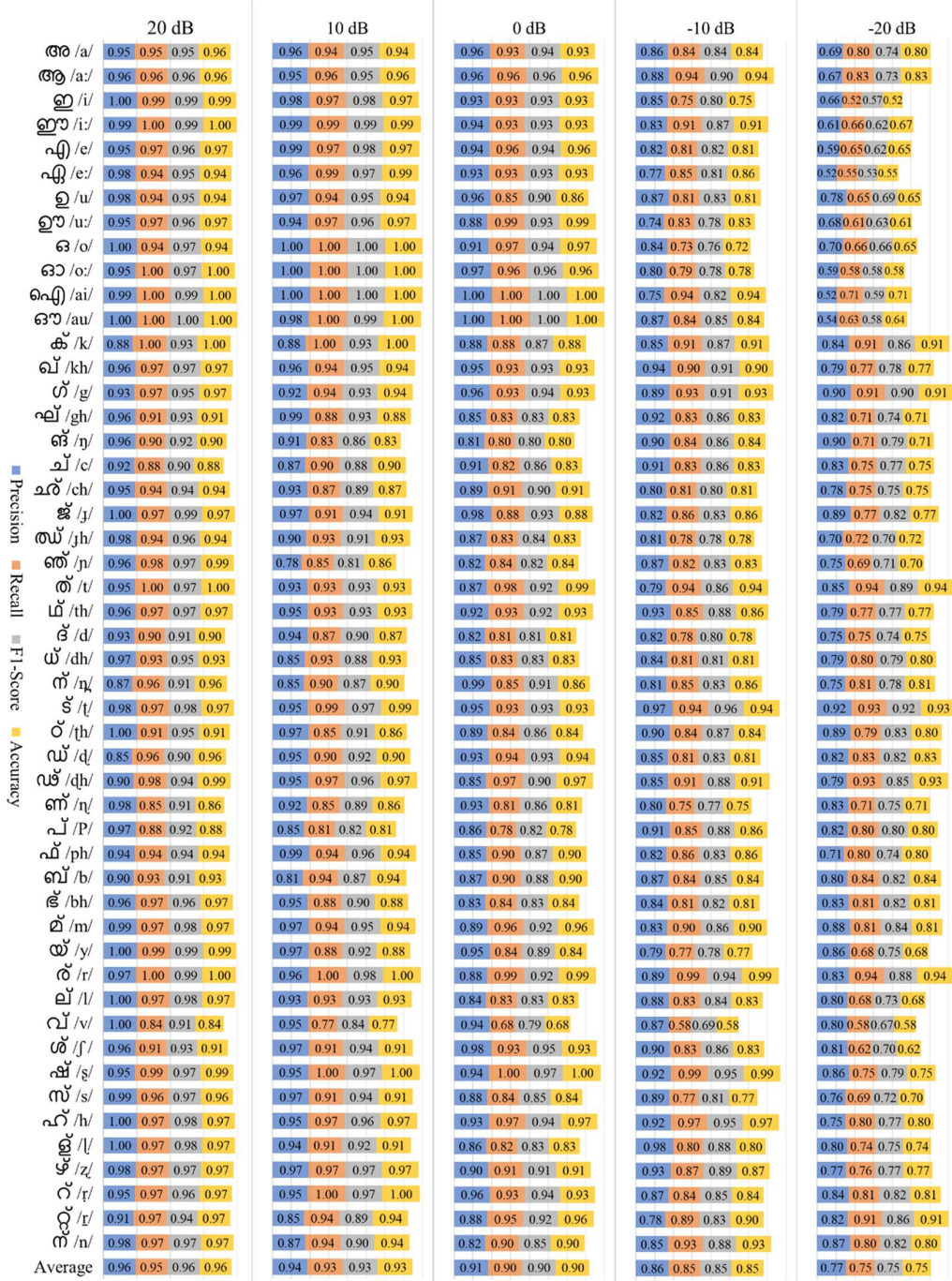


Fig. 6.19 Performance of Audio-only Speech Recognition in Pink Noise

	20 dB	10 dB	0 dB	-10 dB	-20 dB
അ /a/	0.99 0.98 0.99 0.99	1.00 0.99 0.99 0.99	0.97 0.97 0.97 0.97	0.91 0.90 0.90 0.90	0.92 0.90 0.91 0.90
ആ /a:/	0.97 0.99 0.98 0.99	0.99 1.00 0.99 1.00	0.98 0.99 0.98 0.99	0.91 0.94 0.92 0.94	0.87 0.93 0.89 0.93
ഇ /i/	0.96 0.99 0.97 0.99	0.98 0.94 0.96 0.94	1.00 0.97 0.98 0.97	1.00 0.97 0.99 0.97	0.90 0.91 0.90 0.91
ഈ /i:/	0.99 0.99 0.99 0.99	0.95 0.99 0.97 0.99	0.98 1.00 0.99 1.00	1.00 1.00 1.00 1.00	0.97 0.93 0.95 0.93
എ /e/	0.96 0.99 0.97 0.99	0.96 1.00 0.98 1.00	0.96 0.97 0.96 0.97	0.96 0.97 0.96 0.97	0.86 0.88 0.87 0.88
ഏ /e:/	1.00 0.96 0.98 0.96	1.00 0.95 0.98 0.96	0.98 0.96 0.96 0.96	0.97 0.97 0.97 0.97	0.94 0.93 0.93 0.93
ഉ /u/	0.93 0.94 0.94 0.94	0.96 0.94 0.95 0.94	0.95 0.89 0.91 0.90	0.97 0.91 0.94 0.91	0.86 0.88 0.87 0.88
ഊ /u:/	0.95 0.94 0.94 0.94	0.94 0.97 0.96 0.97	0.93 0.95 0.94 0.96	0.92 1.00 0.95 1.00	0.88 0.94 0.91 0.94
ഒ /o/	0.98 0.94 0.96 0.94	1.00 0.97 0.98 0.97	0.96 1.00 0.98 1.00	0.93 0.91 0.92 0.91	0.88 0.74 0.80 0.74
ഓ /o:/	0.95 0.97 0.96 0.97	0.98 0.99 0.98 0.99	0.99 0.96 0.97 0.96	0.92 0.93 0.92 0.93	0.87 0.87 0.86 0.87
ഐ /ai/	0.99 1.00 0.99 1.00	1.00 1.00 1.00 1.00	1.00 1.00 1.00 1.00	0.99 1.00 0.99 1.00	0.86 0.97 0.91 0.97
ഔ /au/	1.00 1.00 1.00 1.00	1.00 1.00 1.00 1.00	0.99 1.00 0.99 1.00	1.00 1.00 1.00 1.00	0.92 0.91 0.91 0.91
ക /k/	0.93 1.00 0.96 1.00	0.88 1.00 0.93 1.00	0.96 0.94 0.95 0.94	0.87 0.95 0.91 0.96	0.83 0.93 0.88 0.93
ഖ /kh/	0.94 0.97 0.95 0.97	0.98 0.94 0.95 0.94	0.93 0.97 0.94 0.97	0.95 0.93 0.93 0.93	0.94 0.86 0.89 0.86
ഗ /g/	0.92 0.96 0.93 0.96	0.96 0.96 0.96 0.96	0.96 0.96 0.96 0.96	0.93 0.94 0.93 0.94	0.91 0.93 0.92 0.93
ഘ /gh/	0.97 0.89 0.92 0.88	0.97 0.90 0.93 0.90	0.94 0.89 0.90 0.88	0.97 0.84 0.90 0.84	0.90 0.84 0.86 0.84
ങ /ng/	0.91 0.96 0.93 0.96	0.98 0.88 0.93 0.88	0.91 0.88 0.89 0.88	0.85 0.83 0.83 0.83	0.87 0.77 0.81 0.77
ച /c/	0.95 0.93 0.93 0.93	0.93 0.85 0.88 0.86	0.96 0.93 0.94 0.93	0.96 0.90 0.92 0.90	0.85 0.82 0.83 0.83
ച്ച /ch/	0.99 0.93 0.95 0.93	0.87 0.94 0.89 0.94	0.95 0.94 0.94 0.94	0.88 0.91 0.89 0.91	0.85 0.90 0.88 0.90
ട /t/	1.00 0.97 0.98 0.97	0.98 0.97 0.98 0.97	0.97 0.99 0.98 0.99	0.96 0.97 0.96 0.97	0.94 0.91 0.92 0.91
ത /th/	0.94 0.97 0.95 0.97	0.96 0.88 0.91 0.88	0.94 0.91 0.92 0.91	0.96 0.94 0.95 0.94	0.93 0.91 0.91 0.91
ത്രി /tr/	0.93 0.95 0.94 0.96	0.94 1.00 0.97 1.00	0.91 0.90 0.90 0.90	0.89 0.88 0.88 0.88	0.81 0.84 0.81 0.84
ത്രി /tr/	0.89 0.97 0.93 0.97	0.97 1.00 0.99 1.00	0.93 0.97 0.95 0.97	0.89 0.95 0.92 0.96	0.87 0.92 0.89 0.93
ദ /d/	0.93 0.94 0.94 0.94	0.92 0.94 0.92 0.94	0.95 0.94 0.94 0.94	0.91 0.94 0.92 0.94	0.94 0.91 0.92 0.91
ധ /dh/	0.95 0.90 0.92 0.90	0.95 0.97 0.96 0.97	0.90 0.90 0.90 0.90	0.86 0.84 0.84 0.84	0.86 0.81 0.83 0.81
ന /n/	0.89 0.94 0.92 0.94	0.89 0.90 0.89 0.90	0.92 0.88 0.90 0.88	0.94 0.88 0.91 0.88	0.82 0.82 0.82 0.83
ന്ദ് /nd/	1.00 0.91 0.95 0.91	0.96 0.93 0.94 0.93	0.94 0.90 0.92 0.90	0.93 0.90 0.90 0.90	0.95 0.85 0.89 0.86
ന്ദ് /nd/	0.99 0.97 0.98 0.97	0.97 0.97 0.97 0.97	0.87 0.94 0.90 0.94	0.94 0.91 0.93 0.91	0.84 0.93 0.88 0.93
ന്ദ് /nd/	0.96 0.93 0.94 0.93	1.00 0.91 0.95 0.91	0.97 0.87 0.91 0.87	0.96 0.93 0.94 0.93	0.92 0.84 0.87 0.84
ന്ദ് /nd/	0.91 0.94 0.92 0.94	0.95 0.93 0.93 0.93	0.95 0.90 0.92 0.90	0.90 0.87 0.88 0.87	0.73 0.83 0.77 0.83
ന്ദ് /nd/	0.96 0.98 0.97 0.99	0.90 0.98 0.94 0.99	0.91 1.00 0.95 1.00	0.85 0.96 0.89 0.96	0.75 0.96 0.83 0.96
ന്ദ് /nd/	0.98 0.90 0.94 0.90	0.92 0.91 0.91 0.91	0.91 0.85 0.88 0.86	0.88 0.81 0.84 0.81	0.95 0.80 0.86 0.80
ന്ദ് /nd/	0.97 0.93 0.95 0.93	0.96 0.91 0.93 0.91	0.95 0.90 0.92 0.90	0.89 0.83 0.85 0.83	0.86 0.81 0.83 0.81
ന്ദ് /nd/	1.00 0.96 0.98 0.96	1.00 0.94 0.97 0.94	1.00 0.93 0.96 0.93	0.96 0.96 0.95 0.96	0.91 0.89 0.89 0.88
ന്ദ് /nd/	0.94 0.97 0.95 0.97	0.92 0.93 0.92 0.93	0.87 0.94 0.90 0.94	0.78 0.90 0.82 0.90	0.83 0.90 0.86 0.90
ന്ദ് /nd/	0.96 0.97 0.97 0.97	0.87 0.94 0.90 0.94	0.96 0.93 0.94 0.93	0.93 0.84 0.87 0.84	0.83 0.81 0.82 0.81
ന്ദ് /nd/	1.00 0.99 0.99 0.99	0.98 1.00 0.99 1.00	0.91 0.96 0.93 0.96	0.91 0.90 0.90 0.90	0.80 0.84 0.82 0.84
ന്ദ് /nd/	0.97 0.99 0.98 0.99	0.98 0.98 0.98 0.99	1.00 0.97 0.98 0.97	0.95 0.85 0.90 0.86	0.84 0.77 0.80 0.77
ന്ദ് /nd/	0.97 0.98 0.98 0.99	1.00 1.00 1.00 1.00	0.96 0.98 0.97 0.99	0.95 1.00 0.97 1.00	0.91 0.93 0.92 0.93
ന്ദ് /nd/	0.99 0.97 0.98 0.97	0.98 0.97 0.98 0.97	0.85 0.90 0.87 0.90	0.73 0.91 0.81 0.91	0.87 0.78 0.80 0.78
ന്ദ് /nd/	0.98 0.85 0.91 0.86	0.92 0.80 0.85 0.80	0.94 0.71 0.80 0.71	0.95 0.62 0.75 0.62	0.94 0.65 0.76 0.65
ന്ദ് /nd/	0.97 0.93 0.95 0.93	0.95 0.91 0.93 0.91	0.97 0.96 0.96 0.96	1.00 0.97 0.98 0.97	0.95 0.93 0.93 0.93
ന്ദ് /nd/	0.93 0.97 0.95 0.97	0.95 1.00 0.97 1.00	0.97 1.00 0.99 1.00	1.00 1.00 1.00 1.00	0.96 1.00 0.98 1.00
ന്ദ് /nd/	1.00 0.99 0.99 0.99	0.97 0.96 0.96 0.96	0.97 0.96 0.96 0.96	0.97 0.88 0.92 0.88	0.95 0.83 0.88 0.83
ന്ദ് /nd/	0.98 0.97 0.98 0.97	0.97 0.97 0.97 0.97	0.93 0.97 0.95 0.97	0.89 0.97 0.93 0.97	0.92 0.94 0.93 0.94
ന്ദ് /nd/	1.00 0.96 0.98 0.96	0.98 0.96 0.97 0.96	0.87 0.85 0.86 0.86	0.88 0.82 0.84 0.83	0.78 0.69 0.73 0.70
ന്ദ് /nd/	0.98 0.97 0.98 0.97	0.98 0.97 0.97 0.97	1.00 0.97 0.98 0.97	0.93 0.91 0.92 0.91	0.89 0.87 0.88 0.87
ന്ദ് /nd/	0.92 0.97 0.95 0.97	0.92 0.94 0.93 0.94	0.89 0.99 0.93 0.99	0.97 0.90 0.93 0.90	0.97 0.87 0.91 0.87
ന്ദ് /nd/	0.98 0.97 0.97 0.97	0.94 0.95 0.95 0.96	0.87 0.95 0.91 0.96	0.82 0.95 0.88 0.96	0.73 0.91 0.81 0.91
ന്ദ് /nd/	0.96 1.00 0.98 1.00	0.96 0.97 0.96 0.97	0.96 0.97 0.96 0.97	0.90 0.87 0.88 0.87	0.88 0.89 0.87 0.87
Average	0.96 0.96 0.96 0.96	0.96 0.95 0.95 0.95	0.95 0.94 0.94 0.94	0.92 0.91 0.91 0.91	0.88 0.87 0.87 0.87

Fig. 6.20 Performance of Audio-only Speech Recognition in Red Noise

6.5.1 Audio-Visual Speech Recognition

This section deals with the second phase of the AVSR task using two different datasets: the visual representation of 50 phonemes and 50 phonemes

in different noisy conditions. The performance of the second phase of AVSR using the first dataset is shown in fig. 6.21. The displayed figure is a concatenation of the performance of AVSR of phonemes in each broad visemes. The average performance AVSR using a visual representation of 50 phonemes is slightly improved when compared with visual-only speech recognition task as in fig. 6.16. Since viseme classes having single phonemes were skipped in the second phase of the AVSR task, which is reflected as the drastic variation in metrics of models using the visual equivalent of diphthongs. The performance of the second phase of AVSR using 50 phonemes in a clean environment, white Gaussian noise, pink noise and red noise is shown in fig. 6.22, 6.23, 6.24 and 6.25, respectively. The average performance of the AVSR task is improved significantly below 0 dB noise level compared with the performance of the audio-only speech recognition task in the corresponding noisy conditions. Remarkable improvement is noticed for the AVSR task in white Gaussian noise, then in pink noise and less in red noise.

The final decision in the performance of the AVSR system depends mainly on the estimation of stream weight value in different noisy conditions. The average accuracy of the second phase of the AVSR system using 50 phonemes ranged from 96 % (clean environment) to 83 % (white gaussian noise with -20 dB noise level), and that of using a visual representation of 50 phonemes is 45 %. This remarkable difference is reflected in the reliability measurement of streams, thereby in stream weight also. Thus, no combination of streams in Eq. 6.11 can improve the overall performance better than the second phase of the AVSR task using 50 phonemes. Hence, the stream weight value of the audio stream is given unity ($\lambda_A = 1$) for different noise types and noise levels used. In other words, the proposed AVSR system consists of the first phase to recognise the viseme of underlying phoneme from 14 visemes and the second phase to recognise the phoneme from phonemes within the selected broad viseme using audio speech alone. Thus, the broad viseme has improved the accuracy by 9%, 8% and 3% for white Gaussian noise, pink noise and red noise, respectively, at -20 dB compared with its corresponding audio-only

version (fig. 6.18, 6.19 and 6.20). In addition, the hierarchical approach also reduced the computational cost by recognising the phoneme from the phonemes of the selected broad viseme instead of from 50 phonemes.

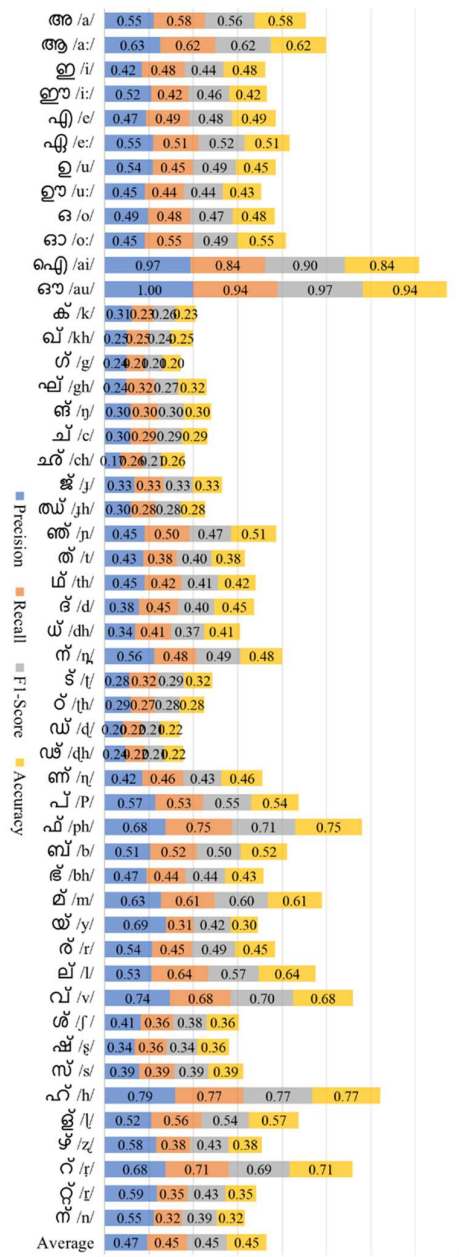


Fig. 6.21 Performance of Second Phase of AV Speech Recognition of Visual representation of 50 Phonemes

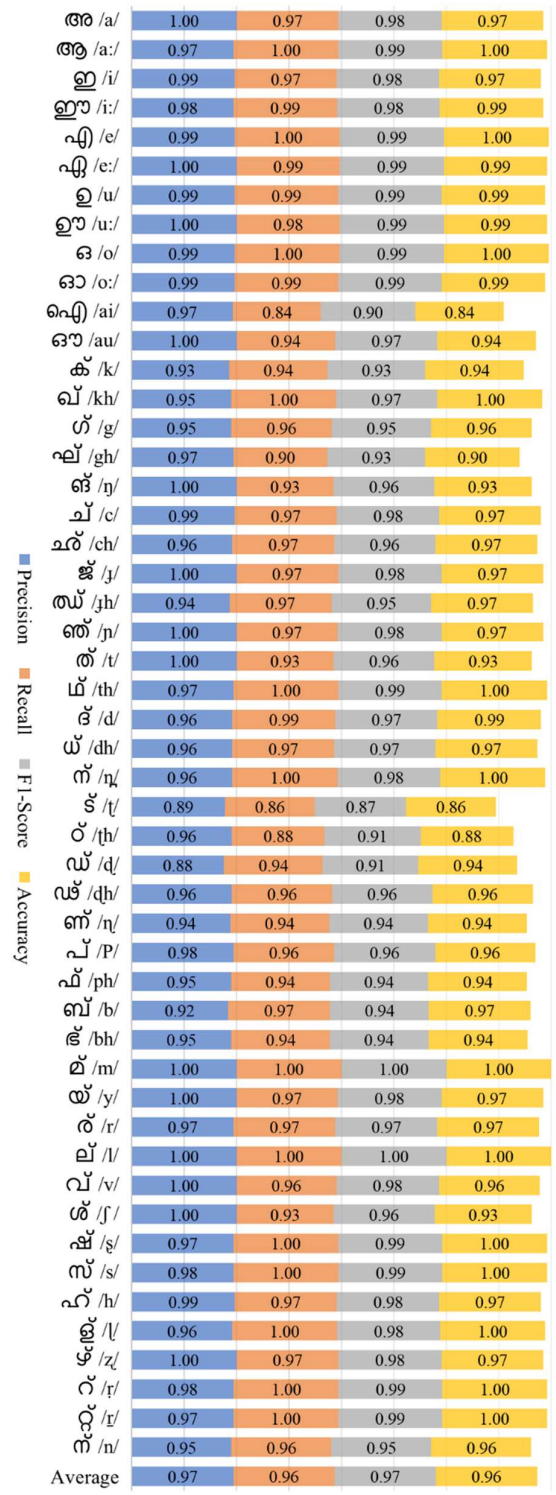


Fig. 6.22 Performance of Second Phase of AV Speech Recognition of 50 Phonemes from Audio

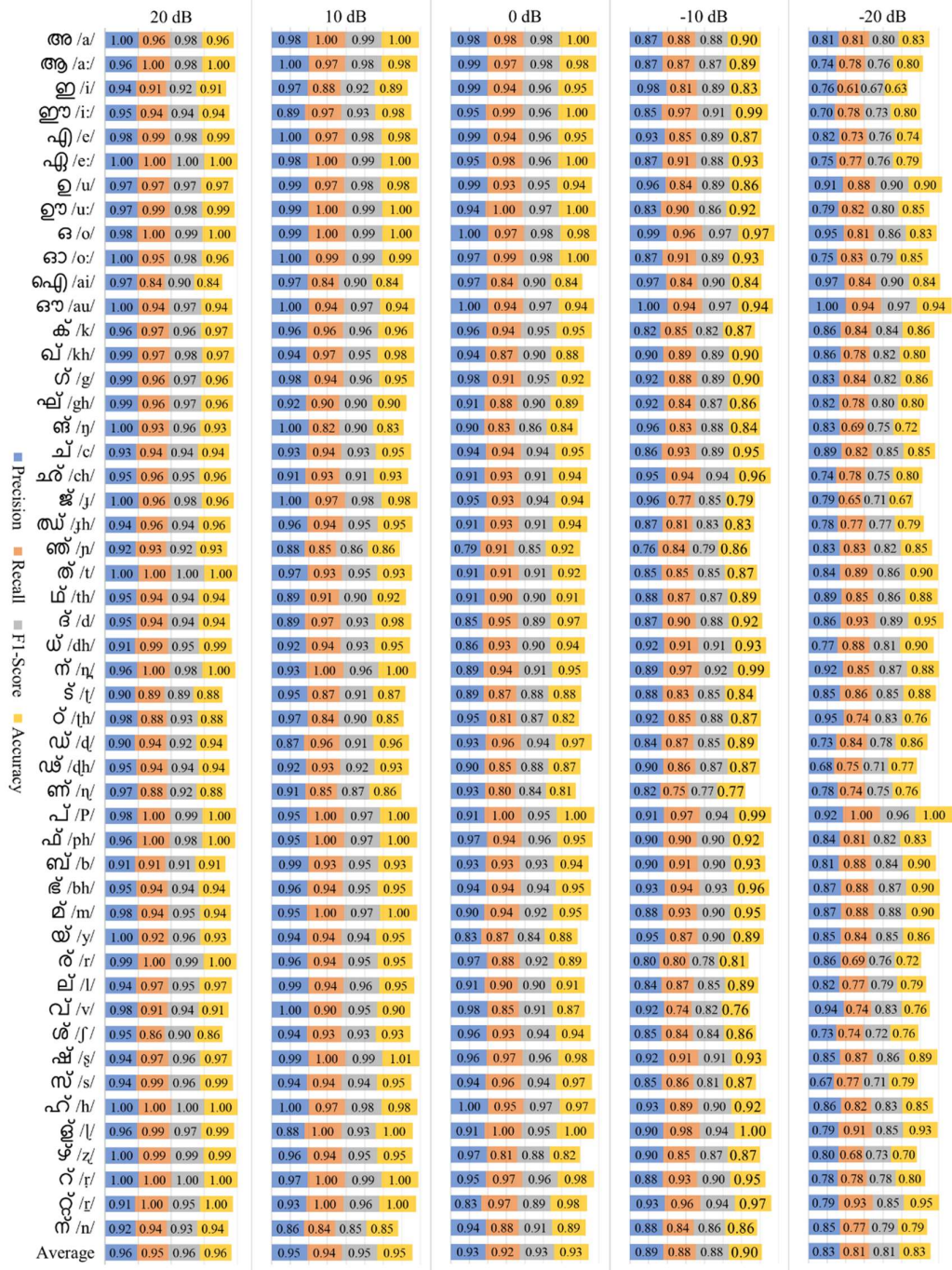


Fig. 6.23 Performance of AV Speech Recognition of 50 Phonemes in White Gaussian Noise

Audio-Visual Speech Recognition Using Support Vector Machine Classifier

	20 dB	10 dB	0 dB	-10 dB	-20 dB
അ /a/	0.98 0.96 0.96 0.96	1.00 0.99 0.99 0.99	0.95 0.93 0.94 0.93	0.89 0.87 0.88 0.87	0.81 0.90 0.84 0.90
ആ /a:/	0.96 0.97 0.96 0.97	0.99 1.00 0.99 1.00	0.93 0.94 0.93 0.94	0.88 0.91 0.89 0.91	0.75 0.85 0.80 0.86
ഇ /i/	0.97 0.93 0.95 0.93	0.97 0.97 0.97 0.97	0.97 0.87 0.91 0.87	0.94 0.78 0.85 0.78	0.75 0.580.640.58
ഈ /i:/	0.94 0.97 0.96 0.97	0.98 0.97 0.97 0.97	0.90 0.95 0.92 0.96	0.87 0.93 0.89 0.93	0.74 0.75 0.74 0.75
എ /e/	0.99 0.94 0.96 0.94	0.99 0.97 0.98 0.97	0.96 0.93 0.94 0.93	0.89 0.88 0.88 0.88	0.80 0.67 0.730.67
ഏ /e:/	0.95 1.00 0.97 1.00	0.97 0.98 0.98 0.99	0.93 0.99 0.96 0.99	0.88 0.91 0.89 0.91	0.77 0.84 0.80 0.84
ഉ /u/	1.00 0.96 0.98 0.96	1.00 0.94 0.97 0.94	0.97 0.94 0.96 0.94	0.99 0.84 0.90 0.84	0.87 0.78 0.81 0.78
ഊ /u:/	0.97 1.00 0.99 1.00	0.95 1.00 0.97 1.00	0.95 1.00 0.97 1.00	0.82 0.93 0.87 0.93	0.84 0.87 0.85 0.87
ഓ /o/	0.99 1.00 0.99 1.00	1.00 0.99 0.99 0.99	0.99 0.96 0.97 0.96	0.95 0.91 0.93 0.91	0.87 0.87 0.87 0.87
ഔ /o:/	1.00 1.00 1.00 1.00	0.99 1.00 0.99 1.00	0.96 0.96 0.96 0.96	0.86 0.88 0.87 0.88	0.85 0.86 0.85 0.86
ഐ /ai/	0.97 0.84 0.90 0.84	0.97 0.84 0.90 0.84	0.97 0.84 0.90 0.84	0.97 0.84 0.90 0.84	0.97 0.84 0.90 0.84
ഔ /au/	1.00 0.94 0.97 0.94	1.00 0.94 0.97 0.94	1.00 0.94 0.97 0.94	1.00 0.94 0.97 0.94	1.00 0.94 0.97 0.94
ക /k/	0.94 0.94 0.94 0.94	0.95 0.94 0.94 0.94	0.88 0.88 0.88 0.88	0.85 0.93 0.88 0.93	0.87 0.93 0.89 0.93
ഖ /kh/	0.99 1.00 0.99 1.00	0.91 0.93 0.92 0.93	0.95 0.94 0.94 0.94	0.91 0.89 0.90 0.88	0.90 0.83 0.86 0.83
ഗ് /g/	0.96 0.97 0.96 0.97	0.92 0.93 0.92 0.93	0.92 0.91 0.91 0.91	0.88 0.90 0.88 0.90	0.88 0.90 0.89 0.90
ഘ /gh/	1.00 0.93 0.96 0.93	0.99 0.90 0.94 0.90	0.97 0.90 0.93 0.90	0.91 0.94 0.92 0.94	0.88 0.84 0.85 0.84
ങ് /ng/	0.98 0.94 0.95 0.94	0.94 0.93 0.93 0.93	0.93 0.81 0.86 0.81	0.91 0.81 0.86 0.81	0.80 0.71 0.75 0.71
ച് /c/	0.96 0.97 0.96 0.97	0.90 0.91 0.91 0.91	0.92 0.93 0.92 0.93	0.91 0.88 0.90 0.88	0.87 0.87 0.87 0.87
ച്ച /ch/	0.98 0.97 0.97 0.97	0.93 0.94 0.94 0.94	0.92 0.97 0.94 0.97	0.90 0.90 0.90 0.90	0.90 0.78 0.83 0.78
ജ് /j/	1.00 0.97 0.98 0.97	0.97 0.88 0.92 0.88	0.97 0.87 0.92 0.87	0.86 0.74 0.79 0.74	0.81 0.75 0.77 0.75
ട് /t/	0.97 0.96 0.96 0.96	0.92 0.93 0.92 0.93	0.95 0.89 0.91 0.88	0.89 0.85 0.87 0.86	0.74 0.79 0.77 0.80
ത് /n/	0.95 0.94 0.94 0.94	0.88 0.90 0.88 0.90	0.89 0.87 0.87 0.87	0.87 0.87 0.86 0.87	0.81 0.75 0.78 0.75
ഥ് /th/	0.97 0.93 0.95 0.93	0.94 0.89 0.91 0.88	0.82 0.90 0.85 0.90	0.87 0.91 0.89 0.91	0.88 0.87 0.87 0.87
ധ് /dh/	0.93 0.97 0.95 0.97	0.85 0.91 0.88 0.91	0.94 0.93 0.93 0.93	0.98 0.85 0.91 0.86	0.87 0.81 0.84 0.81
ന്ദ് /d/	0.95 0.96 0.95 0.96	0.93 0.97 0.95 0.97	0.92 0.94 0.93 0.94	0.85 0.85 0.85 0.86	0.91 0.90 0.90 0.90
ന്ധ് /dh/	0.97 0.97 0.97 0.97	0.93 0.94 0.93 0.94	0.94 0.93 0.93 0.93	0.90 0.90 0.90 0.90	0.80 0.87 0.82 0.87
ൻ /n/	0.98 1.00 0.99 1.00	0.96 0.99 0.97 0.99	0.90 0.96 0.92 0.96	0.93 0.96 0.94 0.96	0.88 0.94 0.90 0.94
ക് /t/	0.93 0.86 0.89 0.86	0.93 0.90 0.91 0.90	0.88 0.89 0.88 0.88	0.87 0.87 0.87 0.87	0.90 0.87 0.88 0.87
ക് /th/	0.91 0.90 0.90 0.90	1.00 0.85 0.92 0.86	0.92 0.80 0.85 0.80	0.91 0.83 0.86 0.83	0.84 0.74 0.77 0.74
ന്ദ് /d/	0.88 0.97 0.92 0.97	0.87 0.94 0.90 0.94	0.98 0.97 0.97 0.97	0.86 0.84 0.85 0.84	0.74 0.85 0.79 0.86
ന്ധ് /dh/	0.92 0.88 0.90 0.88	0.99 0.96 0.97 0.96	0.90 0.93 0.91 0.93	0.81 0.87 0.83 0.87	0.78 0.81 0.79 0.81
ൻ /n/	0.95 0.90 0.92 0.90	0.90 0.84 0.86 0.84	0.86 0.85 0.85 0.86	0.76 0.77 0.76 0.77	0.71 0.71 0.71 0.71
പ് /p/	0.97 0.97 0.97 0.97	0.98 0.96 0.96 0.96	0.95 0.96 0.95 0.96	0.90 0.99 0.94 0.99	0.93 1.00 0.96 1.00
ഫ് /ph/	0.93 0.97 0.94 0.97	0.96 0.97 0.96 0.97	0.92 0.92 0.91 0.93	0.94 0.91 0.92 0.91	0.89 0.88 0.88 0.88
ബ് /b/	0.97 0.93 0.95 0.93	0.89 0.93 0.90 0.93	0.98 0.93 0.95 0.93	0.90 0.87 0.88 0.87	0.86 0.93 0.89 0.93
ഭ് /bh/	0.92 0.94 0.92 0.94	0.95 0.94 0.94 0.94	0.94 0.94 0.93 0.94	0.91 0.96 0.93 0.96	0.94 0.91 0.93 0.91
മ് /m/	0.93 0.97 0.95 0.97	0.96 0.93 0.94 0.93	0.95 1.00 0.97 1.00	0.87 0.96 0.91 0.96	0.92 0.85 0.88 0.86
യ് /y/	1.00 0.95 0.98 0.96	0.91 0.92 0.91 0.93	0.93 0.88 0.90 0.88	0.94 0.88 0.90 0.88	0.87 0.81 0.83 0.81
ര് /r/	1.00 1.00 1.00 1.00	0.97 0.97 0.97 0.97	0.89 0.91 0.89 0.91	0.86 0.83 0.83 0.83	0.78 0.74 0.76 0.74
ല് /l/	1.00 0.96 0.98 0.96	0.92 0.93 0.92 0.93	0.86 0.84 0.85 0.84	0.87 0.90 0.88 0.90	0.75 0.73 0.73 0.72
വ് /v/	1.00 0.90 0.94 0.90	0.96 0.93 0.94 0.93	0.99 0.91 0.95 0.91	0.98 0.75 0.85 0.75	0.86 0.78 0.82 0.78
ശ് /f/	0.98 0.89 0.92 0.88	0.97 0.93 0.95 0.93	0.94 0.99 0.96 0.99	0.86 0.85 0.85 0.86	0.81 0.74 0.77 0.74
ഷ് /s/	0.97 1.00 0.99 1.00	0.97 0.97 0.97 0.97	0.96 0.97 0.96 0.97	0.98 0.93 0.95 0.93	0.77 0.85 0.80 0.86
സ് /s/	0.93 1.00 0.96 1.00	0.94 0.99 0.96 0.99	0.93 0.94 0.93 0.94	0.88 0.90 0.88 0.90	0.80 0.84 0.81 0.84
ഹ് /h/	1.00 0.98 0.99 0.99	1.00 0.97 0.98 0.97	0.99 0.97 0.98 0.97	0.94 0.92 0.93 0.93	0.88 0.84 0.84 0.84
ല് /l/	0.95 1.00 0.97 1.00	0.98 1.00 0.99 1.00	0.96 1.00 0.98 1.00	0.94 0.96 0.94 0.96	0.78 0.85 0.81 0.86
ഴ് /z/	1.00 0.97 0.98 0.97	0.99 0.96 0.97 0.96	0.94 0.93 0.93 0.93	0.87 0.84 0.85 0.84	0.84 0.78 0.80 0.78
റ് /r/	0.99 1.00 0.99 1.00	1.00 1.00 1.00 1.00	0.98 0.96 0.97 0.96	0.79 0.86 0.81 0.86	0.80 0.84 0.82 0.84
ല് /l/	0.92 1.00 0.96 1.00	0.85 1.00 0.92 1.00	0.89 0.97 0.93 0.97	0.82 0.94 0.87 0.94	0.82 0.97 0.89 0.97
ൻ /n/	0.94 0.91 0.92 0.91	0.94 0.89 0.91 0.88	0.90 0.87 0.88 0.87	0.90 0.84 0.87 0.84	0.92 0.75 0.82 0.75
Average	0.96 0.96 0.96 0.96	0.95 0.94 0.94 0.94	0.93 0.92 0.93 0.92	0.89 0.88 0.88 0.88	0.84 0.83 0.83 0.83

Fig. 6.24 Performance of AV Speech Recognition of 50 Phonemes in Pink Noise

	20 dB	10 dB	0 dB	-10 dB	-20 dB
അ /a/	1.00 0.97 0.98 0.97	1.00 1.00 1.00 1.00	0.96 0.94 0.95 0.94	0.92 0.94 0.93 0.94	0.96 0.85 0.90 0.86
ആ /a:/	0.97 1.00 0.99 1.00	1.00 1.00 1.00 1.00	0.94 0.96 0.95 0.96	0.95 0.91 0.93 0.91	0.88 0.97 0.92 0.97
ഇ /i/	0.98 0.95 0.96 0.96	0.97 0.91 0.94 0.91	0.95 0.91 0.92 0.91	0.96 0.88 0.92 0.88	0.94 0.84 0.89 0.84
ഈ /i:/	0.98 0.99 0.98 0.99	0.95 0.97 0.96 0.97	0.93 0.95 0.93 0.96	0.92 0.95 0.93 0.96	0.87 0.97 0.92 0.97
എ /e/	0.98 0.96 0.97 0.96	0.98 0.99 0.98 0.99	1.00 0.94 0.97 0.94	0.95 0.84 0.88 0.84	0.95 0.90 0.91 0.90
ഐ /e:/	0.98 1.00 0.99 1.00	0.99 1.00 0.99 1.00	0.96 1.00 0.98 1.00	0.88 0.97 0.92 0.97	0.93 0.95 0.94 0.96
ഉ /u/	1.00 0.97 0.98 0.97	0.99 1.00 0.99 1.00	0.98 0.94 0.96 0.94	1.00 0.93 0.96 0.93	0.95 0.91 0.92 0.91
ഊ /u:/	0.99 1.00 0.99 1.00	1.00 1.00 1.00 1.00	0.95 0.97 0.96 0.97	0.96 0.98 0.97 0.99	0.93 0.91 0.92 0.91
ഓ /o/	1.00 0.97 0.98 0.97	1.00 0.97 0.98 0.97	1.00 0.99 0.99 0.99	0.98 0.98 0.98 0.99	0.94 0.93 0.93 0.93
ഔ /o:/	0.96 1.00 0.98 1.00	0.98 0.99 0.98 0.99	0.97 0.99 0.98 0.99	0.96 1.00 0.98 1.00	0.91 0.96 0.93 0.96
ഐ /ai/	0.97 0.84 0.90 0.84	0.97 0.84 0.90 0.84	0.97 0.84 0.90 0.84	0.97 0.84 0.90 0.84	0.97 0.84 0.90 0.84
ഔ /au/	1.00 0.94 0.97 0.94	1.00 0.94 0.97 0.94	1.00 0.94 0.97 0.94	1.00 0.94 0.97 0.94	1.00 0.94 0.97 0.94
ക /k/	0.97 0.93 0.94 0.93	1.00 0.93 0.96 0.93	0.99 0.94 0.96 0.94	0.90 0.91 0.90 0.91	0.83 0.91 0.87 0.91
ഖ /kh/	0.96 0.99 0.97 0.99	0.95 0.96 0.95 0.96	0.96 0.93 0.94 0.93	0.96 0.91 0.93 0.91	0.89 0.86 0.86 0.86
ഗ /g/	0.96 0.99 0.97 0.99	0.93 0.99 0.96 0.99	0.98 0.96 0.96 0.96	0.98 0.91 0.94 0.91	0.87 0.91 0.89 0.91
ഘ /gh/	0.99 0.91 0.94 0.91	0.99 0.91 0.94 0.91	0.99 0.93 0.95 0.93	0.91 0.94 0.92 0.94	0.89 0.91 0.90 0.91
ങ /ṅ/	1.00 0.97 0.99 0.97	1.00 0.97 0.98 0.97	1.00 0.88 0.93 0.88	0.95 0.82 0.88 0.83	0.92 0.84 0.88 0.84
ച /c/	0.93 0.94 0.93 0.94	0.95 1.00 0.97 1.00	0.94 0.91 0.93 0.91	0.96 0.94 0.95 0.94	0.89 0.91 0.90 0.91
ച്ച /ch/	0.93 0.96 0.94 0.96	0.94 0.97 0.95 0.97	0.93 0.94 0.94 0.94	0.90 0.94 0.92 0.94	0.93 0.93 0.93 0.93
ജ /j/	1.00 0.94 0.97 0.94	0.98 0.91 0.95 0.91	1.00 0.91 0.95 0.91	1.00 0.96 0.98 0.96	0.95 0.87 0.91 0.87
യ്യാ /jya/	0.96 0.96 0.96 0.96	0.95 0.93 0.93 0.93	0.93 0.96 0.94 0.96	0.96 0.93 0.94 0.93	0.86 0.90 0.87 0.90
യ്യാ /ja/	1.00 0.97 0.98 0.97	1.00 0.94 0.97 0.94	0.98 0.96 0.96 0.96	0.90 0.90 0.89 0.90	0.88 0.87 0.87 0.87
യ്യാ /ya/	1.00 0.93 0.96 0.93	1.00 0.97 0.98 0.97	0.95 0.87 0.91 0.87	0.92 0.86 0.87 0.86	0.87 0.84 0.85 0.84
യ്യാ /ya/	0.91 0.99 0.94 0.99	0.93 0.97 0.95 0.97	0.83 0.97 0.89 0.97	0.87 0.93 0.90 0.93	0.86 0.88 0.87 0.88
യ്യാ /ya/	0.96 0.99 0.97 0.99	0.99 0.99 0.99 0.99	0.96 0.96 0.96 0.96	0.96 0.96 0.96 0.96	0.96 0.90 0.92 0.90
യ്യാ /ya/	0.96 0.97 0.96 0.97	0.95 0.97 0.96 0.97	0.92 0.94 0.93 0.94	0.92 0.93 0.92 0.93	0.91 0.91 0.91 0.91
യ്യാ /ya/	1.00 1.00 1.00 1.00	0.98 1.00 0.99 1.00	0.92 1.00 0.95 1.00	0.89 1.00 0.94 1.00	0.94 0.96 0.95 0.96
യ്യാ /ya/	0.93 0.86 0.89 0.86	0.94 0.90 0.92 0.90	0.89 0.87 0.88 0.87	0.91 0.87 0.88 0.87	0.88 0.89 0.88 0.88
യ്യാ /ya/	0.97 0.90 0.93 0.90	0.95 0.91 0.92 0.91	0.95 0.87 0.91 0.87	1.00 0.90 0.94 0.90	0.89 0.81 0.85 0.81
യ്യാ /ya/	0.85 0.96 0.90 0.96	0.89 0.98 0.93 0.99	0.88 0.93 0.90 0.93	0.90 0.94 0.92 0.94	0.82 0.91 0.86 0.91
യ്യാ /ya/	0.98 0.94 0.96 0.94	0.91 0.91 0.90 0.91	0.99 0.97 0.98 0.97	0.92 0.91 0.91 0.91	0.86 0.87 0.86 0.87
യ്യാ /ya/	0.92 0.93 0.92 0.93	0.94 0.93 0.93 0.93	0.93 0.87 0.88 0.87	0.89 0.85 0.86 0.86	0.90 0.77 0.82 0.77
യ്യാ /ya/	0.97 0.99 0.98 0.99	0.97 0.97 0.97 0.97	0.95 0.94 0.94 0.94	0.94 0.97 0.95 0.97	0.94 0.93 0.93 0.93
യ്യാ /ya/	0.95 0.97 0.96 0.97	0.92 0.94 0.93 0.94	0.93 0.94 0.93 0.94	0.94 0.94 0.93 0.94	0.93 0.95 0.94 0.96
യ്യാ /ya/	0.98 0.95 0.96 0.96	0.96 0.95 0.96 0.96	0.88 0.91 0.89 0.91	0.87 0.97 0.92 0.97	0.84 0.90 0.87 0.90
യ്യാ /ya/	0.96 0.96 0.96 0.96	0.94 0.94 0.93 0.94	0.93 0.94 0.93 0.94	0.92 0.93 0.92 0.93	0.96 0.90 0.93 0.90
യ്യാ /ya/	1.00 1.00 1.00 1.00	1.00 0.99 0.99 0.99	0.99 0.97 0.98 0.97	0.93 0.91 0.92 0.91	0.88 0.94 0.91 0.94
യ്യാ /ya/	1.00 0.95 0.98 0.96	1.00 0.94 0.96 0.94	0.98 0.97 0.97 0.97	0.90 0.92 0.91 0.93	0.88 0.84 0.85 0.84
യ്യാ /ya/	0.93 0.97 0.95 0.97	1.00 0.97 0.98 0.97	0.99 0.94 0.96 0.94	0.92 0.90 0.91 0.90	0.90 0.78 0.82 0.78
യ്യാ /ya/	1.00 0.94 0.97 0.94	0.98 0.97 0.97 0.97	0.96 0.94 0.95 0.94	0.83 0.83 0.83 0.83	0.77 0.83 0.79 0.83
യ്യാ /ya/	1.00 0.97 0.98 0.97	1.00 0.97 0.98 0.97	0.97 0.88 0.92 0.88	0.97 0.78 0.86 0.78	0.94 0.81 0.86 0.81
യ്യാ /ya/	1.00 0.96 0.98 0.96	0.96 0.96 0.96 0.96	0.96 0.97 0.96 0.97	1.00 0.91 0.95 0.91	0.95 0.93 0.94 0.93
യ്യാ /ya/	0.96 1.00 0.98 1.00	0.99 0.97 0.98 0.97	0.97 0.97 0.97 0.97	0.95 1.00 0.97 1.00	0.96 0.97 0.96 0.97
യ്യാ /ya/	1.00 1.00 1.00 1.00	0.98 0.99 0.98 0.99	1.00 1.00 1.00 1.00	0.95 0.96 0.95 0.96	0.94 0.94 0.94 0.94
യ്യാ /ya/	0.99 0.97 0.98 0.97	0.97 0.97 0.97 0.97	0.95 0.95 0.95 0.96	0.98 0.95 0.96 0.96	0.96 0.88 0.91 0.88
യ്യാ /ya/	0.95 1.00 0.97 1.00	0.94 1.00 0.97 1.00	0.96 1.00 0.98 1.00	0.93 0.99 0.95 0.99	0.80 0.94 0.86 0.94
യ്യാ /ya/	0.99 0.97 0.98 0.97	1.00 0.97 0.98 0.97	0.96 0.96 0.96 0.96	0.96 0.91 0.93 0.91	0.93 0.81 0.86 0.81
യ്യാ /ya/	0.98 1.00 0.99 1.00	1.00 1.00 1.00 1.00	0.95 1.00 0.97 1.00	0.95 0.94 0.94 0.94	0.89 0.93 0.91 0.93
യ്യാ /ya/	0.92 1.00 0.96 1.00	0.95 1.00 0.97 1.00	0.87 1.00 0.93 1.00	0.84 0.96 0.89 0.96	0.89 0.97 0.93 0.97
യ്യാ /ya/	0.97 0.94 0.95 0.94	0.99 0.96 0.97 0.96	0.94 0.91 0.93 0.91	0.89 0.82 0.85 0.83	0.92 0.87 0.89 0.87
Average	0.97 0.96 0.96 0.96	0.97 0.96 0.96 0.96	0.95 0.94 0.94 0.94	0.93 0.92 0.92 0.92	0.91 0.89 0.90 0.90

Fig. 6.25 Performance of AV Speech Recognition of 50 Phonemes in Red Noise

The performance of visual-only, audio-only and AVSR recognition systems in different acoustical noisy conditions based on accuracy metrics is shown in table 6.7. For better readability, table 6.7 is visually presented in fig. 6.26. The number between the lines denotes the accuracy in percentage at different dB levels.

Table 6.7 Comparison of Performance of Visual-only, Audio-only and Proposed AVSR in different Acoustical Noisy Conditions

		Audio-Visual Speech Recognition						
Noise Type	Audio Noise Level	Visual-only Speech Recognition (Phonemes)	Audio-only Speech Recognition (Phonemes)	Phase 1		Phase 2		Combined ($\lambda_A = 1$ & $\lambda_V = 0$)
				Visemes from Video	Phonemes from Audio	Phonemes from Video		
Accuracy (%)								
Clean		40	96	90	96	45	96	
	20 dB	40	96	90	96	45	96	
	10 dB	40	93	90	95	45	95	
White Noise	0 dB	40	90	90	93	45	93	
	-10 dB	40	85	90	90	45	90	
	-20 dB	40	74	90	83	45	83	
	20 dB	40	96	90	96	45	96	
Pink Noise	10 dB	40	93	90	94	45	94	
	0 dB	40	90	90	92	45	92	
	-10 dB	40	85	90	88	45	88	
	-20 dB	40	75	90	83	45	83	
	20 dB	40	96	90	96	45	96	
Red Noise	10 dB	40	95	90	96	45	96	
	0 dB	40	94	90	94	45	94	
	-10 dB	40	91	90	92	45	92	
	-20 dB	40	87	90	90	45	90	

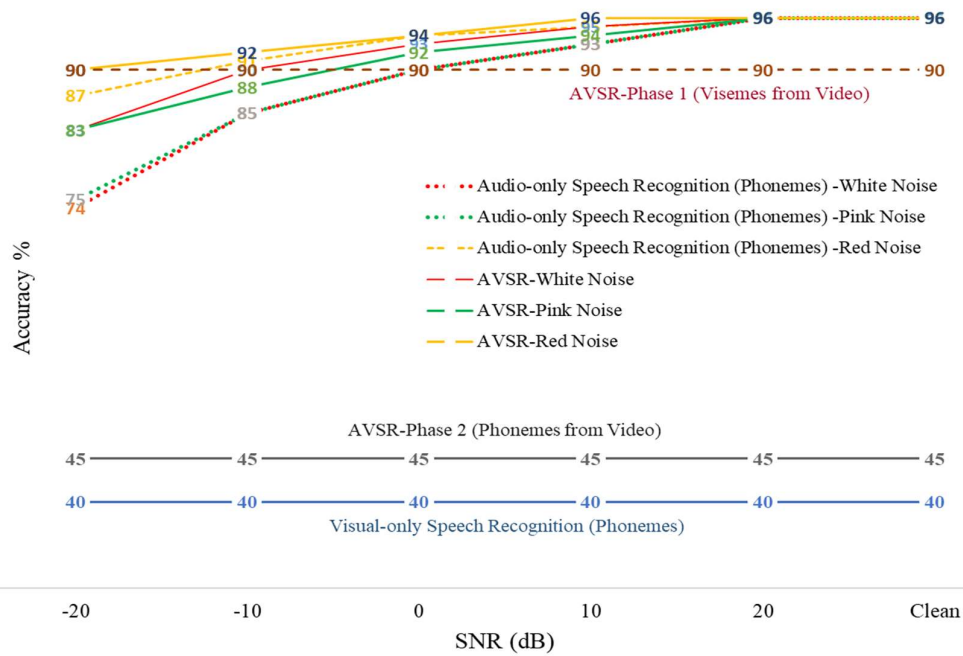


Fig. 6.26 Graphical Representation of Performance of Proposed AVSR system

6.6 Conclusions

The implementation of an Audio-Visual Malayalam speech recognition (phoneme) system using an SVM classifier is presented. Nested stratified 5-fold cross-validation is utilised to select the model hyperparameters and evaluate the model's performance. The audio-visual integration is carried out with a modified hierarchical approach and stream weight. The dataset contains two visual speech datasets comprising of 14 visemes and visual representation of 50 phonemes and one audio speech dataset comprising 50 phonemes in different noisy conditions. The performance of the audio-only, visual-only and AVSR systems is evaluated using accuracy, precision, recall and F1-score. Finally, the second phase of the proposed AVSR system is used to recognise the underlying phoneme from the phonemes of selected broad viseme using audio speech alone. Visual-only speech recognition system performs with an average of 90% in all metrics for 14 visemes. Based on the broad viseme

approach, the performance of the second phase of the AVSR system has improved the average accuracy by 9%, 8% and 3% for audio-only speech recognition in white Gaussian noise, pink noise and red noise even at -20 dB noise level. The better performance of the proposed system even at intense background noise depends on the process involved in the feature extraction algorithm and implementation part of the SVM classifier.

CHAPTER 7

CONCLUSIONS AND FUTURE DIRECTIONS

7.1 Conclusions

Recognising what is spoken, especially in noisy conditions, is significantly aided by looking at the talker's face. During the last decades, significant effort has been made to improve speech-based applications like the AVSR system's performance by using visual cues. Effective merging of noise-tolerant acoustic speech features with visual information is the crucial factor of success behind such a system. However few works, especially in resourced languages, have addressed speech recognition in intense background noise and struggled to meet satisfactory performance. This thesis presented an approach that firstly utilises the visual speech unit (viseme) to recognise the Malayalam speech (phoneme) in intense background noise, which performs better.

The backbone of any speech-based application is the availability of a phonetically balanced audio-visual speech database in the concerned language. This work created and presented an audio-visual speech database in Malayalam named "MOZHI", captured in various environments for various research goals. The first category of this database is dedicated to audio-visual speech processing in a controlled environment. Two video lamps are used to highlight the visual features of the speaker's mouth in a linguistically rich environment. The second category is dedicated to audio-only speech processing which is utilized to study the effects of real-time noise in speech processing and ageing on human speech organs by capturing data from wide age groups. The third category is dedicated to audio-visual speech processing in an acoustically and visually realistic environment. This database is intended to facilitate research in audio-visual speech processing in clean and noisy conditions, lip-reading, speech synthesis, among other topics. The first category of the database is used

in this work which consists of 30 speakers uttering 50 phonemes and 207 connected words comprising of all allophonic variations captured in a controlled environment to capture the in-depth information from the speaker's mouth. To perform the proposed work in noisy conditions, three different noises were added to the segmented and labelled clean speech with a noise level ranging from 20 dB to -20 dB. Statistical analysis on the duration of all phonemes and allophones is carried out to study thereby estimated audio-visual asynchrony to study the coarticulation nature of the Malayalam language. In Malayalam, it is observed that anticipatory coarticulation is prominent and ranges from small early audio to large early video.

Since the proposed work primarily depends on visual speech, an in-depth study of viseme is carried out to decode the visual information from the lips. A linguistic involved data-driven approach is implemented to build a connection between phoneme and viseme. Linguistic expertise is utilised to select the relevant frame of the visual appearance of the underlined phoneme. Three different visual representations were examined, out of which a viseme represented by two proceeding and following frames from the chosen frame is selected. Lip region is extracted from manually selecting two lip corner points and hue channel information from the HSV colour model and is also made rotational and translational invariant. Linguistically 50 phonemes were grouped into 14 viseme classes. This viseme set will act as the foundation for future studies of visual speech analysis in Malayalam. The data-driven approach is carried out by clustering the DCT features using K-means, and an optimum 16 viseme class were selected using the gap statistic method. Allophone-to-viseme mapping is performed by the data-driven approach, which highlights the need for linguistic classification of allophones.

After extracting relevant information from visual speech, the following procedure captures noise-robust features from audio speech. Fundamental

frequency (F0), formant frequencies (F1 and F2) and most popular Mel frequency cepstral coefficients (MFCCs) is used in this work. Instead of using these features, its noise-tolerant version was proposed by using the autocorrelation function. F0 estimation from ACR, F1 and F2 estimation from ACR Cepstrum and ACR MFCC. The proposed feature extraction method is compared with the most popular algorithms. The ACR method outperforms in white Gaussian noise but falls behind in red noise compared with Praat, YIN, RAPT and PEFAC. ACR Cepstrum is noise-robust in all noise but less precise when compared with LPC. ACR MFCC outperforms in all noise but performs similarly in clean speech when compared with MFCC. The proposed audio speech features have displayed their noise-tolerant property over the chosen methods.

The next task is to effectively merge the audio and visual speech feature by optimally selecting the stream weight based on the noisy condition, thereby improving the overall performance compared to individual stream performance. SVM classifier based audio-visual speech recognition system is proposed and compared with audio-only and video-only speech recognition systems. The empirical and structural error minimisation power of the SVM classifier, especially for noisy data and easiness on the implementation side, made to choose better than other popular models. A nested stratified 5-fold cross-validation approach is used to select an optimum value for the hyperparameters of the SVM classifier and evaluate the model performance, which minimises the overfitting issues and imbalance in the distribution of target classes. As the proposed work initially depends on the performance of visual speech, a slight variation in the visual appearance of a phoneme may wrongly judge the underlying phoneme. A modified hierarchical approach is implemented, which tackles this issue by considering a modified phoneme-to-viseme mapping or broad viseme class. To implement this approach, three different datasets are used. In the first phase, a dataset of 14 viseme classes is

utilised to identify the visual appearance of the underlying phoneme. The next phase is utilised to identify the underlying phoneme from the phoneme list of selected broad viseme classes by using two datasets: one contains 50 phonemes, and another has a visual appearance of 50 phonemes. The performance of the proposed systems is evaluated with precision, recall, F1-score and accuracy. Visual-only and audio-only speech recognition system performs with an average of 90% and 96% (clean speech) in all metrics, respectively. Based on the broad viseme approach, the performance of the second phase of the AVSR system has improved the average accuracy by 9%, 8% and 3% for audio-only speech recognition in white Gaussian noise, pink noise and red noise even at -20 dB noise level. Due to the efficient performance of extracted noise-robust acoustical features, relevant visual features and the algorithm of modified hierarchical approach nullifies the need for stream weight for decision making. This is the first work that addresses audio-visual speech recognition in Malayalam.

7.2 Future Research Directions

Even though the outcomes of this research work give promising results, certain areas are untouched or need modifications to be addressed in future research. Some of them are listed below.

- The proposed database needs an extension to incorporate continuous speech while capturing all dialectal variations of Malayalam language. Also, extend the existing database and make it available online for research purposes only.
- Category 2 and 3 from the proposed database must be utilised for its underlying applications.
- In this work, the visual speech unit (viseme) is identified by rendering the expertise from the linguistic people. In a real-time application, these

visemes are to be selected automatically from the visual speech. An initiative work towards this direction is already started but not presented in this work.

- The proposed work utilises the lips corner information to facilitate the lip segmentation process. It is necessary to automate this process and propose a prominent method invariant to shadows and other visual noises by exploring the features of the third category of the proposed database.
- The main reason for limiting this work to phoneme-level recognition is the lack of linguistic information for creating the allophone-to-viseme mapping like phoneme-to-viseme mapping. Constructing such a relation will trigger the audio-visual speech processing to word-level or even to continuous speech.
- Implementing the proposed AVSR task using Hidden Markov Model (HMM) or Deep Neural Network (DNN) in continuous speech has great potential.

REFERENCES

- [1] R. K. Moore, "Spoken language processing: Piecing together the puzzle," *Speech Communication*, vol. 49, pp. 418–435, 2007, doi: 10.1016/j.specom.2007.01.011.
- [2] Z. Peri, M. Seč, J. Nikoli, N. Simi, and T. Deli, "Speech Technology Progress Based on New Machine Learning Paradigm," *Computational intelligence and neuroscience*, 2019.
- [3] H. Mcgurk and J. Macdonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976, doi: 10.1038/264746a0.
- [4] G. Potamianos *et al.*, "Audio and visual modality combination in speech processing applications," *Handb. Multimodal-Multisensor Interfaces Found. User Model. Common Modality Comb. - Vol. 1*, pp. 489–543, 2017, doi: 10.1145/3015783.3015797.
- [5] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532-556, 1976.
- [6] S. Davis, and M. Paul, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357-366, 1980.
- [7] Sumbly, H. William, and P. Irwin, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of America*, vol. 26, no. 2, pp. 212-215, 1954.
- [8] Lee, K-F., H-W. Hon, and R. Raj, "An overview of the SPHINX speech recognition system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, pp. 35-45, 1990.
- [9] A. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," In *Nips workshop on deep learning for speech recognition and related applications*, vol. 1, no. 9, pp. 39, 2009.
- [10] J. R. Movellan, "Visual Speech Recognition with Stochastic Networks," *Adv. Neural Inf. Process. Syst.* 7, pp. 851–858, 1995.

-
- [11] C. C. Chibelushi, S. Gandon, J. S. D. Mason, F. Deravi, and R. D. Johnston, "Design issues for a digital audio-visual integrated database," *IEE Colloq.*, no. 213, pp. 2–8, 1996.
- [12] S. P.-L. Vandendorpe, "The M2VTS Multimodal Face Database," *First Int. Conf. Audio-and Video-based Biometric Pers. Authentication*, pp. 403–409, 1997.
- [13] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB : The Extended M2VTS Database University of Surrey 1 Introduction 2 Database Specification 3 The Database Acquisition System," *Proc. Second Int. Conf. audio video-based biometric Pers. authentication*, no. April 2016, pp. 1–6, 1999.
- [14] D. H. Brooks, E. L. Miller, C. A. Dimarzio, M. Kilmer, and R. J. Gaudette, "Audiovisual Speech Processing-Lip Reading and Lip Synchronization-IEEE-2001," no. November, pp. 57–75, 2001.
- [15] 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, doi: 10.1109/34.982900.
- [16] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Cuave: A new audio-visual database for multimodal human-computer interface research," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2, 2002.
- [17] C. Sanderson and K. K. Paliwal, "The vidtimit database," *Idiap Commun.*, vol. 06, pp. 2–6, 2002.
- [18] A. G. Chițu and L. J. M. Rothkrantz, "Building a data corpus for audio-visual speech recognition," *EUROMEDIA 2007 - 13th Annu. Sci. Conf. Web Technol. New Media Commun. Telemat. Theory Methods, Tools Appl. D-TV*, vol. 1, no. Movellan 1995, pp. 88–92, 2007.
- [19] B. Bailli re, Enrique, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mari thoz, and J. Matas, "The BANCA database and evaluation protocol," In *International conference on Audio-and video-based biometric person authentication*, pp. 625-638. Springer, 2003.
- [20] B. Lee *et al.*, "AVICAR: Audio-Visual Speech Corpus in a Car Environment," *8th Int. Conf. Spok. Lang. Process. ICSLP 2004*, pp. 2489–2492, 2004.
-

-
- [21] 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," *ICMI'04 - Sixth Int. Conf. Multimodal Interfaces*, pp. 235–242, 2004.
- [22] R. Göcke and J. B. Millar, "A Detailed Description of the AVOZES Data Corpus," *Technology*, pp. 486–491, 2004.
- [23] L. Liang, Y. Luo, F. Huang, and A. V. Nefian, "A multi-stream audio-video large-vocabulary Mandarin Chinese speech database," *2004 IEEE Int. Conf. Multimed. Expo*, vol. 3, pp. 1787–1790, 2004.
- [24] N. A. Fox, B. A. O'Mullane, and R. B. Reilly, "VALID: A new practical audio-visual database, and comparative results," *Lect. Notes Comput. Sci.*, vol. 3546, pp. 777–786, 2005.
- [25] P. Císař, J. Zelinka, M. Železný, A. A. Karpov, and A. L. Ronzhin, "Audio-Visual Speech Recognition for Slavonic Languages (Czech and Russian) Department of Cybernetics , University of West Bohemia in Pilsen (UWB), Czech Republic Speech Informatics Group , St . Petersburg Institute for Informatics and Automation of th," *Specom*, no. June, pp. 493–498, 2006.
- [26] 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, doi: 10.1121/1.2229005.
- [27] S. Cox, R. Harvey, and Y. Lan, "The Challenge of multispeaker lip-reading," *Audio Vis. Speech Process. AVSP, Brisbane, 2008*, 2008.
- [28] J. Trojanová, M. Hruží, P. Campr, and M. Železný, "Design and recording of Czech audio-visual database with Impaired conditions for continuous speech recognition," *Proc. 6th Int. Conf. Lang. Resour. Eval. Lr. 2008*, pp. 1239–1243, 2008.
- [29] D. Petrovska-Delacrétaz *et al.*, "The IV2 multimodal biometric database (including Iris, 2D, 3D, stereoscopic, and talking face data), and the IV2-2007 evaluation campaign," *BTAS 2008 - IEEE 2nd Int. Conf. Biometrics Theory, Appl. Syst.*, no. June, 2008, doi: 10.1109/BTAS.2008.4699323.
- [30] D. Teferi and J. Bigun, "Evaluation protocol for the DXM2VTS database and performance comparison of face detection and face tracking on video," *Proc. - Int. Conf. Pattern Recognit.*, pp. 1–4, 2008.
-

-
- [31] P. Lucey, G. Potamianos, and S. Sridharan, "(IBMSmart-Room Language DB) Patch-Based Analysis of Visual Speech from Multiple Views," *Proc. Int. Conf. Audit. Speech Process.*, no. Section 4, pp. 1–6, 2008.
- [32] X. Lin, H. Yao, X. Hong, and Q. Wang, "HIT-AVDB-II: A New Multi-view and Extreme Feature Cases Contained Audio-Visual Database for Biometrics," pp. 1–7, 2008, doi: 10.2991/jcis.2008.61.
- [33] G. Zhao, M. Barnard, and M. Pietikäinen, "Lipreading with local spatiotemporal descriptors," *IEEE Trans. Multimed.*, vol. 11, no. 7, pp. 1254–1265, 2009, doi: 10.1109/TMM.2009.2030637.
- [34] A. Vorwerk, X. Wang, D. Kolossa, S. Zeiler, and R. Orglmeister, "WAPUSK20 - A database for robust audiovisual speech recognition," *Proc. 7th Int. Conf. Lang. Resour. Eval. Lr. 2010*, pp. 3016–3019, 2010.
- [35] A. Bastanfard, M. Fazel, and A. A. Kelishami, "The Persian Linguistic Based Audio-Visual Data Corpus , AVA II , Considering Coarticulation," pp. 284–294.
- [36] Y. Benezeth and G. Bachman, "BL-Database: A French audiovisual database for speech driven lip animation systems," no. August, 2011.
- [37] Y. W. Wong *et al.*, "A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities," *Pattern Recognit. Lett.*, vol. 32, no. 13, pp. 1503–1510, 2011, doi: 10.1016/j.patrec.2011.06.011.
- [38] D. Burnham *et al.*, "Building an audio-visual Corpus of Australian English: Large Corpus collection with an economical portable and replicable black box," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. May 2014, pp. 841–844, 2011.
- [39] C. McCool *et al.*, "Bi-modal person recognition on a mobile phone: Using mobile phone data," *Proc. 2012 IEEE Int. Conf. Multimed. Expo Work. ICMEW 2012*, no. July 2012, pp. 635–640, 2012, doi: 10.1109/ICMEW.2012.116.
- [40] Y. Lan, B. J. Theobald, and R. Harvey, "View independent computer lip-reading," *Proc. - IEEE Int. Conf. Multimed. Expo*, pp. 432–437, 2012, doi: 10.1109/ICME.2012.192.
- [41] S. Antar, A. Sagheer, S. Aly, and M. F. Tolba, "AVAS: Speech database for
-

-
- multimodal recognition applications,” *13th Int. Conf. Hybrid Intell. Syst. HIS 2013*, pp. 123–128, 2014, doi: 10.1109/HIS.2013.6920467.
- [42] A. Biswas and M. Chandra, “Audio Visual Isolated Oriya Digit Recognition Using HMM and DWT,” *Conf. Adv. Commun. Control Syst. 2013*, no. April, pp. 234–238, 2013.
- [43] P. Żelasko, B. Ziółko, T. Jadczyk, and D. Skurzok, “AGH corpus of Polish speech,” *Lang. Resour. Eval.*, vol. 50, no. 3, pp. 585–601, 2016, doi: 10.1007/s10579-015-9302-y.
- [44] I. Anina, Z. Zhou, G. Zhao, and M. Pietikainen, “OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis,” *2015 11th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2015*, 2015, doi: 10.1109/FG.2015.7163155.
- [45] N. Harte and E. Gillen, “TCD-TIMIT: An audio-visual corpus of continuous speech,” *IEEE Trans. Multimed.*, vol. 17, no. 5, pp. 603–615, 2015, doi: 10.1109/TMM.2015.2407694.
- [46] P. Upadhyaya, O. Farooq, M. R. Abidi, and P. Varshney, “Comparative Study of Visual Feature for Bimodal Hindi Speech Recognition,” *Arch. Acoust.*, vol. 40, no. 4, pp. 609–619, 2015, doi: 10.1515/aoa-2015-0061.
- [47] P. Borde, “‘ vVISWa ’ – A Multilingual Multi-Pose Audio Visual Database for Robust Human Computer Interaction,” vol. 137, no. 4, pp. 25–31, 2016.
- [48] A. P. Kandagal and V. Udayashankara, “Visual Speech Recognition Based on Lip Movement for Indian Languages,” vol. 13, no. 8, pp. 2029–2041, 2017.
- [49] M. R. A. R. Maulana and M. I. Fanany, “An Audio-Visual Corpus for Multimodal Automatic Speech Recognition-2017,” *2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017*, vol. 2018-Janua, pp. 381–385, 2018, doi: 10.1109/ICACSIS.2017.8355062.
- [50] M. R. A. R. Maulana and M. I. Fanany, “Indonesian audio-visual speech corpus for multimodal automatic speech recognition,” *2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017*, vol. 2018-Janua, pp. 381–385, 2018, doi: 10.1109/ICACSIS.2017.8355062.
- [51] A. H. Abdelaziz, “NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition,” *Proc. Annu. Conf. Int. Speech*
-

-
- Commun. Assoc. INTERSPEECH*, vol. 2017-Augus, pp. 3752–3756, 2017, doi: 10.21437/Interspeech.2017-860.
- [52] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 3444–3450, 2017, doi: 10.1109/CVPR.2017.367.
- [53] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, “A corpus of audio-visual Lombard speech with frontal and profile views,” *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. EL523–EL529, 2018, doi: 10.1121/1.5042758.
- [54] Ephrat, Ariel, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation,” *arXiv preprint arXiv:1804.03619*, 2018.
- [55] L. A. Elrefaei, T. Q. Alhassan, and S. S. Omar, “An Arabic Visual Dataset for Visual Speech Recognition,” *Procedia Comput. Sci.*, vol. 163, pp. 400–409, 2019, doi: 10.1016/j.procs.2019.12.122.
- [56] L. R. Kishline, S. W. Colburn, and P. W. Robinson, “A multimedia speech corpus for audio visual research in virtual reality (L),” *J. Acoust. Soc. Am.*, vol. 148, no. 2, pp. 492–495, 2020, doi: 10.1121/10.0001670.
- [57] N. Ahmed, M. Lataifeh, and I. Junejo, “Dynamic Facial Dataset Capture and Processing for Visual Speech Recognition using an RGB-D Sensor,” vol. 47, no. 4, 2020.
- [58] A. Kashevnik *et al.*, “Multimodal Corpus Design for Audio-Visual Speech Recognition in Vehicle Cabin,” vol. 9, 2021, doi: 10.1109/ACCESS.2021.3062752.
- [59] K. Swarna and M. Shahidur Rahman, “A model of diphone duration for speech synthesis in Bangla,” *2019 Int. Conf. Bangla Speech Lang. Process. ICBSLP 2019*, no. September, pp. 27–28, 2019, doi: 10.1109/ICBSLP47725.2019.201469.
- [60] B. Chen, T. Bian, and K. Yu, “Discrete Duration Model For Speech Synthesis,” pp. 789–793, 2017.
- [61] M. O. M. Khelifa, A. Yousfi, Y. Ould, M. Elhadj, and M. Belkasm, “Enhancing Arabic Phoneme Recognizer using Duration Modeling
-

-
- Techniques,” no. December, 2016, doi: 10.15224/978-1-63248-113-9-53.
- [62] V. Delic, I. Sovilj-Nikic, and M. Markovic, "Tree-based phone duration modelling of the Serbian language," *Elektronika ir Elektrotehnika*, vol. 20, no. 3, pp. 77-82, 2014.
- [63] M. Igras and B. Zi, "Length of Phonemes in a Context of their Positions in Polish Sentences," pp. 59–64, 2013, doi: 10.5220/0004503500590064.
- [64] 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*, pp. 1-7, 2012.
- [65] K. U. Ogbureke, 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.
- [66] J. Romportl, "Prosody modelling in Czech text-to-speech synthesis Prosody Modelling in Czech Text-to-Speech Synthesis," no. May, 2014.
- [67] A. Lazaridis, P. Zervas, N. Fakotakis, and ... G., "A CART Approach for Duration Modeling of Greek Phonemes," *Proc. of SPECOM*, 2007.
- [68] G. Norkevičius, G. Raškinis, and A. Kazlauskien, "Knowledge-based grapheme-to-phoneme conversion of Lithuanian words Knowledge-based grapheme-to-phoneme conversion of Lithuanian words," no. January 2005, 2018.
- [69] J. Pylkkönen, and M. Kurimo, "Duration modeling techniques for continuous speech recognition," In *Interspeech*, 2004.
- [70] O. Şayli, L. M. Arslan, and A. S. Özsoy, "Duration properties of the Turkish phonemes," In *11th International Conference on Turkish Linguistics (ICTL 2002)*, pp. 7-9, 2002.
- [71] S. Roy, "Duration Modeling in Hindi Duration Modeling in Hindi," no. July, 2014, doi: 10.5120/17015-7296.
- [72] D. Govind, S. Mahanta, and S. R. M. Prasanna, "Significance of Duration in the Prosodic Analysis of Assamese Significance of Duration in the Prosodic Analysis of Assamese," no. February, 2015.
- [73] B. L. Kanth, V. Keri, and K. S. Prahallad, "Durational Characteristics of Indian Phonemes for Language Discrimination Durational Characteristics of Indian
-

-
- Phonemes for,” no. May, 2014, doi: 10.1007/978-3-642-19403-0.
- [74] D. P. Gopinath, J. Divya Sree, R. Mathew, S. J. Rekhila, and A. S. Nair, “Duration analysis for Malayalam text-to-speech systems,” *Proc. - 9th Int. Conf. Inf. Technol. ICIT 2006*, pp. 129–132, 2006, doi: 10.1109/ICIT.2006.48.
- [75] K. S. Rao and B. Yegnanarayana, “Modeling Syllable Duration in Indian Languages Using Support Vector,” pp. 258–263, 2005.
- [76] N. S. Krishna and H. A. Murthy, “Duration Modeling of Indian Languages Hindi and Telugu,” pp. 197–202, 2004.
- [77] K. Samudravijaya, “Durational Characteristics of Hindi Stop Consonants Tata Institute of Fundamental Research,” pp. 81–84, 2003.
- [78] S. R. Savithri, “Durational analysis of Kannada Vowels,” 1986.
- [79] P. Koziarski, T. Sadalla, S. Drgas, and A. Dabrowski, “Allophones in automatic whispery speech recognition,” *2016 21st Int. Conf. Methods Model. Autom. Robot. MMAR 2016*, pp. 811–815, 2016, doi: 10.1109/MMAR.2016.7575241.
- [80] 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)*, pp. 1-4. IEEE, 2015.
- [81] J. Xu, Y. Si, J. Pan, and Y. Yan, “Automatic allophone deriving for korean speech recognition,” *Proc. - 9th Int. Conf. Comput. Intell. Secur. CIS 2013*, pp. 776–779, 2013, doi: 10.1109/CIS.2013.169.
- [82] N. G. Fundador, T. Iwata, L. Sciences, A. Science, and K. Word, “A Comparison Between Allophone, Syllable, and Diphone Based TTS Systems for Kurdish Language-2009,” vol. 60, no. 2, pp. 5387–5388, 2011.
- [83] L. Nguyen, G. Xuefeng, and J. Makhoul, "Modeling Frequent Allophones in Japanese Speech Recognition," In *Seventh International Conference on Spoken Language Processing*. 2002.
- [84] P. A. Skrelin, “Allophone-based concatenative speech synthesis system for russian,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 1692, pp. 156–159, 1999, doi: 10.1007/3-540-48239-3_28.
-

-
- [85] P. Vivek, S. E. Pa, and V. L. Lajish, "Durational Characteristics of Allophonic Variations in Malayalam Vowel Phonemes," vol. 6, no. 3, pp. 1795–1801, 2018.
- [86] E. Marcheret, G. Potamianos, J. Vopicka, and V. Goel, "Scattering vs . Discrete Cosine Transform Features in Visual Speech Processing," pp. 175–180, 2015.
- [87] K. Kumar, G. Potamianos, J. Navratil, E. Marcheret, and V. Libal, "Audiovisual speech synchrony detection by a family of bimodal linear prediction models," *Multibiometrics Hum. Identif.*, vol. 9780521115964, pp. 31–50, 2011, doi: 10.1017/CBO9780511921056.004.
- [88] E. A. Rúa, H. Bredin, C. G. Mateo, G. Chollet, and D. G. Jiménez, "Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden markov models," *Pattern Anal. Appl.*, vol. 12, no. 3, pp. 271–284, 2009, doi: 10.1007/s10044-008-0121-2.
- [89] M. F. Woodward, and G. B. Carroll, "Phoneme perception in lipreading," *Journal of Speech and Hearing Research*, vol. 3, no. 3, pp. 212–222, 1960.
- [90] C. G. Fisher, "Confusions among visually perceived consonants.," *J. Speech Hear. Res.*, vol. 11, no. 4, pp. 796–804, 1968.
- [91] H. L. Bear, R. W. Harvey, B. J. Theobald, and Y. Lan, "Which phoneme-to-viseme maps best improve visual-only computer lip-reading?," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8888, pp. 230–239, 2014.
- [92] 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, doi: 10.1044/jshd.4104.530.
- [93] 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, doi: 10.1121/1.389537.
- [94] A. J. Goldschen, E. Petajan, G. Washington, M. Avenue, and M. Hill, "Continuous Optical Automatic Speech Recognition by Lipreading," 1995.
-

-
- [95] L. Cappelletta and N. Harte, "Viseme definitions comparison for visual-only speech recognition," *Eur. Signal Process. Conf.*, no. Eusipco, pp. 2109–2113, 2011.
- [96] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments.," *Interspeech*, pp. 1293–1296, 2003.
- [97] S. Lee and D. Yook, "Audio-to-visual conversion using hidden Markov models," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2417, pp. 563–570, 2002.
- [98] C. Aschenberner, B. Weiss, "Phoneme-Viseme Mapping for German Video-Realistic Audio-Visual-Speech-Synthesis," *Inst. für Kommun. und Phonetik, Univ. Bonn*, pp. 1–11, 2005.
- [99] J. Melenchón, J. Simó, G. Cobo, E. Martínez, A. La, and U. R. Llull, "Objective Viseme Extraction and Audiovisual Uncertainty: Estimation Limits between Auditory and Visual Modes," vol. 2007, 2007.
- [100] A. G. Chitu and L. J. M. Rothkrantz, "Visual Speech Recognition Automatic System for Lip Reading of Dutch," *Inf. Technol. Control*, vol. year viii, no. 3, pp. 2–9, 2009.
- [101] P. Damien, N. Wakim, and M. Egéa, "Phoneme-viseme mapping for modern, classical arabic language," *2009 Int. Conf. Adv. Comput. Tools Eng. Appl. ACTEA 2009*, vol. 2, no. 1, pp. 547–552, 2009, doi: 10.1109/ACTEA.2009.5227875.
- [102] D. Yu, O. Ghita, A. Sutherland, and P. F. Whelan, "A novel visual speech representation and HMM classification for visual speech recognition," *IPSJ Trans. Comput. Vis. Appl.*, vol. 2, pp. 25–38, 2010, doi: 10.2197/ipsjtcva.2.25.
- [103] F. Z. Chelali, K. Sadeddine, and A. Djeradi, "Visual speech analysis application to Arabic phonemes," *Special Issue of International Journal of Computer Applications (0975–8887) on Software Engineering, Databases and Expert Systems-SEDEXS*, pp. 29–34, 2012.
- [104] 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, doi: 10.1016/j.specom.2013.02.005.
-

-
- [105] T. Seko, N. Ukai, S. Tamura, and S. Hayamizu, "Improvement of Lipreading Performance Using Discriminative Feature and Speaker Adaptation," *Avsp*, 2013.
- [106] M. Aghaahmadi, M. M. Dehshibi, A. Bastanfard, and M. Fazlali, "Clustering Persian viseme using phoneme subspace for developing visual speech application," *Multimed. Tools Appl.*, vol. 65, no. 3, pp. 521–541, 2013, doi: 10.1007/s11042-012-1128-7.
- [107] P. Varshney, O. Farooq, and P. Upadhyaya, "Hindi viseme recognition using subspace DCT features," *Int. J. Appl. Pattern Recognit.*, vol. 1, no. 3, p. 257, 2014, doi: 10.1504/ijapr.2014.065768.
- [108] H. L. Bear, R. W. Harvey, and Y. Lan, "Finding phonemes: improving machine lip-reading," *arXiv preprint arXiv:1710.01142*, pp. 115-120, 2017.
- [109] S. Taylor, B. J. Theobald, and I. Matthews, "A mouth full of words: Visually consistent acoustic redubbing," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2015-Augus, pp. 4904–4908, 2015, doi: 10.1109/ICASSP.2015.7178903.
- [110] 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, pp. 1–12, 2015.
- [111] A. Brahme and U. Bhadade, "Phoneme visem mapping for Marathi language using linguistic approach," *Proc. - Int. Conf. Glob. Trends Signal Process. Inf. Comput. Commun. ICGTSPICC 2016*, pp. 152–157, 2017, doi: 10.1109/ICGTSPICC.2016.7955288.
- [112] 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, doi: 10.1007/s11760-014-0615-x.
- [113] M. A. Bakhshali and M. Shamsi, "Segmentation of color lip images by optimal thresholding using bacterial foraging optimization (BFO)," *J. Comput. Sci.*, vol. 5, no. 2, pp. 251–257, 2014, doi: 10.1016/j.jocs.2013.07.001.
- [114] 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, doi: 10.1007/s11760-015-0834-9.
-

-
- [115] 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, doi: 10.1109/TIP.2003.818116.
- [116] S. Lucey, S. Sridharan, and V. Chandran, "Chromatic lip tracking using a connectivity based fuzzy thresholding technique," *ISSPA 1999 - Proc. 5th Int. Symp. Signal Process. Its Appl.*, vol. 2, pp. 669–672, 1999, doi: 10.1109/ISSPA.1999.815761.
- [117] M. Li and Y. M. Cheung, "Automatic segmentation of color lip images based on morphological filter," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6352 LNCS, no. PART 1, pp. 384–387, 2010, doi: 10.1007/978-3-642-15819-3_51.
- [118] 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, doi: 10.1109/TCSVT.2004.826754.
- [119] M. Lievin, P. Delmas, P. Y. Coulon, F. Luthon, and V. Fristot, "Automatic lip tracking: Bayesian segmentation and active contours in a cooperative scheme," *Int. Conf. Multimed. Comput. Syst. -Proceedings*, vol. 1, pp. 691–696, 1999.
- [120] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 266–277, 2001, doi: 10.1109/83.902291.
- [121] P. Gacon, P. Y. Coulon, and G. Bailly, "Non-linear active model for mouth inner and outer contours detection," *13th Eur. Signal Process. Conf. EUSIPCO 2005*, no. 1, pp. 1111–1114, 2005.
- [122] J. Luetin, N. A. Thacker, and S. W. Beet, "Active Shape Models for Visual Speech Feature Extraction," no. 95, pp. 383–390, 1996, doi: 10.1007/978-3-662-13015-5_28.
- [123] 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," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 3, pp. 1485–1488, 2001.
- [124] P. Dalka, P. Bratoszewski, and A. Czyzewski, "Visual lip contour detection for the purpose of speech recognition," *2014 Int. Conf. Signals Electron. Syst. ICSES 2014*, pp. 1–4, 2014, doi: 10.1109/ICSES.2014.6948716.
-

- [125] I. O. P. C. Series and M. Science, "Feature extraction for face recognition via Active Shape Model (ASM) and Active Appearance Model (AAM) Feature extraction for face recognition via Active Shape Model (ASM) and Active Appearance Model (AAM)," 2018, doi: 10.1088/1757-899X/332/1/012032.
- [126] L. B. Babu, "Continuous Speech Recognition System for Malayalam Language Using Kaldi," *2018 Int. Conf. Emerg. Trends Innov. Eng. Technol. Res.*, pp. 1–4, 2018.
- [127] S. K. Mukundan, "Shreshta Bhasha ' Malayalam Speech Recognition using HTK," vol. 1, no. 1, pp. 1–5, 2014.
- [128] Rajis. U. A and Babu. Anto. P, "Intelligent Query Processing in Malayalam," no. 2, pp. 51–59, 2013.
- [129] C. Kurian and K. Balakrishnan, "Connected digit speech recognition system for Malayalam language," *Sadhana - Acad. Proc. Eng. Sci.*, vol. 38, no. 6, pp. 1339–1346, 2013, doi: 10.1007/s12046-013-0160-2.
- [130] A. V Anand, P. S. Devi, J. Stephen, and V. K. Bhadrans, "Malayalam Speech Recognition System and Its Application for visually impaired people," pp. 619–624, 2012, doi: 10.1109/INDCON.2012.6420692.
- [131] V. R. Vimal Krishnan, A. Jayakumar, and P. Babu Anto, "Speech recognition of isolated Malayalam words using wavelet features and Artificial Neural Network," *Proc. - 4th IEEE Int. Symp. Electron. Des. Test Appl. DELTA 2008*, pp. 240–243, 2008, doi: 10.1109/DELTA.2008.88.
- [132] A. N. Mishra, N. Awasthy, V. Verma, and S. Malhotra, "Hindi speech audio visual feature recognition," vol. 29, no. 5, pp.1734-1743, 2020.
- [133] A. N. Mishra, M. Chandra, A. Biswas, and S. N. Sharan, "Hindi phoneme-viseme recognition from continuous speech," *Int. J. Signal Imaging Syst. Eng.*, vol. 6, no. 3, pp. 164–171, 2013, doi: 10.1504/IJSISE.2013.054793.
- [134] B. J. Shannon and K. K. Paliwal, "Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition," *Speech Commun.*, vol. 48, no. 11, pp. 1458–1485, 2006, doi: 10.1016/j.specom.2006.08.003.
- [135] H. Meutzner, N. Ma, R. Nickel, C. Schymura, and D. Kolossa, "Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal*
-

-
- Process. - Proc.*, pp. 5320–5324, 2017, doi: 10.1109/ICASSP.2017.7953172.
- [136] M. Gurban, J. Thiran, and S. Member, “Information Theoretic Feature Extraction for Audio-Visual Speech Recognition,” vol. 57, no. 12, pp. 4765–4776, 2009.
- [137] W. Yu, S. Zeiler, and D. Kolossa, "Large-vocabulary Audio-visual Speech Recognition in Noisy Environments," *arXiv preprint arXiv:2109.04894*, 2021.
- [138] K. Thangthai, R. Harvey, S. Cox, and B. Theobald, “Improving Lip-reading Performance for Robust Audiovisual Speech Recognition using DNN s,” pp. 127–131, 2015.
- [139] A. S. Saudi, M. I. Khalil, and H. M. Abbas, “Improved features and dynamic stream weight adaption for robust Audio-Visual Speech Recognition framework,” *Digit. Signal Process.*, vol. 1, 2019, doi: 10.1016/j.dsp.2019.02.016.
- [140] I. Mcloughlin, Z. Xie, Y. Song, H. Phan, and R. Palaniappan, “Time – Frequency Feature Fusion for Noise Robust Audio Event Classification,” *Circuits, Syst. Signal Process.*, 2019, doi: 10.1007/s00034-019-01203-0.
- [141] V. Estellers, M. Gurban, and J. P. Thiran, “On dynamic stream weighting for audio-visual speech recognition,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 4, pp. 1145–1157, 2012, doi: 10.1109/TASL.2011.2172427.
- [142] H. L. Bear and R. Harvey, “Phoneme-to-viseme mappings: the good, the bad, and the ugly,” *Speech Commun.*, vol. 95, pp. 40–67, 2017, doi: 10.1016/j.specom.2017.07.001.
- [143] U. Bhattacharjee, S. Gogoi, and R. Sharma, “A statistical analysis on the impact of noise on MFCC features for speech recognition,” *2016 Int. Conf. Recent Adv. Innov. Eng. ICRAIE 2016*, 2017, doi: 10.1109/ICRAIE.2016.7939548.
- [144] H. L. Bear and R. Harvey, “DECODING VISEMES: IMPROVING MACHINE LIP-READING Helen L . Bear and Richard Harvey,” *Icassp 2016*, pp. 2009–2013, 2016.
- [145] S. L. Taylor, M. Mahler, B. J. Theobald, and I. Matthews, “Dynamic units of visual speech,” *Comput. Animat. 2012 - ACM SIGGRAPH / Eurographics*
-

-
- Symp. Proceedings, SCA 2012*, pp. 275–284, 2012.
- [146] H. L. Bear and R. Harvey, “Comparing heterogeneous visual gestures for measuring the diversity of visual speech signals,” *Comput. Speech Lang.*, vol. 52, pp. 165–190, 2018, doi: 10.1016/j.csl.2018.05.001.
- [147] P. Lucey and G. Potamianos, “Lipreading using profile versus frontal views,” *2006 IEEE 8th Work. Multimed. Signal Process. MMSP 2006*, pp. 24–28, 2007, doi: 10.1109/MMSP.2006.285261.
- [148] A. Blokland and A. H. Anderson, “Effect of low frame-rate video on intelligibility of speech,” *Speech Commun.*, vol. 26, no. 1–2, pp. 97–103, 1998, doi: 10.1016/S0167-6393(98)00053-3.
- [149] T. Saitoh and R. Konishi, “A Study of Influence of Word Lip-reading by Change of Frame Rate,” *Word J. Int. Linguist. Assoc.*, pp. 400–407, 2010.
- [150] D. Jachimski, A. Czyzewski, and T. Ciszewski, “A comparative study of English viseme recognition methods and algorithms,” *Multimed. Tools Appl.*, vol. 77, no. 13, pp. 16495–16532, 2018, doi: 10.1007/s11042-017-5217-5.
- [151] D. S. Alexandre and J. M. R. S. Tavares, “Introduction of human perception in visualization,” *Int. J. Imaging*, vol. 4, no. 10 A, pp. 60–70, 2010.
- [152] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010, doi: 10.1016/j.patrec.2009.09.011.
- [153] M. Mohajer, K.-H. Englmeier, and V. J. Schmid, "A comparison of Gap statistic definitions with and without logarithm function," *arXiv preprint arXiv:1103.4767*, 2011.
- [154] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel, “Towards unrestricted lip reading,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 14, no. 5, pp. 571–585, 2000, doi: 10.1142/S0218001400000374.
- [155] S. D. Lalitha and K. K. Thyagarajan, “A study on lip localization techniques used for lip reading from a video,” *Int. J. Appl. Eng. Res.*, vol. 11, no. 1, pp. 611–615, 2016.
- [156] C. Berry and N. Harte, “Region of interest extraction using colour based methods on the CUAVE Database,” pp. 38–38, 2010, doi: 10.1049/cp.2009.1715.
- [157] B. D. Baswaraj, a Govardhan, and P. Premchand, “Active Contours and Image
-

-
- Segmentation: The Current State of the Art,” *Glob. J. Comput. Sci. Technol. Graph. Vis.*, vol. 12, no. 11, 2012.
- [158] R. Yogamangalam and B. Karthikeyan, “Segmentation Techniques Comparison in Image Processing,” *Int. J. Eng. Technol.*, vol. 5, no. 1, pp. 307–313, 2013.
- [159] A. ElMaghraby *et al.*, “Detect and Analyze Face Parts Information using Viola- Jones and Geometric Approaches,” *Int. J. Comput. Appl.*, vol. 101, no. 3, pp. 23–28, 2014, doi: 10.5120/17667-8494.
- [160] P. Kuo, P. Hillman, and J. Hannah, “Improved lip fitting and tracking for model-based multimedia and coding,” *IEE Int. Conf. Vis. Inf. Eng. (VIE 2005)*, no. 2005–10882, pp. 251–258, 2005.
- [161] Y. Lan, R. W. Harvey, B.-J. Theobald, E.-J. Ong, and R. Bowden, “Comparing Visual Features for Lipreading,” *Avsp 2009*, no. 2, pp. 102–106, 2009.
- [162] A. A. Abdulrahman, “Edge detection for lips area using RGB color space,” no. 209, pp. 149–158, 2014.
- [163] D. J. Liu and C. T. Lin, “Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 609–621, 2001, doi: 10.1109/89.943339.
- [164] A. B. A. Hassanat and S. Jassim, “Color-based lip localization method,” *Mob. Multimedia/Image Process. Secur. Appl. 2010*, vol. 7708, no. April, p. 77080Y, 2010, doi: 10.1117/12.850629.
- [165] 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, doi: 10.1121/1.2936016.
- [166] D. Stewart, R. Seymour, and J. Ming, “Comparison of image transform-based features for visual speech recognition in clean and corrupted videos,” *Eurasip J. Image Video Process.*, vol. 2008, no. 2008, pp. 1–9, 2008, doi: 10.1155/2008/810362.
- [167] P. S. Aleksic and A. K. Katsaggelos, “Comparison of low- and high-level visual features for audio-visual continuous automatic speech recognition,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 5, 2004, doi: 10.1109/icassp.2004.1327261.
-

-
- [168] D. Websdale and B. Milner, "Analysing the importance of different visual feature coefficients," *Faavsp*, no. 3, pp. 137–142, 2015.
- [169] O. Farooq, P. Upadhyaya, O. Farooq, P. Varshney, and A. Upadhyaya, "Enhancement of VSR Using Low Dimension Visual Feature Enhancement of VSR Using Low Dimension Visual Feature," no. November, 2013, doi: 10.1109/MSPCT.2013.6782090.
- [170] N. Puviarasan and S. Palanivel, "Lip reading of hearing impaired persons using HMM," *Expert Syst. Appl.*, vol. 38, no. 4, pp. 4477–4481, 2011, doi: 10.1016/j.eswa.2010.09.119.
- [171] S. S. Morade, "Visual Lip Reading using 3D-DCT and 3D-DWT and LSDA," vol. 136, no. 4, pp. 7–15, 2016.
- [172] S. S. Morade and S. Patnaik, "Lip Reading by Using 3-D Discrete Wavelet Transform with Dmey Wavelet," *Int. J. Image Process.*, no. 8, pp. 384–396, 2014.
- [173] R. Rajavel and P. S. Sathidevi, "Static and Dynamic Features for Improved HMM based Visual Speech Recognition," *Proc. First Int. Conf. Intell. Hum. Comput. Interact.*, pp. 184–194, 2009, doi: 10.1007/978-81-8489-203-1_17.
- [174] H. Xiaopeng, Y. Hongxun, W. Yuqi, and C. Rong, "A PCA based visual DCT feature extraction method for lip-reading," *Proc. - 2006 Int. Conf. Intell. Inf. Hiding Multimed. Signal Process. IHH-MSP 2006*, no. December 2006, pp. 321–324, 2006, doi: 10.1109/IHH-MSP.2006.265008.
- [175] S. Alizadeh, R. Boostani, and V. Asadpour, "Lip feature extraction and reduction for hmm-based visual speech recognition systems," *Int. Conf. Signal Process. Proceedings, ICSP*, pp. 561–564, 2008, doi: 10.1109/ICOSP.2008.4697195.
- [176] J. He and H. Zhang, "Research on visual speech feature extraction," *Proc. - 2009 Int. Conf. Comput. Eng. Technol. ICCET 2009*, vol. 2, pp. 499–502, 2009, doi: 10.1109/ICCET.2009.63.
- [177] A. Biswas, P. K. Sahu, A. Bhowmick, and M. Chandra, "VidTIMIT audio visual phoneme recognition using AAM visual features and human auditory motivated acoustic wavelet features," *2015 IEEE 2nd Int. Conf. Recent Trends Inf. Syst. ReTIS 2015 - Proc.*, no. 2004, pp. 428–433, 2015, doi: 10.1109/ReTIS.2015.7232917.
-

-
- [178] N. Li, N. Lefebvre, and R. Lengellé, “Kernel hierarchical agglomerative clustering: Comparison of different gap statistics to estimate the number of clusters,” *ICPRAM 2014 - Proc. 3rd Int. Conf. Pattern Recognit. Appl. Methods*, no. January, pp. 255–262, 2014, doi: 10.5220/0004828202550262.
- [179] M. G. H. Omran, A. P. Engelbrecht, and A. Salman, “An overview of clustering methods,” *Intell. Data Anal.*, vol. 11, no. 6, pp. 583–605, 2007.
- [180] S. Miglani and K. Garg, “Factors Affecting Efficiency of K-means Algorithm,” vol. 2, pp. 85–87, 2013.
- [181] T. Tibshirani, R. Walther, G., & Hastie, “Estimating the number of clusters in a data set via the gap statistic,” *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, vol. 63, no. 2, pp. 411–423., 2001.
- [182] M. McLaren and Y. Lei, "Improved speaker recognition using DCT coefficients as features," In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4430–4434, 2015.
- [183] S. Hilder, B. Theobald, and R. Harvey, “In pursuit of visemes,” *Proc. Int. Conf. Audit. Speech Process.*, pp. 154–159, 2010.
- [184] A. K. Katsaggelos, S. Bahaadini, and R. Molina, “Audiovisual Fusion: Challenges and New Approaches,” *Proc. IEEE*, vol. 103, no. 9, pp. 1635–1653, 2015, doi: 10.1109/JPROC.2015.2459017.
- [185] P. Borde, A. Varpe, R. Manza, and P. Yannawar, “Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition,” *Int. J. Speech Technol.*, vol. 18, no. 2, pp. 167–175, 2015, doi: 10.1007/s10772-014-9257-1.
- [186] F. Kurth, A. Cornaggia-Urrigshardt, and S. Urrigshardt, “Robust F0 estimation in noisy speech signals using shift autocorrelation,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 1468–1472, 2014, doi: 10.1109/ICASSP.2014.6853841.
- [187] S. Strömbergsson, “Today’s most frequently used F0 estimation methods, and their accuracy in estimating male and female pitch in clean speech,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 525–529, 2016, doi: 10.21437/Interspeech.2016-240.
- [188] H. De Cheveigné, A., & Kawahara, “YIN, a fundamental frequency estimator
-

-
- for speech and music,” *Int. J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002, doi: 10.1121/1.1458024.
- [189] D. Talkin, W. B. Kleijn, and K. K. Paliwal, “A Robust Algorithm for Pitch Tracking (RAPT),” *Speech Coding Synth. Eds. Amsterdam, NetherlandsElsevier*, pp. 495–518, 1995.
- [190] S. Gonzalez and M. Brookes, “PEFAC - A pitch estimation algorithm robust to high levels of noise,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 22, no. 2, pp. 518–530, 2014, doi: 10.1109/TASLP.2013.2295918.
- [191] K. Kasi and S. A. Zahorian, “Yet another algorithm for pitch tracking,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 1, pp. 361–364, 2002, doi: 10.1109/icassp.2002.5743729.
- [192] X. Sun, “Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 1, pp. 333–336, 2002, doi: 10.1109/icassp.2002.5743722.
- [193] T. Drugman and A. Alwan, “Joint robust voicing detection and pitch estimation based on residual harmonics,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 1973–1976, 2011.
- [194] A. Camacho and J. G. Harris, “A sawtooth waveform inspired pitch estimator for speech and music,” *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, 2008, doi: 10.1121/1.2951592.
- [195] H. Huang and J. Pan, “Speech pitch determination based on Hilbert-Huang transform,” *Signal Processing*, vol. 86, no. 4, pp. 792–803, 2006, doi: 10.1016/j.sigpro.2005.06.011.
- [196] R. Daido and Y. Hisaminato, “A fast and accurate fundamental frequency estimator using recursive moving average filters,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 2160–2164, 2016, doi: 10.21437/Interspeech.2016-394.
- [197] L. N. Tan and A. Alwan, “Multi-band summary correlogram-based pitch detection for noisy speech,” *Speech Commun.*, vol. 55, no. 7–8, pp. 841–856, 2013, doi: 10.1016/j.specom.2013.03.001.
- [198] W. Chu and A. Alwan, “SAFE: A statistical approach to F0 estimation under clean and noisy conditions,” *IEEE Trans. Audio, Speech Lang. Process.*, vol.
-

-
- 20, no. 3, pp. 933–944, 2012, doi: 10.1109/TASL.2011.2168518.
- [199] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirovic, “SPICE: Self-Supervised Pitch Estimation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1118–1128, 2020, doi: 10.1109/taslp.2020.2982285.
- [200] J. Tabrikian, S. Dubnov, and Y. Dickalov, “Maximum A-Posteriori Probability Pitch Tracking in Noisy Environments Using Harmonic Model,” *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76–87, 2004, doi: 10.1109/TSA.2003.819950.
- [201] M. Wu, D. L. Wang, and G. J. Brown, “A multipitch tracking algorithm for noisy speech,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 229–241, 2003, doi: 10.1109/TSA.2003.811539.
- [202] K. Han and D. L. Wang, “Neural network based pitch tracking in very noisy speech,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 2158–2168, 2014, doi: 10.1109/TASLP.2014.2363410.
- [203] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, “Crepe: A Convolutional Representation for Pitch Estimation,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 161–165, 2018, doi: 10.1109/ICASSP.2018.8461329.
- [204] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, “A pitch tracking corpus with evaluation on multipitch tracking scenario,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. August, pp. 1509–1512, 2011.
- [205] W. A. Plante, F., Meyer, G. F., & Ainsworth, “A pitch extraction reference database,” *Fourth Eur. Conf. Speech Commun. Technol.*, no. September, pp. 837–840, 1995.
- [206] “Enhanced pitch tracking and the processing of f0 contours for computer and intonation teaching,” p. 400, 1993.
- [207] Q. Zhao, T. Shimamura, and J. Suzuki, “a Robust Algorithm for Formant Frequency,” *Computer (Long. Beach. Calif.)*, pp. 534–537, 1998.
- [208] M. A. Kammoun, D. Gargouri, M. Frikha, and A. Ben Hamida, “Cepstrum vs. LPC: A Comparative Study for Speech Formant Frequencies Estimation,” *GESTS Int’l Trans. Commun. Signal Process.*, vol. 9, no. 1, pp. 87–102, 2006.
-

-
- [209] J. Le Roux, H. Kameoka, N. Ono, A. De Cheveigné, and S. Sagayama, “Single and multiple F0 contour estimation through parametric spectrogram modeling of speech in noisy environments,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 4, pp. 1135–1145, 2007, doi: 10.1109/TASL.2007.894510.
- [210] M. Chougala and K. Shridhar, “Novel Formant Estimation Techniques for Speech Processing,” vol. 3, no. 6, pp. 229–234, 2015.
- [211] M. A. Kammoun, D. Gargouri, M. Frikha, and A. Ben Hamida, “Cepstral method evaluation in speech formant frequencies estimation,” *Proc. IEEE Int. Conf. Ind. Technol.*, vol. 3, no. January, pp. 1612–1616, 2004, doi: 10.1109/ICIT.2004.1490808.
- [212] S. A. Majeed, H. Husain, S. A. Samad, and T. F. Idbeaa, “Mel frequency cepstral coefficients (Mfcc) feature extraction enhancement in the application of speech recognition: A comparison study,” *J. Theor. Appl. Inf. Technol.*, vol. 79, no. 1, pp. 38–56, 2015.
- [213] A. Carlos and G. Thomé, “SVM Classifiers – Concepts and Applications to Character Recognition,” 2012.
- [214] G. Maragatham, and S. Rajendran, "Improving the classifier accuracy with an integrated approach using medical data—a study," *International Journal of Medical Engineering and Informatics*, vol. 12, no. 4, pp. 313-321, 2020.
- [215] M. Frikha and A. Ben Hamida, “A Comparative Survey of ANN and Hybrid HMM/ANN Architectures for Robust Speech Recognition,” *Am. J. Intell. Syst.*, vol. 2, no. 1, pp. 1–8, 2012, doi: 10.5923/j.ajis.20120201.01.
- [216] N. Smith and M. Gales, “Speech recognition using SVMS,” *Adv. Neural Inf. Process. Syst.*, 2002.
- [217] C. Y. Fook, H. Muthusamy, L. S. Chee, S. Bin Yaacob, and A. H. B. Adom, “Comparison of speech parameterization techniques for the classification of speech disfluencies,” *Turkish J. Electr. Eng. Comput. Sci.*, vol. 21, no. SUPPL. 1, pp. 1983–1994, 2013, doi: 10.3906/elk-1112-84.
- [218] S. N. Srihari, “Machine Learning : Generative and Discriminative Models”, 2010.
- [219] M. Sarma and K. K. Sarma, “Recent Trends in Intelligent and Emerging Systems,” no. May 2015, pp. 173–187, 2015, doi: 10.1007/978-81-322-2407-5.
-

-
- [220] R. Mukherjee, T. Islam, and R. Sankar, "Text dependent speaker recognition using shifted MFCC," *Conf. Proc. - IEEE SOUTHEASTCON*, pp. 1–4, 2013, doi: 10.1109/SECON.2013.6567398.
- [221] K. Saenko, K. Livescu, J. Glass, and T. Darrell, "Production domain modeling of pronunciation for visual speech recognition," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. V, pp. 473–476, 2005, doi: 10.1109/ICASSP.2005.1416343.
- [222] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Learning dynamic stream weights for coupled-HMM-based audio-visual speech recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 5, pp. 863–876, 2015, doi: 10.1109/TASLP.2015.2409785.
- [223] H. Pahuja, P. Ranjan, and A. Ujlayan, "Audio visual automatic speech recognition using multi-tasking learning of deep neural networks," *2017 Int. Conf. Infocom Technol. Unmanned Syst. Trends Futur. Dir. ICTUS 2017*, vol. 2018-Janua, no. December 2017, pp. 455–458, 2018, doi: 10.1109/ICTUS.2017.8286052.
- [224] A. Mohamed and K. N. R. Nair, "HMM/ANN hybrid model for continuous Malayalam speech recognition," *Procedia Eng.*, vol. 30, pp. 616–622, 2012, doi: 10.1016/j.proeng.2012.01.906.
- [225] M. Gurban and J. P. Thiran, "Audio-visual speech recognition with a hybrid SVM-HMM system," *13th Eur. Signal Process. Conf. EUSIPCO 2005*, pp. 728–731, 2005.
- [226] R. Solera-Ureña, J. Padrell-Sendra, D. Martín-Iglesias, A. Gallardo-Antolín, C. Peláez-Moreno, and F. Díaz-de-María, "SVMs for automatic speech recognition: A survey," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4391 LNCS, pp. 190–216, 2007, doi: 10.1007/978-3-540-71505-4_11.
- [227] A. Ganapathiraju, J. E. Hamaker, and J. Picone, "Applications of support vector machines to speech recognition," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2348–2355, 2004, doi: 10.1109/TSP.2004.831018.
- [228] B. A. Sonkamble and D. D. Doye, "An overview of speech recognition system based on the support vector machines," *Proc. Int. Conf. Comput. Commun. Eng. 2008, ICCCE08 Glob. Links Hum. Dev.*, pp. 768–771, 2008, doi: 10.1109/ICCCE.2008.4580709.
-

-
- [229] K. Aida-Zade, A. Xocayev, and S. Rustamov, "Speech recognition using Support Vector Machines," *Appl. Inf. Commun. Technol. AICT 2016 - Conf. Proc.*, vol. 1, 2017, doi: 10.1109/ICAICT.2016.7991664.
- [230] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey," *Proc. IEEE*, vol. 98, no. 10, pp. 1692–1715, 2010, doi: 10.1109/JPROC.2010.2057231.
- [231] N. Tatbul, M. Buller, R. Hoyt, S. Mullen, and S. Zdonik, "Confidence-based data management for personal area sensor networks," in *ACM International Conference Proceeding Series*, 2004, vol. 72, no. January 2015, pp. 24–31, doi: 10.1145/1052199.1052204.
- [232] Hsu, WH-M., and Shih-Fu Chang, "Generative, discriminative, and ensemble learning on multi-modal perceptual fusion toward news video story segmentation," In *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, vol. 2, pp. 1091-1094, 2004.
- [233] J.-S. Lee and C. Hoon, "Adaptive Decision Fusion for Audio-Visual Speech Recognition," *Speech Recognit.*, no. November, 2008, doi: 10.5772/6364.
- [234] M. Gurban, and J.-P. Thiran, "Using entropy as a stream reliability estimate for audio-visual speech recognition," In *2008 16th European Signal Processing Conference*, pp. 1-5, 2008.
- [235] S. Pachoud, G. Shaogang, and A. Cavallaro, "Space-time audio-visual speech recognition with multiple multi-class probabilistic support vector machines," In *AVSP*, pp. 155-160, 2009.
- [236] J. Platt and others, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. large margin Classif.*, vol. 10, no. 3, pp. 61–74, 1999.
- [237] and C.-J. L. Chih-Wei Hsu, Chih-Chung Chang, "A Practical Guide to Support Vector Classification-2016," *Theory, Cult. Soc.*, vol. 17, no. 1, pp. 39–61, 2016, doi: 10.1177/02632760022050997.
- [238] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, 2010.
-

LIST OF PUBLICATIONS

Book Chapter

1. **Bibish Kumar, K. T.**, Sunil John, Muraleedharan, K. M., & Sunil Kumar, R. K. (2021). “*Linguistically involved Data-driven Approach for Malayalam Phoneme-to-Viseme Mapping*”. In Applied Speech Processing (pp. 117-145). Academic Press.
<https://doi.org/10.1016/B978-0-12-823898-1.00003-5>

Journals

1. **Bibish Kumar, K. T.**, Sunil Kumar, R. K., Sandesh, E. PA., & Lajish, V. L. (2020). “*A Comparative study of lip region segmentation in different colour space for lip reading in Indian context*”. International Journal of Tomography and Simulation, 33(1), 73-89.
2. **Bibish Kumar, K. T.**, Sunil Kumar, R. K., Sandesh, E. P.A. et al. (2019). “*Viseme set identification from Malayalam phonemes and allophones*”. International Journal of Speech Technology, 22(4), 1149–1166.
<https://doi.org/10.1007/s10772-019-09655-0>
3. **Bibish Kumar, K. T.**, Sunil John, Muraleedharan, K. M., & Sunil Kumar, R. K. (2019). “*Hierarchical Picture of existing Audio-Visual Speech Database*”. International Journal of Recent Technology and Engineering, 8(3), 8372-8379.
<https://doi.org/10.35940/ijrte.C6483.098319>
4. **Bibish Kumar, K. T.**, Sunil John, Muraleedharan, K. M., & Sunil Kumar, R. K. (2019). “*Audio-Visual Asynchrony in Malayalam phonemes and allophones*”. International Journal of Recent Technology and Engineering, 8(3), 8359-8362.
<https://doi.org/10.35940/ijrte.C6468.098319>

5. Sandesh, E. PA., and Lajish, V.L, **Bibish Kumar, K. T.**, and Sunil Kumar, R. K. (2018). “*A Comparative Study of Colour Spaces for Mouth Region Segmentation in Indian Context*”. International Journal of Research in Advent Technology, 6(8), 2108-2117.
6. **Bibish Kumar, K. T.**, Sunil Kumar, R. K., Sunil John, & Muraleedharan, K. M. “MOZHI-An Audio-Visual Malayalam Speech Database for diverse Speech-based Applications”. (Communicated to International Journal of Computer Speech and Language, Under Review.)

Conferences

1. **Bibish Kumar, K. T.**, Sunil John, Muraleedharan, K. M., & Sunil Kumar, R. K. (2021). “*Viseme Classification Using Support Vector Machines*”. 2nd International Conference on Smart Technologies for Smart Nation (SmartTechCon 2021), Malaysia. (Communicated)
2. Aljinu Khadar, K. V., **Bibish Kumar, K. T.**, Muraleedharan. K.M., Sunil John., and Sunil Kumar, R. K. (2020). “*Vocal Tract Length and Formant Frequency Analysis for Forensic Speech Applications*”. National Conference on Computational Intelligence, Practices and Technologies (ConCIPT 2020) organized by Department of Computer Science, University of Calicut, Kerala, for 9 -11 January 2020.
3. **Bibish Kumar, K. T.**, Sunil Kumar, R. K., Sandesh, E. PA., and Lajish, V.L. (2017). “*Color Thresholding based Approaches to Lip Segmentation for Visual Speech Recognition*”. 7th National Conference on Indian Language Computing (NCILC 2017), organized by Department of Computer Applications, CUSAT, Kerala, during 17-18 February 2017.