HIDDEN MARKOV MODEL BASED KEYWORD SPOTTING FOR MALAYALAM SPEECH ANALYTICS

A Thesis Submitted by

VIVEK P.

Under the Guidance of

Dr. LAJISH V. L.

In Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY
in
COMPUTER SCIENCE

Under the Faculty of Science

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF CALICUT
KERALA, INDIA – 673 635

NOVEMBER 2017

UNIVERSITYOF CALICUT
DEPARTMENT OF COMPUTER SCIENCE

Dr. Lajish. V. L.
Assistant Professor

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

Certificate

This is to certify that the thesis entitled **"HIDDEN MARKOV MODEL BASED KEYWORD SPOTTING FOR MALAYALAM SPEECH ANALYTICS "** is a report of the original work carried out by Mr. VIVEK P. under my supervision and guidance in the Department of Computer Science, University of Calicut, Kerala and that no part thereof has been presented for the award of any other degree.

Calicut University
November 6, 2017

Dr. LAJISH. V.L
(Supervising Guide)

Declaration

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

Calicut University
November 6, 2017

Vivek. P.

Acknowledgements

This thesis is the result of many experiences I have encountered at the University of Calicut from dozens of remarkable individuals who I wish to acknowledge. My experience at University campus has been nothing short of amazing. First and foremost I wish to thank my research supervisor Dr.Lajish.V.L, Assistant Professor, Department of Computer Science, University of Calicut, Kerala. It has been an honor to be his first PhD student. I wish to express my deep gratitude towards him for the insightful guidance and his invaluable help in steering the course of study. And during the most difficult times throughout the course of study, he gave me moral support and the freedom I needed to move on.

I am deeply thankful to Mrs.Manjula K. A, Head, Department of Computer Science, University of Calicut, Kerala, for her encouragement and providing right resources and facilities in the Department for the fulfilment of this research work.

I extend my gratitude towards Dr. K Samudravijaya, Scientific Officer, Tata Institute of Fundamental Research, Mumbai, India: Dr.P.Laxminarayana, Professor and Director, Research and Training Unit for Navigational Electronics, Osmania University, Hyderabad, India and Dr.KumarRajamani, Architect, Robert Bosch Engineering and Business Solutions Ltd., Bangalore, India for their valuable support and inspiration throughout the work. I am deeply obliged to Dr.P.Nagabhushan, Professor and Director, Indian Institute of Information Technology (IIIT), Allahabad, India for his inestimable support.

I thank Mr. K. Suresh Babu, and Dr. P. Ramakrishnan, former faculty member of the Post Graduate Department of Physics, Govt. College, Madappally, Kerala who have extended their invaluable contributions. I also extend my gratitude to Dr. R.K. Sunilkumar, andDr.G.Harikrishnan,

faculty members of the same institute, who have supported me in all sphere of my work. I acknowledge Mr.K.Jayakumar IAS, former Vice-Chancellor, ThunchathEzhuthachan Malayalam University, Kerala and Dr.M.Sreenathan, Faculty Dean and Professor, Department of Linguistics, ThunchathEzhuthachan Malayalam University, Kerala for being the discussants especially on the study of Malayalam phonology.

My special thanks are due to my colleagues Mr.Sandesh E.PA, Mr.Baiju K.B, Mrs.Reshma P.K, Mr.Benson C.C, Ms.Habeebath K.P, Mr.Anoop K and Ms.Manjary P Gangan for their wholehearted co-operation throughout this endeavor. I would like to express my special indebtedness to my family and friends whose encouragement and support was unremitting source of inspiration for this work.

Vivek. P.

To my beloved family for allways standing by me

Contents

List of Tables

List of Figures

Abbreviations

ANN : Artificial Neural Network
ASR : Automatic Speech Recognition
AUC-PR : Area Under Curve Precision
AUC-ROC : Area Under Curve Receiver Operating Characteristics
CD : Capacity Dimension
CRD : Correlation Dimension
CV : Consonant-Vowel
DCT : Discrete Cosine Transformation
EM : Expectation Maximization
EMM : Exact Matching Method
FFL : Fixed Frame Length
FFMLP : Feed Forward Multilayer Perceptron
FFR : Fixed Frame Rate
FFT : Fast Furrier Transform
G2P : Grapheme to Phoneme
HMM : Hidden Markov Model
IPA : International Phonetic Alphabet
IRSTLM : IRST Language Modeling
KBPT-M : Knowledge Base Preparation Tool for Malayalam
KE : Kolmogorov Entropy
KNN : K-Nearest Neighbour
KWS : Key Word Spotting
LLE : Largest Lyapunov Exponent
LTI : Linear Time Invariant
MFCC : Mel Frequency Cepstral Coefficients
MIL : Multiple Instance Learning
ML : Maximum Likelihood
PLP : Perceptual Linear Prediction

QbE : Query by Example
RKHS : Reproducing Kernel Hilbert Space
RMM : Relaxed Matching Method
ROC : Receiver Operating Characteristics
RPS : Reconstructed Phase Space
RPS-EV : Eigen Value of Reconstructed Phase Space
SD : Standard Deviation
SDC : Spectral Decay Coefficien
STD : Spoken Term Detection
SV : Speaker Verification
SVM : Support Vector Machine

Chapter 1

Introduction

1.1 Background

Speech is the original medium in which human language evolved and is the most natural means of communication. Speech processing is a unique discipline, which encompasses a broad range and variety of technologies and applications that allow humans to interact naturally with intelligent computing systems. Digital Speech Processing (DSP) deals primarily with the processing of speech signal for Automatic Speech Recognition (ASR), Speech Synthesis (text-to-speech), Speech Coding and Speech Analytics. The art of automatic speech recognition has advanced remarkably in the past decade. Hidden Markov Models (HMMs) provide a simple and effective framework for modelling time-varying spectral vector sequences. As a consequence, almost all present day Large Vocabulary Continuous Speech Recognition (LVCSR) systems are based on HMMs. Acoustical (acoustic–phonetic) modelling, lexical modelling (pronunciation lexicon/vocabulary) and language modelling are the important parts of HMM based speech recognition algorithms. Language models play an important role in automatic speech recognition systems, particularly in modelling morphologically rich languages. Effective implementation of

ASR systems requires transcribed speech recordings from many speakers, pronunciation dictionaries which cover the full vocabulary of the training corpus, and massive amounts of text data to reliably train statistical language models.

With the rapid growth of information and communication technologies, broadcast, cellular and other wireless technologies, improved audio compression technology and the advent of low-cost large storage systems, a vast volume of speech data is being collected every day. A new emerging application in which ASR engines can be effectively used is speech analytics. In such applications, the ASR engine is used for Keyword Spotting (KWS) and provides the input for advanced surveillance and analytical application. Keyword Spotting is the technology of searching and detecting keywords of interest in audio streams using an ASR engine. The concept of speech analytics is to use various core technologies and develop targeted applications that use the KWS engine's capabilities to mine valuable information from the massive speech data. Speech-analytics solutions can provide an efficient way to access and leverage the valuable information derived from the speech data such as what is captured in telephone calls. Although the concept of KWS is not new, the ability to deliver such a functionality for large scale systems with acceptable levels of performance and accuracy has to be investigated further for improved results.

Malayalam is a Dravidian language spoken mainly in the south-western region of India. Like other Indian languages, it also possesses phonological features which are rich in vowel and consonant realizations. Malayalam is syllabic in nature and has a one-to-one correspondence between spoken and written syllables. Although speech technology has been the focus of research in India for a number of years and the technology itself has matured for real-world applications, the main obstacle in speech

research for Malayalam is the deficiency of standard benchmark speech, text corpora and associated computational linguistic resources. The present research and development efforts to shape innovative solutions in Malayalam ASR is still in a nascent stage and hence it is necessary that concerted efforts need to be initiated in this direction so that a complete Malayalam ASR-based speech analytics system will evolve in the public domain.

1.2 Motivation

The idea of interaction between computers and humans in natural language has reached the realm of reality. The active research in the field of speech processing in the last two decades is highly motivated by the increasing need for common people to interact naturally with the computing machines in regional languages. In addition to speech recognition and speech synthesis applications, speech analytics is also an evolving frontier research area. Keyword spotting (KWS) is a powerful technology tool used in speech analytics, which detects speech segments that contain a given query word in the large audio dataset. Keyword spotting is particularly well suited for applications such as real-time speech monitoring and large vocabulary audio indexing. Many prominent research findings have been reported in several foreign languages in the area of speech analytics. However, speech analytics research in Indian languages, particularly in Malayalam, is still in its infancy.

Malayalam has a rich set of phones and allophones. Durational and spectral properties of allophones have been studied in many languages, but Malayalam lacks comprehensive studies in this direction. The characterization of the allophonic variations in Malayalam phones can efficiently be used in speech processing applications. Since the durational properties

are language specific, detailed analysis has to be performed for each language, based on which suitable language models can be built to support speech recognition. This work contributes to both the characterization of allophonic variations and durational-spectral analysis of all the phones and allophones in Malayalam.

Language-specific pre-processing applications are essential for developing language computing tools. Grapheme-to-Phoneme (G2P) converter is one of the most needed foundation applications which is essential for language computing initiatives. Another objective of this study has been to develop a G2P transcriptor for Malayalam and to attempt for a detailed probabilistic analysis of Malayalam phonemes that occur in sentence and word corpora. The result of this analysis can be directly used to improve the efficiency of various speech processing applications.

The prime motivation of the study is to develop a modest speech analytic system based on a Malayalam KWS engine, which works with the help of Hidden MArkov Model based Automatic Speech Recognition, which can incorporate domain-specific linguistic knowledge and also a wide range of linguistic variation. In continuation of the above, another motivation for this study is formulate algorithms based on the Malayalam KWS engine which can attempt audio category prediction and speaker spotting to support speech analytics research.

1.3 Outline of the Thesis Organisation

The intent of chapter 2 is to establish the necessary background for the following chapters. A quick review of durational analysis, Grapheme-to-Phoneme transcription, audio classification and speaker spotting are discussed in this chapter. Different supervised and unsupervised KWS techniques were discussed in detail. A quick review on Indian language

speech recognition research is also included in the last part of this chapter.

Chapter 3 deals with the study of Malayalam phones and allophones. An allophone is a class of phones corresponding to a specific variant of a phoneme. A comprehensive ruleset for the formation of Malayalam allophones is discussed in detail. Then a detailed analysis is also performed in the allophonic variability of Malayalam vowels. An extensive statistical analysis is then performed on the durational properties and the first two formant frequencies of the vowel allophones. The characterisation of the allophonic variations in Malayalam vowel phone derived out of this research work can be efficiently used in automatic speech processing applications. This work is first of its kind in Malayalam language.

Chapter 4 proposes a rule-based Grapheme-to-Phoneme (G2P) transcription algorithm for Malayalam. The transcriptor converts input text into a sequence of phonemes with corresponding International Phonetic Alphabet (IPA) symbols. The set of Malayalam graphemes is divided into six subsets consisting of vowels, diphthongs,/*anusaram*/ and /*chandrakkala*/, consonant classes, compound letters and /*chillukal*/. A set of rules is defined based on the language-specific knowledge that helps to perform G2P transcription. The G2P transcription for each subset of graphemes is implemented with the help of separate processing routines. The system can be effectively used in the automatic speech processing applications including text-to-speech converter and speech recognizer. Phoneme frequency based on word corpus and own developed sentence corpus are evaluated using the proposed G2P conversion tool. The phoneme statistics derived out of the experiment can be a salient factor in designing various language processing tools. Then phoneme statistics is widely used to improve the performance of language model based speech recognition systems. The phonotactic observations about the

permissible phoneme combinations in a language can also be derived from the phoneme level statistical information.

In this work, keyword spotting experiments are conducted on a large news audio corpus. Chapter 5 deals with data preparation and Keyword Spotting experiments. The Malayalam news audio database is created by recording speech samples in a normal acoustic environment. News sentences were collected from popular online news portals of leading Malayalam dailies for dataset creation. For the processing efficiency, news audios were categorised into five classes *viz.* State news, National news, International news, Sports news, and news with Cultural importance. Malayalam continuous speech database is created by recording the news sentences spoken by 35 male and female speakers of different age groups. Each speaker uttered 150 sentences taken from the above mentioned five news categories. All these, 5250 spoken sentences are generated and these speech samples are labelled with a category id indicating the news category, sample number, speaker id and speaker gender that they belong to. The news text dataset is transcripted using the proposed rule-based G2P transcription algorithm. A discriminative method for detecting and spotting keywords in spoken utterances is described subsequently. The most widely used speech recognition algorithm, Hidden Markov Models (HMM) is used for its implementation. The block diagram of the proposed system is shown in figure 1.1. In the speech analytics module, a novel method for audio classification based on Multiple Instance Learning (MIL) algorithm is proposed. A nonlinear speaker modelling approach is also proposed to facilitate the speech analytics research and the performance of the proposed speaker spotting system is verified experimently.

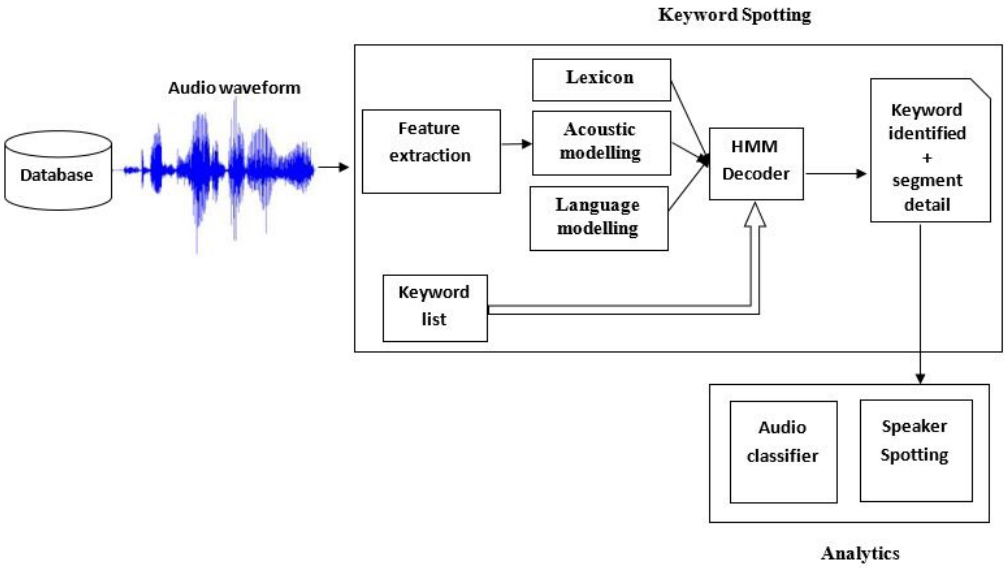


Fig. 1.1 Block diagram of the proposed system

Preparation of knowledge base is an important pre-processing step performed as part of the HMM-based speech modelling and recognition. A Knowledge Base Preparation Tool for Malayalam (KBPT-M) is proposed a part of this work. This tool generates language model and related components required for acoustic modelling of Malayalam language. The purpose of KBPT is to simplify the processing of knowledge base generation which is essential for HMM-based speech modelling. HMM decoder is trained using the news audio data set and knowledge base generated using KBPT. The recogniser first converts the audio waveform into a sequence of fixed size acoustic feature vectors.

In this work, the Keyword Spotting (KWS) system is implemented based on two different approaches. The first one is Automatic Speech Recognition based approach (ASR-KWS) and the second one is the filler model based acoustic approach. The KWS system evaluation is performed using two different methods namely Exact Matching Method (EMM) and Relaxed Matching Method (RMM). The exact matching

method uses exact keyword targeting criteria, whereas in the relaxed matching method the inflected (FMA-KWS) forms are allowed in the keyword targeting criteria. The results obtained from both methods are compared and discussed. The evaluation dataset consisting of 5,250 Malayalam news audio samples. There are approximately 570 mentions of keywords in the dataset. The average length of audio samples is 5.35 seconds, and the average number of characters in the keyword set is 5.6875. Keyword is considered to be recognised correctly if it is present in the transcription of the selected audio file fragment. The proposed KWS system is implemented using two different methods. The first one is ASR based Keyword Spotting technique (ASR-KWS) and the second one is the Filler Model based Acoustic approach (FMA-KWS). A word lattice consists of a set of nodes representing points in time and a set of spanning arcs representing word hypotheses is used for optimizing the KWS search. The experiment results obtained are analysed based on the performance scores *viz.* precision, recall, and F1. It is observed that the relaxed matching method gives better performance than exact matching method. It is also observed that ASR based KWS, with the optimized Lattice Search (ASRLS-KWS) provides improved KWS performance

In chapter 6, a novel method for content-based news audio classification using Multiple Instance Learning (MIL) approach is proposed. The content-based analysis provides useful information for audio classification as well as segmentation. Audio content understanding is an active research problem in speech analytics. A key step taken in this direction is to propose a classifier that can predict the category of the input audio sample. Classification of the audio samples are conducted based on the keywords identified from the input audio sample using the proposed KWS system. In MIL, the labels are assigned to bags of instances. The binary classifier labels an audio bag positive if no less than one instance

in that bag is positive. Otherwise the bag is labelled as negative. The classification results obtained using the MIL approches are evaluated using different performance metrics. Two types of features are used for audio content detection, namely Mel-Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP). Two variants of MIL based techniques *viz.* mi-Graph and mi-SVM are used for classification purpose. From the experimental results, it is evident that the MIL based approach works efficiently for the audio classification. It is also verified that mi-Graph with MFCC feature provides a better F1 score of 0.91 compared to other methods.

Chapter 7 describes the speaker spotting method proposed in this study with the emphasis on nonlinear speaker modelling techniques. These research findings can be effectively used in speech analytics approaches when speaker specific short listing of keyword spotted audios are required. In general, the speech signal is considered as non-stationary in nature as it is produced from a time-varying vocal tract system with time-varying excitation. However, most of the signal processing algorithms used in speech encoding like LPC, MFCC *etc.* considers and models speech as a Linear Time-Invariant (LTI) system. In this work, an attempt is made to model the vocal tract using different nonlinear features. Nonlinearities are included in attempts to model the physical process of vocal cord vibration, which has focused on more than one model. The traditional nonlinear features like capacity dimension, correlation dimension, Kolmogorov entropy and largest Lyapunov exponents of audio data are extracted and analysed. Two novel nonlinear features, Eigenvalues of the Reconstructed Phase Space (RPS-EV) and Spectral Decay Coefficients (SDC) which is extracted from the power spectrum of the speech samples are proposed in this work. The speaker spotting capabilities of the proposed nonlinear features are investigated using

Feed Forward Multilayer Perceptron (FFMLP) classifier simulated using the error backpropagation learning algorithm. The experimental results indicate that the speaker spotting results obtained using combined non-linear features can be efficiently used to improve the speaker spotting results obtained based on the conventional acoustic parameters including MFCC, LPC *etc.* The speaker spotting results as part of the study can be used for speaker specific shortlisting of the KWS result.

Finally, chapter 8 concludes this work and suggests a few directions for future research. The complete outline of the chapter organization and research work reported in this thesis is shown in figure 1.2 which consists of 8 chapters followed by a brief description of the contents on each chapter.

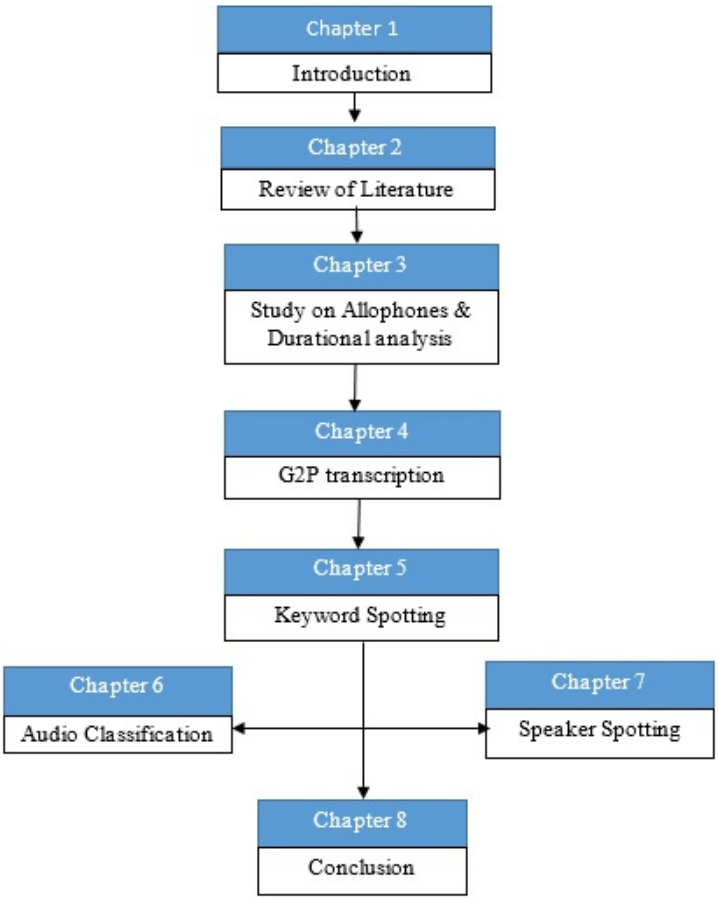


Fig. 1.2 Block diagram of the thesis structure

Chapter 2

Review of Previous Works

2.1 Introduction

Automatic speech recognition was regarded as an important and challenging research area by scientists and engineers even before the advent of modern electronic gadgets. The initial research works on speech recognition reported on early 1920s. A toy named Radio Rex was the first machine that manufactured to recognize speech [1]. Bell labs introduced a speech synthesis machine at the New York World Fair in 1939. Davis, K. H *et al.* from Bell Laboratory demonstrated a speaker dependent system for isolated speech recognition in 1952 [2]. In 1956, at RSA laboratory, Olson and Belar recognized 10 distinct syllables of a single speaker [3]. Rapid growth in speech recognition research was reported from 1960 onwards. Different representational techniques were introduced during this time like spectral analysis, statistical methods, wavelet based techniques *etc.* Among them, the statistical framework, Hidden Markov Model is the most widely accepted method. Recently, different categories of speech recognition systems have come into existence. Nowadays network of Deep Neural Networks are widely used for commercial speech processing products [4, 5]. Speech recognition and

speech synthesis in Malayalam is an active research area [6–8]. But the attempts have not yet reached the realm of successful development of Malayalam speech technology solutions that are useful for the masses. More study is needed on the basic structure and peculiarities of the language in the perspective of Malayalam language computing. Malayalam is an alphasyllabary language with the aksharam (character) as its core. Almost one to one correspondence exists between orthographic symbols in the alphabet and phoneme in Malayalam [9].

Speech assisted human-machine interaction helps easy communication in native language, particularly in a multi-lingual country like India, where a large majority of the people will not be comfortable by means of communicating in English. Extensive research works for developing systems that enable human-machine interaction in Indian languages for Hindi [10–13], Bangla [14], Telugu [15], Tamil [16], Kannada [17] have been reported. A promising contribution in this direction by Samudravijaya *et al.* [18] presents a speaker independent, continuous speech recognition system for Hindi. The proposed system recognizes spoken queries in Hindi in the context of a railway reservation inquiry task and recognizes sentences spoken naturally (without the need for a pause between words).

The recognition of Malayalam speech has been studied by many researchers [6–8, 19]. Lajish V.L *et al.* [20] proposed a recognition scheme for unconstraint, large vocabulary, Malayalam words using lexicon based heuristics and HMMs. The hierarchical probabilities of occurrence of fifty-one letters in the Malayalam alphabets after the starting letter of the selected word were estimated and compared for the implementation of the proposed recognition scheme. The Malayalam lexicon *Sabdhatharavali* [21] is used in this study and considered 88,413 words. It may be noted that the work on Keyword spotting as well as speech analytics

in Malayalam is still in its infancy. The following section 2.2 presents a summary of the research works reported in speech processing based on allophonic variations and phone duration modelling. Section 2.3 gives a review of methods and main investigations towards Grapheme-to-Phoneme (G2P) transcription. Section 2.4 describes various keyword spotting strategies adopted in speech processing. Section 2.5 and 2.6 presents a special review on audio classification and speaker spotting techniques which are widely used in speech analytics. The section 2.5 discusses different content based audio classification techniques and the section 2.6 emphasis on nonlinear speaker spotting methods.

2.2 Review on Speech Processing based on Allophonic Variations and Phone Duration Modeling

Allophone based speech processing studies are reported in many languages. Piotr Kozierski *et al.* used allophones instead of phonemes for polish language speech recognition [22]. Here, instead of using the entire set of allophones, a proper subset of allophones is used based on the frequency of occurrence. Some rarely occurring allophones are omitted in their study. A variation of the same is also used for the Japanese language by Long Nguyen *et al.* [23]. Allophonic properties are explored in Korean speech recognition in a work reported by Xu, Ji *et al.* [24]. Much concatenative text to speech synthesis systems use allophones as the basic unit. Imedjdouben *et al.* introduced a phone to allophone conversion algorithm for speech synthesis dedicated to the Arabic language [25]. A report by Pavel A. Skrelin describes the principles of the allophone extraction from Russian natural speech flow, ways of forming synthesized speech, modification of the acoustic

parameters [26]. Barkhoda *et al.* implemented three synthesis systems for the Kurdish language based on syllable, allophone, and diphone and showed all systems' Intelligibility is acceptable using various tests [27].

As indicated by Louis C.W. Pols *et al.*, the recognition performance can be further improved by incorporating "specific knowledge" (such as duration and pitch) into the recognizer [28]. They conducted a detailed study on phone duration modeling and its potential benefit on ASR. Durational models are developed in many languages to address the dynamic nature of phoneme duration. Rule-based approaches built on experimental data and model-based approaches are reported in various languages such as English, Korean, Turkish and Czech [29–32]. Durational modelling and characteristic analysis were performed in various Indian languages. Samudravijaya, K. studied the durational characteristics of Hindi phonemes as well as stop consonants in detail [33, 34]. K. Sreenivasa Rao and B. Yegnanarayana proposed a syllable duration prediction system for Indian languages [35]. The analysis is performed in Hindi, Telugu and Tamil languages, in order to predict the duration of syllables in these languages using SVM regression model. N. Sridhar Krishna *et al.* reported a preliminary attempt on data-driven modeling of segmental (phoneme) duration for two Indian languages, Hindi and Telugu [36]. S. R. Savithri identified some of the variables influencing the durations of Kannada vowels in the initial position [37]. Deepa P. Gopinath *et al.* proposed a preliminary Malayalam phoneme duration model for speech synthesis system [38]. As Malayalam is a language with a rich set of allophones, the study of allophonic variations and duration analysis is considered to be more relevant for alpha syllabic languages like Malayalam.

2.3 Review on Grapheme-to-Phoneme transcription methods

Phonetization of text and associated problems in various languages is an active research area in natural language processing (NLP). Phonetization of text has important applications in both speech synthesis as well as in speech recognition. In speech synthesis, the transcript data is used to derive the correspondence between the orthography and the sounds. In speech recognition, it is used for modeling the speech to enhance the quality of the recognizer and generating pronunciations for new words, which are not in the original vocabulary of the speech recognition system. There are mainly three methods that have been used for grapheme-to-phoneme transcription. They are (1) dictionary-based methods in which maximum phonological information about morphemes in a lexicon are stored in a dictionary (2) Rule-based methods whereby linguistic and phonetic knowledge is used to develop a set of letter-to-sound rules and (3) the relatively newer trained data-driven methods.

Considering the alpha syllabic nature and the regular relationship between its spelling and pronunciation of Malayalam language, a comprehensive set of grapheme-to-phoneme conversion rules are proposed in this work. The initial works reported in this area were for English and French. Ainsworth in 1973 introduced a system for converting English text into speech [39]. He segmented the text into breath groups, an utterance or part of an utterance produced between pauses for breath, and the orthography is converted into a phonemic representation. For speech synthesis, he assigned the lexical stress to appropriate syllables. Michel Divay *et al.* developed a Grapheme-to-Phoneme transcription system for French in 1977 [40]. They introduced a rule system based on the application of a partially ordered set of phonological rules. The left-hand side of each

rule indicates the graphemes involved by the rule and the right-hand side of each rule specifies the corresponding phonemes. In the following years, different rule-based transcription algorithms were proposed in English, French *etc.* [41–45]. Askars Salimbajevs *et al.* implemented a large vocabulary automatic speech recognition for Latvian language using a rule-based grapheme to phoneme converter [46]. They archived the first-rate result with the grapheme-based approach extended with several straightforward rules.

A Grapheme-to-phoneme transcription in the Hungarian language is proposed by Attila Novak *et al.* [47]. Besides the implementation of transcription rules, the proposed tool incorporates the knowledge of a Hungarian morphological analyzer so as to detect morpheme and compound boundaries. Daniela Braga *et al.* introduced a rule-based grapheme-to-phone converter for TTS systems in European Portuguese [48]. A rule-based Korean Grapheme to Phoneme Conversion is proposed by Yu-Chun Wang *et al.* at 2009 [49].

A Rule-Based Grapheme to Phoneme Mapping for Hindi Speech Synthesis is proposed by Monojit Choudhury [50]. A computational framework for rule-based grapheme to phoneme mapping for Hindi has been described. Sumi S Nair *et al.* made an attempt to frame Rule-Based Grapheme to Phoneme Converter for Malayalam [51]. In this work, all exception rules in the algorithm are not discussed and the performance of the transcriptor is also not evaluated.

2.4 Review on Keyword Spotting Techniques

Keyword spotting finds the occurrences of specific words in a speech utterance. Content-based indexing and retrieval of speech data can be made powerful with the approach of keyword spotting (KWS). Earlier

works in keyword spotting can be found in the late 1980s. In 1989 J.R. Rohlicek *et al.* presented a Gaussian hidden Markov model based KWS system [52]. The results are reported on the use of various signal processing and feature transformation techniques. The authors have observed that performance can be greatly affected by the choice of features used, the covariance structure of the Gaussian models, and transformations based on energy and feature distributions.

A work by R. Wohlford *et al.* on enhancement of speaker independent word spotting in conversational, telephone bandwidth speech from a variety of talkers. This work involves the comparison of five LPC based parameter sets with three levels of additive white noise using a dynamic programming based techniques [53]. In 1990, a speaker-independent hidden Markov model (HMM) KWS for continuous speech recognition is proposed by R. C. Rose and D. B. Paul [54]. The system provided 82% probability of detection for KWS for a 20-keyword vocabulary standard conversational speech. 1n 2009 Joseph Keshet *et al.* proposed a new approach for keyword spotting, which is based on large margin and kernel methods [55]. The proposed method employs a discriminative learning procedure, in which the learning phase aims at achieving a high area under the ROC curve.

In 2017, Ganapathiraju *et al.* proposed a real-time KWS system for speech analytics is proposed by feeding real-time audio along with a keyword model, into a recognition engine [56]. The audio stream data probability is computed by recognition engine and compared to a threshold to determine whether or not the keyword has been spotted. In the same year Chen, Zhehuai, *et al.* proposed a phone synchronous decoding method by removing tremendous search redundancy caused by blank frames resulting in compact phone-level acoustic space representation: Connectionist temporal classification (CTC) lattice [57]. KWS with

CTC lattice achieved reduced KWS model size and increased the search speed.

The major challenges in this area are detection performance, detection of out of vocabulary words, speed of search, handling of variants of a keyword, *etc.* The methods in keyword spotting can be broadly classified into supervised techniques and unsupervised techniques which are shown in figure 2.1.

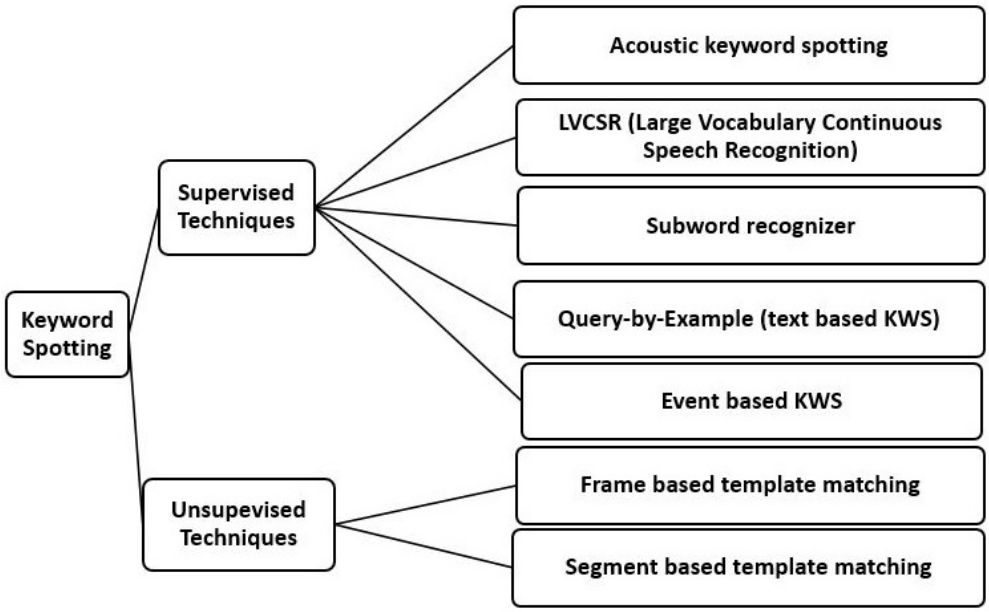


Fig. 2.1 Classification of keyword spotting techniques

There are mainly 4 types of keyword spotting techniques in supervised method. The following sections review the most widely used Acoustic and Large Vocabulary Continuous Speech Recognition (LVCSR) based Keyword Spotting techniques in detail. The subword recognizer builds indices with different subword units such as phone n-grams, multi-grams, syllables, segments or lattice representations for subword units [58]. Query-By-Example (QBE) uses phone lattice representation of keyword examples to be matched against a similar representation of the target utterance [59]. Text-based STD techniques are applied on phone lat-

tices during the process of matching. Event based keyword spotting is motivated by the fact that a keyword can be characterized by a set of phonetic events and a faster processing can be achieved by minimizing the set of phonetic events used to represent a keyword [60]. Unsupervised techniques are template matching methods [59]. Such methods are based on template matching paradigm where the queried keyword template is matched with the target utterance for detecting a possible presence of the same. These approaches do not require the availability of any kind of labelled resources. Hence they are most suitable for under-represented languages.

2.4.1 Acoustic Keyword Spotting Methods

Earlier works in keyword spotting started with the acoustics based keyword spotting method. The process of acoustic keyword spotting is depicted in Figure 2.2. Acoustic KWS is performed in only one stage [61]. Recognition vocabulary is constructed with the set of designated keywords. Non-keyword speech is modelled using the general speech models.

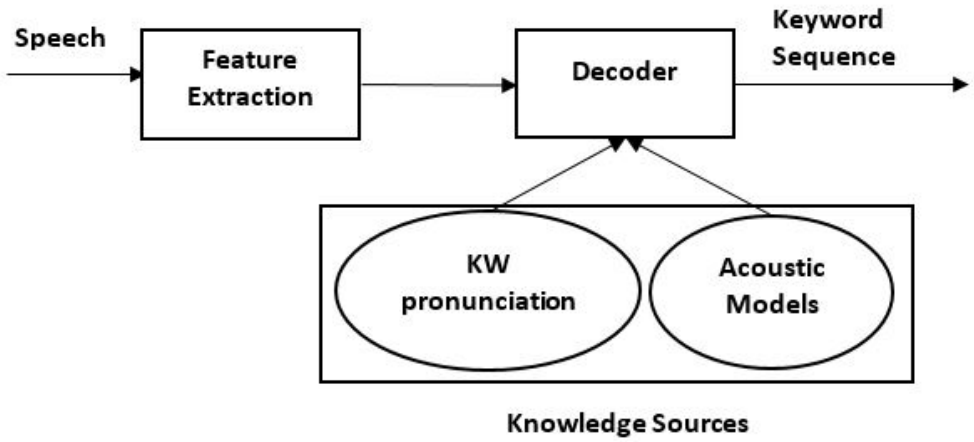


Fig. 2.2 Acoustic keyword spotting approach

In 1990 R.C. Rose and Douglas B. Paul presented a Hidden Markov Model (HMM) based keyword recognizer (KWR) using continuous-speech-recognition model [54]. The acoustic models are trained for providing the representation of non-vocabulary speech.

A comparison of acoustic keyword spotting against large vocabulary continuous speech recognition based keyword spotting and a hybrid approach making use of phoneme lattices generated by a phoneme recognizer is conducted by Igor Szoke *et al.* in 2005 [62]. A two-stage process for utterance verification is presented by Rafid Sukkar A. *et al.* [63]. The HMM-based speech recognizer produces recognition hypothesis. The training is attained by a procedure to minimize the verification error.

Eric D. Sandness and I. Lee Hetherington in 2000 proposed a training technique for training only segments of speech relevant to keyword errors [64]. In this work it is observed that the keyword error rate reduces highly and the overall accuracy is improved for keyword-based training. An out-of-vocabulary word detection method using confidence measures and support vector machines is proposed by Y. Benayed *et al.* in 2003 [65]. The phone level information provides the confidence measures by a Hidden Markov Model (HMM) based speech recognizer.

An approach for keyword spotting based on mapping the acoustic representation of the speech into a vector space is proposed by J. Keshet *et al.* [55]. The method describes an iterative algorithm for training a keyword spotter. The disadvantage of the acoustic keyword spotting is in handling new keywords. The target utterance is to be decoded freshly with the new keyword list. This results in high search time. This limitation is addressed by LVCSR systems [66].

2.4.2 LVCSR based Keyword Spotting Methods

LVCSR system is used to generate word level transcription corresponding to the input speech signal. Then the transcription generated is indexed using information retrieval techniques available for text. These indices are examined for the presence of the keyword to be searched. LVCSR can be found in many of the recent keyword spotting systems. LVCSR utilizes a large amount of annotated speech for supervised training of HMMs. LVCSR engine first transforms the speech signal into text. It searches the possible sequence of words using acoustic models based on the Viterbi search algorithm. Next, it utilizes the search methods to locate the keywords within the text. Figure 2.3 shows the block diagram of a typical LVCSR keyword spotting system.

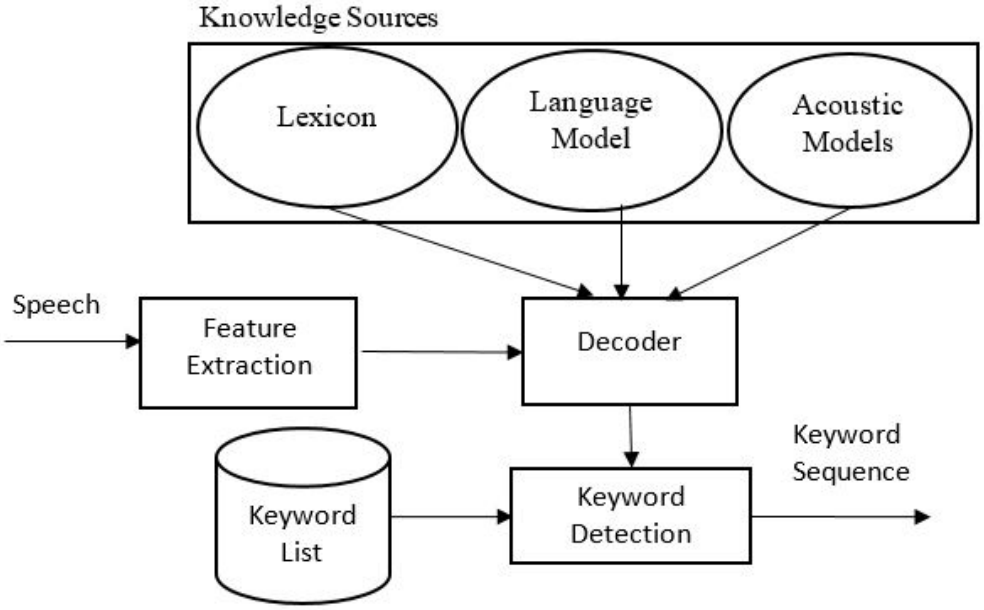


Fig. 2.3 An LVSCR keyword spotting system

R. Rose and D. Paul in 1990 proposed a Hidden Markov Model (HMM) Based keyword Recognition System [54]. A speaker-independent HMM keyword recognizer (KWR) based on a continuous-speech-recognition

model is implemented. John S. Garofolo *et al.* proposed a Spoken Document Retrieval (SDR) technology within a broadcast news domain [67]. SDR involves the search and retrieval of excerpts from spoken audio recordings using a combination of automatic speech recognition and information retrieval technologies.

A framework for improving the accuracy of speech recognition is proposed in 2000 by Mangu, L. *et al.* [68]. The method minimizes word error rate by extracting the highest posterior probabilities of word lattices. The method proposes a representation with the sequence of word-level confusions in a compact lattice format So as to improve the accuracy. A much more compact word lattice representation known as Position Specific Posterior Lattice (PSPL) is proposed by Chelba, C. and Acero. A in 2005 [69]. The posterior word probability at a specific position in a lattice is given by PSPL by the summation of all weights of the path.

A comparison of Position Specific Posterior Lattice (PSPL) and word confusion Network (WCN) is discussed by Pan, Y. C., & shan Lee, L. in 2010 [70]. Even though more space is required for index storage, It is found that PSPLs outperform WCNs. Modeling of multilingual words is one of the limitations, especially for Indian languages. LVCSRs cannot naturally handle Out Of Vocabulary (OOV) words. It is due to the effect of language model an LVCSR based approach gives high word level accuracy. Thus LVCSR is not much effective for keyword spotting approaches whose appropriate language model is not available for training [66].

2.5 Review on audio classification using Multiple Instance Learning

Automatic audio analytics and classification is an emerging research area in multimedia stream. Erling Wold *et al.* in 1996 introduced an engine for content-based classification, search, and retrieval of audio data [71]. This analysis allows the sounds to be classified or queried by their audio content. Queries can be based on any one or a combination of the acoustical features, either by specifying previously learned classes based on these features, or by selecting or entering reference sounds and asking the engine to retrieve sounds that are similar or dissimilar to them. A system to retrieve audio documents by acoustic similarity is introduced by Jonathan T. Foote in 1997 [72]. The similarity measure used is based on statistics derived from a supervised vector quantizer.

In the early days, classification experiments were conducted on simple cases such as speech-music classification, speech-silent classification *etc.* Pfeiffer *et al.* presented a theoretical framework and application of automatic audio content analysis using some perceptual features [73]. In a work, D. Kimber *et al.* classified audio recordings into speech, silence, laughter, and non-speech sounds, to segment discussion recordings in meetings [74]. Zhang and Kuo introduced a method to discriminate audio recordings into classes such as songs and speeches over music, based on a heuristic-based model [75].

Audio-based approaches are also found more in video classification literature rather than the texts and video based approach. Advantages of audio approaches includes the usage of fewer computational resources than visual methods and more dependable than text. In the earlier works time domain features were used widely and later, researches started using combined features from both time and frequency domains for better

recognition accuracy [76, 77]. Among these features MFCC is identified as the most used and trustable one [78].

Liu *et al.* in 1998 considered the problem of discriminating five types of news namely commercial, basketball game, football game, news report, and weather forecast [79]. They have designed an ergodic HMM using the clip based features as observation vectors. A filter predictor for audio event classification and extraction is introduced in 2017 by Visser *et al.* [80]. Deep Neural Networks (DNNs) is used to extract underlying target audio events from unlabelled training data.

The MIL problem was first formulated for the task of digit recognition, Here a neural network was trained with the information on the presence of a given digit without specifying its position [81]. Another early application of MIL was to the problem of drug discovery in which the bags were molecules of the drug and the instances were conformations of those molecules [82]. MIL has also been applied to object detection in images [83], video classification to match names and faces [84], and to text classification [85], in which the bags were documents and the instances were sentences or paragraphs. Many approaches have been introduced for MIL, including mi-Graph, Gaussian Process Multiple Instance Learning (GPMIL), MILBoost, mi-SVM and Bag key instance SVM (B-KI-SVM) [86]. The use of weakly supervised machine learning technique can reduce the computational cost in a large manner. So far no work has been reported on the use of Multiple Instance Learning (MIL) approach on speech audio classification.

2.6 Review on Speaker Spotting methods with emphasis on nonlinear properties of the speech signal

It was Lawrence Kersta who made the first major step towards speaker identification by computers as he developed spectrographic voice identification at Bell Labs in the early 1960s [87]. His identification procedure was based on the visual comparison of spectrogram, which was generated by a complicated electro-mechanical device. Although the visual comparison method cannot cope with the physical and linguistic variation in speech, this work encouraged the introduction of automatic speaker recognition. The global shape of the Discrete Fourier Transform (DFT) magnitude spectrum which is known as spectral envelope contains information about the resonance properties of the vocal tract and has been found out to be the most informative part of the spectrum in speaker recognition [88].

In the subsequent four decades, speaker recognition research has advanced a lot. Speech production has been assumed to be linear in the past, and the parameters such as the Mel Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Cepstral Coefficients (LPCCs) are used in speaker spotting system. Nowadays this linear model has been challenged, and the importance of the nonlinearities emphasised [78]. Nonlinearity is routinely included in attempts to model the physical process of vocal cord vibration, which have focused on two or more mass models [89].Observations of glottal waveform reinforce this evidence, where it has been shown that this wave form can change shape at different amplitudes. Such change would not be possible in a strictly linear system where the wave form shape is unaffected by the amplitude changes. Nonlinear signal processing techniques have several poten-

tial advantages over traditional linear signal processing methodologies [90, 91]. Different nonlinear speaker spotting experiments are conducted based on phase space reconstruction in 2003, dynamic modeling in 2002 and chaotic modelling in 2002 [92–94].

AdrianoPetry *et al.* in 2002, combined commonly used feature parameters, such as LP-derived cepstral coefficients and pitch, with nonlinear dynamic features, namely, fractal dimension, entropy and largest Lyapunov exponent [93]. They observed that the best results are obtained when the nonlinear dynamic features are included and suggest that the information added with these features was not present in the others so far. Niko Brummer *et al.* conducted a comparison of linear and non-linear calibrations for speaker recognition [95]. It is observed that the non-linear methods provide wider ranges of optimal accuracy and it can be trained without having resort to objective function tailoring. In 2017, Shabnam Gholamdokht Firooz *et al.* evaluated some useful features from the Recurrence Plot (RP) of the embedded speech signals in the RPS; the proposed features are evaluated via applying a two-dimensional wavelet transform to the resulted RP diagrams [96].

2.7 A special review on Indian language speech recognition

The diversity in languages spoken in India truly reflects the diversity in Indian life and culture. The structure and hence the language modelling for Indian languages are different from English and other foriegn languages and even carries significant differences within themselves.

A significant work on database development and speech recognition in Indian language is reported at Tata Institute of Fundamental Research

(TIFR), Mumbai [97]. A Speech Recognition system for agriculture purpose using cell phones and landline in Marathi Language is developed based on the dataset. The speech data for the project was collected using two dedicated phone line. For the development of database two volunteers have been appointed by the TIFR and IIT Bombay. They visited various districts of Maharashtra and Collected the Speech Sample by calling the dedicated phone lines. This database consists of speech samples recorded from approximately 1500 speakers. As the phone recorded data is narrow band speech along with background noise, wide band speech are also recorded when the speaker speaks on the phone line.

There are various standard speech databases available for foreign Languages but very fewer for Indian Languages. Table 2.1 lists various standard benchmark Speech Databases developed in different Indian Languages to support speech processing research.

Table 2.1 List of standard benchmark speech databases avaialable in various Indian languages

Sl.No.	Language	Database Devoloped by	Application of Database
1	Marathi	TIFR Mumbai and IIT Bombay	Speech Recognition System for Agriculture Purpose
2	Hindi	C-DAC Noida	Speech Recognition System for Travel Domain

Continue on the next page

Table 2.1 List of standard benchmark speech databases avaialable in various Indian languages(cont.).

Sl.No.	Language	Database Devoloped by	Application of Database
3	Telugu	IIIT Hyderabad	Speech Recognition System for Agricultural commodity Price Enquiry
4	Telugu, Hindi & English	IIIT Hyderabad	Travel and Emergency Services
5	Garhwali	Government P.G. College, Rishikesh	Speech Recognition System
6	Punjabi	Punjabi University, Patiala	Text to Speech Synthesis System
7	Hindi, Odiya, Bengali & Telugu	Utkal University, Bhubaneswar	Text to Speech Synthesis System
8	Hindi	TIFR Mumbai and C-DAC Noida	General Purpose
9	Hindi, Telugu, Tamil, & Kannada	IIT Kharagpur	General Purpose
10	Konkani (Goan)	Islampur, Maharashtra	Text to Speech Synthesis System
11	Kannada	SJ College of Engineering, Mysore	Text to Speech Synthesis System

Continue on the next page

Table 2.1 List of standard benchmark speech databases avaialable in various Indian languages(cont.).

Sl.No.	Language	Database Devoloped by	Application of Database
12	Hindi & Indian Spoken English	KIIT, Bhubaneswar	Mobile based speech recognition
13	Marathi, Tamil & Telugu	IIIT Hyderabad and HP Labs Bangalore	General Purpose
14	Malayalam	TEMU, Kerala	General Purpose

Table 2.2 summarises the automatic speech recognition works reported in Indian Languages in this millennium. Most of the recent ASR works are developed on Hidden Markov Modelling based implementations. The MFCC, LPCC, PLP *etc.* are the commonly used parameters for speech modelling in conventional ASR systems. Table 2.2 summarizes the major recent speech recognition research outcomes reported for Indian languages. The table includes quick references of 11 works from Hindi and from other Indian languages including Punjabi (3), Tamil (10), Assamese(2), Bengali (4), Marati (8), Urdu (2), Oriya (2), Kannada (5), Telugu (7), Gujarati (4), Bodo (2) and Malayalam (11).

Table 2.2 Summary of recent ASR research in Indian language

Sl. No.	Language	Work	Methods Used	Year	Authors
1	Hindi	continuous speech recognition	MFCC feature based dynamic time wrapping	2000	Samudravijaya [18]
2	Hindi	continuous speech recognition	IBM speech recognizer	2002	Chalapathy Neti *et al.* [98]
3	Hindi	Digit recognition	HMM	2006	Gupta *et al.* [99]
4	Hindi	Digit recognition	LPC coefficientsBased HMM	2006	Venkitaramani [100]
5	Hindi	Large Vocabulary Continuous Speech Recognition	HMM trained using forward backward Algorithm and Trigram Language Model	2004	N.Rajput M *et al.* [13]
6	Hindi	Continuous Speech Recogniser	Julius Speech Recognition Engine	2010	Mathur *et al.* [101]
7	Hindi	Travel Domain AR	Julius Speech Recognition Engine	2010	Sunitha *et al.* [102]

Continue on the next page

Table 2.2 Summary of recent ASR research in Indian language(cont.)

Sl. No.	Language	Work	Methods Used	Year	Authors
8	Hindi	Isolated Word Speech Recogniser	MFCC feature based HMM	2011	Kumar and Agarwal [103]
9	Hindi	Statistical Analysis of Classification	Extended MFCC based HMM	2011	Agarwal and Dave [104]
10	Hindi	Connected digit recognizer	HMM	2011	Mishra *et al.* [105]
11	Hindi	Speech Recognition with speaker Adapatation	HMM for recognition and Maximum Likely Hood Linera Regression for Speaker Adaptation	2011	Sivaram *et al.* [106]
12	Punjabi	Isolated and connected word recognizer	Acoustic Template Matching	2004	Kumar *et al.* [107]

Continue on the next page

Table 2.2 Summary of recent ASR research in Indian language(cont.)

Sl. No.	Language	Work	Methods Used	Year	Authors
13	Punjabi	Speech Recognition	Group of techniques including ANN and Vector Quantisation	2012	Ghai W *et al.* [108]
14	Punjabi	Isolated Word Recognition System	HMM and DWT	2010	Kumar *et al.* [109]
15	Tamil	Domain Specific Speech Recognition	HMM	2001	Nayeenmulla *et al.* [110]
16	Tamil	Language model for speech recognizer	Morpheme based and other Models comparison	2004	Saraswathy *et al.* [111]
17	Tamil	Continuous speech recognizer	Automatically annotated recognizer and HMM	2006	Lakshmi A *et al.* [112]

Continue on the next page

Table 2.2 Summary of recent ASR research in Indian language(cont.)

Sl. No.	Language	Work	Methods Used	Year	Authors
18	Tamil	Continuous Speech Recognition - Word and triphone based	HMM	2008	R. Thangarajan *et al.* [113]
19	Tamil	Continuous Speech Recognition	MFCC and HMM	2009	R. Thangarajan *et al.* [114]
20	Tamil	Recognition for Words and Numerals	MFCC and DTW	2012	V.S.Dharun *et al.* [115]
21	Tamil	Continuous Speech Recognition, monophone based	HMM	2012	V. Radha *et al.* [116]
22	Tamil	Continuous Speech Recognition	HMM	2012	Vimala C. *et al.* [116]
23	Tamil	Digits Recognition	Vector Quantisation	2012	S. Karpagavalli *et al.* [117]

Continue on the next page

Table 2.2 Summary of recent ASR research in Indian language(cont.)

Sl. No.	Language	Work	Methods Used	Year	Authors
24	Tamil	Word Based Speech Recognition	HMM	2013	A. Akila *et al.* [118]
25	Assameese	Digit Recognition	LPC and HeterogeneousANNs	2010	M. P. Sarma *et al.* [119]
26	Assameese	Digit Recognition	Cooperative LVQ	2011	M. P. Sarma *et al.* [120]
27	Bangali	Phoneme Recognition	Neural Network based Aproach	2003	M. R. Hassan [121]
28	Bangali	Isolated Word Recognition	HMM	2006	Mahmudul Hoque *et al.* [122]
29	Bangali	Triphone and monphone based language models	Decision tree based clustering	2008	Banerjee *et al.* [123]
30	Bangali	E-mail Application for blind	HMM based	2010	Mandal *et al.* [124]

Continue on the next page

Table 2.2 Summary of recent ASR research in Indian language(cont.)

Sl. No.	Language	Work	Methods Used	Year	Authors
31	Marati	Digit Recognition	Linear Predictive Coding (LPC) coefficients decomposed using Discrete Wavelet Transform (DWT) With Continuous Density HMM	2009	N. S. Nehe *et al.* [125]
32	Marati	Isolated Word Recogniser	Multi HMM	2012	Kayte *et al.* [126]
33	Marati	acoustic features for aspiration detection in voiced and unvoiced stops of Marathi	Feature Detection	2011	Vaishali Patil *et al.* [127]

Continue on the next page

Table 2.2 Summary of recent ASR research in Indian language(cont.)

Sl. No.	Language	Work	Methods Used	Year	Authors
34	Marati	The automatic detection of aspiration in Marathi word-initial voiced stops and affricates	Feature Detection	2013	Vaishali Patil *et al.* [128]
35	Marati	Language Modelling for ASR Systems to be used in rural areas	HMM	2013	Tejas Godambe [129]
36	Marati	Phoneme Recognition On Android OS platform	DTW	2014	Saroj B *et al.* [130]
37	Marati	Isolated Word Recognition	MFCC and DTW features	2010	Gawali *et al.* [131]
38	Marati	Medical Enquiry System	Emphasis on Model	2011	Gaikwad S *et al.* [132]
39	Oriya	Digit Recognition	Bakis Model of HMM	2011	Mohanty and Swain [133]

Continue on the next page

Table 2.2 Summary of recent ASR research in Indian language(cont.)

Sl. No.	Language	Work	Methods Used	Year	Authors
40	Oriya	Isolated Word Recognition	HMM	2010	Mohanty and Swain [134]
41	Urdu	ContinuousSpeech Recognition	HMM , developed Urdu Auto completer and Lexicon Development Utility	2009	Raza *et al.* [135]
42	Urdu	ContinuousSpeech Recognition	HMM based	2010	Sarfraz H [136]
43	Kannada	Isolated Word Recognition	DWT and PCA	2010	M. A. Anusuya *et al.* [137]
44	Kannada	Isolated Word Recognition	SVM	2012	Sarika Hegde *et al.* [138]
45	Kannada		Vector Quantisation	2012	M A Anusuya *et al.* [139]
46	Kannada	Digit Recognition	DWA and PCA	2013	Shiva Kumar [140]

Continue on the next page

Table 2.2 Summary of recent ASR research in Indian language(cont.)

Sl. No.	Language	Work	Methods Used	Year	Authors
47	Kannada	Isolated Word Recognition	SVM	2013	Sarika Hegde [141]
48	Telugu	Isolated Vowel Recognition	Neural Networks	2003	Sai Prasad P. S. V. S *et al.* [142]
49	Telugu	Continuous Speech Recognition	Group Delay Function and MFCC	2004	Hegde R.M *et al.* [143]
50	Telugu	Continuous Speech Recognition	Modified Group Delay Function	2005	Hegde R.M et.al. [144]
51	Telugu	Dictation Sytem	Statistical analysis for framing syllabication rules	2009	Sunitha .K.V.N *et al.* [145]
52	Telugu	Syllable based Recognition Analysis	Emphasis on Syllable extraction	2012	Sunitha .K.V.N *et al.* [146]
53	Telugu	Speech Recognition	HMM	2012	Vijai Bhaskar P *et al.* [15]

Continue on the next page

Table 2.2 Summary of recent ASR research in Indian language(cont.)

Sl. No.	Language	Work	Methods Used	Year	Authors
54	Telugu	Speech Recognition	Error Analysis based improvements	2012	Usha Rani N *et al.* [147]
55	Gujarati	IsolatedWord Recognition	Adapted English Speech Recognition System	2011	Himanshu N. Patel *et al.* [148]
56	Gujarati	Telephone Command Processing	DTW	2014	Pandit P. *et al.* [149]
57	Gujarati	Digit Recognition	DTW	2014	Pandit P. *et al.* [149]
58	Gujarati	Speech Recognition	Neural Network	2013	Patel Pravin *et al.* [150]
59	Bodo	Noise Robust Speech Recognition	ANN	2011	M.K.Deka *et al.* [151]
60	Bodo	TonalWord Recognition	RNN based Phoneme Recogniser	2013	Utpal Bhattacharjee [152]

Continue on the next page

Table 2.2 Summary of recent ASR research in Indian language(cont.)

Sl. No.	Language	Work	Methods Used	Year	Authors
61	Malayalam	Large Vocabulary word recognition	lexicon based heuristics and HMMs		Lajish V.L et al. [153]
62	Malayalam	Isolated Word Recognition	HMM	2006	Lajish V.L *et al.* [154]
63	Malayalam	Isolated Word Recognition	ANN with wavelet basd features	2008	V.R.V. Krishnan *et al.* [8]
64	Malayalam	Speech Recognition	HMM based	2008	Syama R *et al.* [155]
65	Malayalam	Question Word Recognition for Speech Query System	ANN	2010	A.R. Sukumar *et al.* [156]
66	Malayalam	Continuous Speech Recognition	HMM	2010	Anuj Mohamed *et al.* [157]
67	Malayalam	Continuous Speech Recognition	HMM – Hybrid Model ANN	2012	Anuj Mohamed *et al.* [158]
68	Malayalam	Isolated Word Recognition	ANN with wavelet basd features	2012	Sonia Sunny *et al.* [7]

Continue on the next page

Table 2.2 Summary of recent ASR research in Indian language(cont.)

Sl. No.	Language	Work	Methods Used	Year	Authors
69	Malayalam	Short Query based Directory Access	HMM	2013	Lajish V.L *et al.* [159]
70	Malayalam	Keyword Spotting	HMM	2015	Vivek P *et al.* [160]

In the last few years, speech to speech systems and multilingual speech recognition approaches has been emerging to accommodate the diversity in the Indian Language domain. Sweety *et al.*, attempts bilingual speech recognition in Assamese and English language. The proposed approach uses MFCC and ANN for speech recognition [161]. A multilingual speech processing system for answering travel related and other emergency queries was developed by IIIT Hyderabad [162] to support the tourism industry. Tejaswi *et al.* developed acoustic models using deep neural networks (DNN) for low resource languages [163]. Experiments were conducted on Kannada (low resource) and Telugu data and 46% relative improvement is observed in multi-task framework over its mono-lingual DNN. Thus in recent works on speech recognition, various speech processing tasks are integrated to realise multilingual speech signals.

2.8 Conclusion

The studies on recent developments in the evaluation of keyword spotting system based on different strategies for feature selection and classifi-

cation techniques are reviewed in this chapter. Studies on allophonic properties and durational analysis of phones in different languages are also considered and reviewed. A detailed discussion has been carried out on the progress and current status of research in Grapheme-to-Phoneme methods. It is observed that not many research works in these area have been carried out in Malayalam language. It is marked that the durational information and pre-processing tools like G2P converter has a relevant role in regional language computing. A quick review on the keyword spotting and speech analytics techniques have been performed. Significant works in two emerging speech analytics research areas, audio classification and speaker spotting are also reviewed in detail. Finally a special review on recent advances in speech research in Indian languages is also conducted with a special emphasis on Malayalam speech processing.

Chapter 3

A Study on Malayalam Phonology based on Durational and Spectral Characteristics of Allophonic Variations in Vowel Phonemes

3.1 Introduction

Malayalam, the mother tongue of the south Indian state Kerala is spoken by more than 40 million people in Kerala, Lakshadweep Islands, Mahe *etc.* [164]. It is the youngest in the Dravidian language family and conferred the classical language status by the Government of India in 2013. Malayalam is included in the list of top 30 languages spoken in the world and top 10 languages in India. The dual rooted evolution of Malayalam from Sanskrit and Tamil make it unique in Indian languages [165].

Speech recognition and speech synthesis in Malayalam is an active research area [166, 167, 6]. But the attempts have not yet reached the realm of successful development of Malayalam speech technology solutions that are useful for the masses. Further study is essential on

the basic structure and peculiarities of the language in the perspective of Malayalam language computing. Malayalam is an alphasyllabary language with the *aksharam* (character) as its core. It may be also noted that almost one to one correspondence exists between orthographic symbols in the alphabet and phoneme in Malayalam [9].

Phones are considered to be the basic unit in speech and a phone is a physical sound produced when a particular phoneme is articulated. There will be an infinite number of phones corresponding to a phoneme due to the underlying variability of vocal tract configurations. Allophone is a class of phones corresponding to a specific variant of a phoneme. The utterance of a phoneme is affected by the context it appears. This co-articulation effect is mainly responsible for allophonic variations. Forward or anticipatory co-articulation refers to the changes due to anticipatory phonemes. Backward or preservatory co-articulation is caused by the inertia of the uttered phonemes [168]. It is evident that each phoneme existing in speech will be in any of its allophonic forms.

A well-defined rule-based structure aids for almost every allophone formation in Malayalam [164, 9]. These contextual rules explains the allophonic variability in terms of the place of occurrence of the phone (including initial, middle and final) and the neighbouring phonemes. These rules can be effectively used for speech synthesis where concatenating allophones are employed. The sequence of graphical signs, named as graphemes, can also be converted to the corresponding sequence of allophones (Grapheme to Allophone conversion) based on these rules. To the best of our knowledge, no detailed study has been reported in Malayalam which places allophones as the basic unit in speech recognition or synthesis.

In this work, long and short vowel allophones in Malayalam are identified and listed. Detailed analysis of the durational and the spectral

differences of the first two formants are performed on the allophonic inventory. Rest of this chapter is organized as follows. Section 3.2 describes the properties of Malayalam phones and allophones. Section 3.3 discuss the process of finding the extensive rule set for the formation of Malayalam vowel and consonant allophones. This section also introduces the TEMU Malayalam phonetic archive. Section 3.4 discusses durational properties of Malayalam vowel allophones and the session 3.5 presents the studies on spectral differences of vowel allophones based on formant frequencies. Section 3.6 discusses allophonic clustering of Malayalam vowels and section 3.7 concludes the work.

3.2 Malayalam Phoneme Set and Allophones

A phone refers to the instances of phonemes in the actual utterances. The smallest meaningful distinctive sound unit in a language is called a phoneme. The phoneme set varies considerably from one language to another. International Phonetic Alphabet (IPA) defines around 150 phones among all languages. American English has around 40 phonemes. Malayalam has 11 vowel phonemes, 2 diphthongs, and 37 consonant phonemes together constitute a 50 member phoneme set [27]. The following section describes the characteristics of phonetic inventory of Malayalam language and its allophonic variations in detail followed by an introduction to the the structure of the Malayalam phonetic archive used in this study.

3.2.1 Malayalam Vowel Phones

The vowels are produced by exciting an essentially fixed vocal tract shape with quasi-periodic pulses of air caused by the vibration of the

vocal cords [169]. The sound difference is produced by changing the shape and position of lip and tongue. Diphthongs are considered as the combination of vowels as an in-between smooth transition happens between the vowel configurations.Malayalam has 11 monophthongs and 2 diphthongs [170]. Based on articulation, among the 10 vowels, 4 are classified as front vowels (ഇ [i], ഈ [i:], എ [e], ഏ [e:]), two as central vowels (അ [a], ആ [a:]) and four classified as back vowels (ഉ [u], ഊ [u:], ഒ[o], ഓ [o:]). Ideally, the frequently used semivowel (ˇ / ə /) can either be treated as a separate phoneme or as an allophonic variation of ഉ /u/. In this work we have considered ˇ/ə/ as an allophone of /u/. Another sound ഋ /ṛ/ is not included in the study as it is rarely used in normal speech. The complete list of Malayalam vowels and diphthongs with its allophones are shown in figure 3.1.

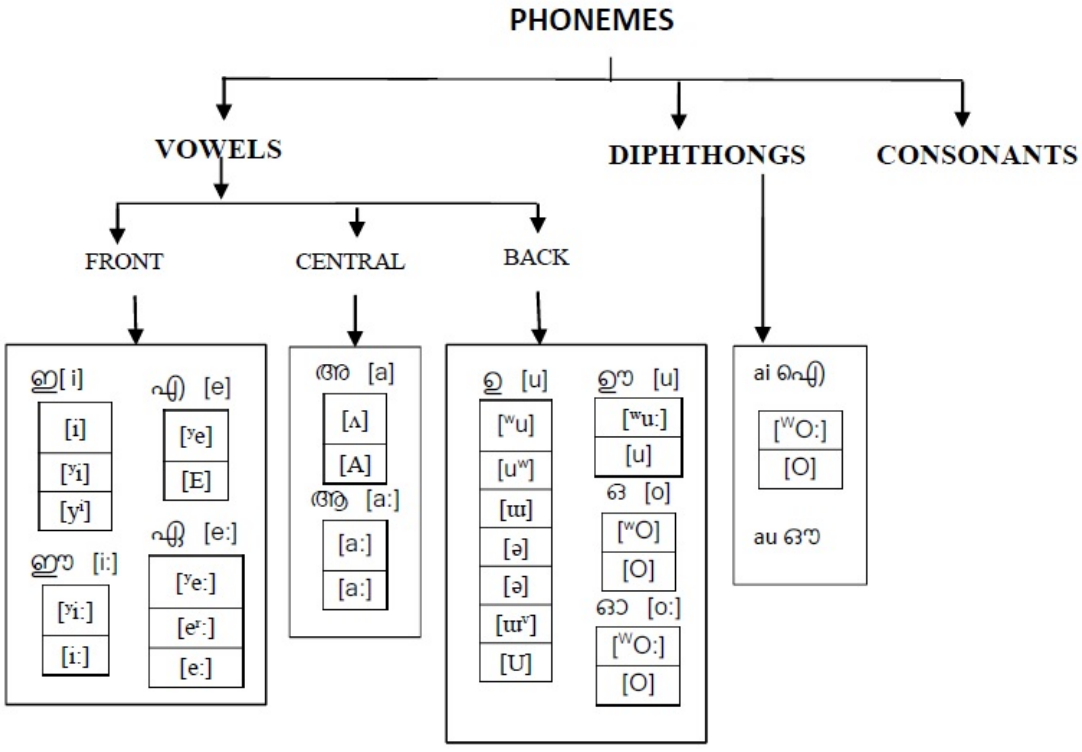


Fig. 3.1 Malayalam Vowels and Diphthongs and its allophonic variations

As part of this study, 107 allophones in Malayalam which include 76 consonant allophones, 28 vowel allophones and 3 allophones corresponding to diphthongs are identified. The vowels have relatively high degree of allophonic variability. It is also reported that, while comparing with the ideal positions of vowels in the *Cardinal Vowel Figure* [169], Malayalam vowels are slightly displaced towards inside from the vowel area borders [9].

3.2.2 Malayalam Consonant Phones

A consonant is a speech sound that is articulated with the complete or partial closure of the vocal tract [169]. The classification problem of consonants is more challenging than that of vowels. There are more parameters of contrast with the neighbours on consonant classification. Consonants can be broadly categorized into plosive (stop consonant), nasal, fricative, flapped, lateral, approximant and glide. Malayalam has 38 consonants in which 21 of them are plosives. Plosives are again classified based on voice and aspiration. The classification of Malayalam plosives are given in table 3.1 and the consonants other than plosives are listed in table 3.2.

Table 3.1 Classification of Malayalam plosives

	Voice	**Bilabial**	**Dental**	**Alveolar**	**Retroflex**	**Palatal**	**Velar**
Plosive	Voiceless unaspirated	പ [p]	ത [t]	റ [r̪]	ട [ʈ]	ച [c]	ക [k]
	Aspirated	ഫ [p^{h}]	ഥ [t^{h}]		ഠ [ʈ^{h}]	ഛ [c^{h}]	ഖ [k^{h}]
	VoicedUnaspirated	ബ [b]	ദ [d]		ഡ [ɖ]	ജ [ɟ]	ഗ [g]
	Aspirated	ഭ [b^{h}]	ധ [d^{h}]		ഢ [ɖ ^{h}]	ഝ [ɟ^{h}]	ഘ [g^{h}]

Table 3.2 Classification of Malayalam consonants other than plosives

	Bilabial	Labiodentals	Dental	Alveolar	Retroflex	Palatal	Velar	Glottal
Nasal	മ [m]		ന [n̪]	ന [n]	ണ [ɳ]	ഞ [ɲ]	ങ [ŋ]	
Fricative				സ [s]	ഷ [ʂ]	ശ [ʃ]		ഹ [h]
Trill/flapped				ര[r] റ[ɾ]				
Lateral				ല [l]/ ൽ	ള/ ൾ [ɭ]			
Approximant					ഴ [ɻ]			
Glide		വ [ʋ]				യ [j]		

3.3 Founding an Exhaustive Rule Set for Malayalam Allophone Formation

The variation in the duration of a phoneme can be attributed to many factors. Some factors such as contextual variability can be detected from the text while some others such as dialect cannot be detected from the text [22]. In their separate works, Asher and V.R. Prabodhachandran Nair have described the language rules of Malayalam allophone formation in detail [165, 9]. They have proposed linguistic descriptions for defining the allophones of each Malayalam phones. It further reveals that these findings can be transformed to a position and neighborhood information-based rule set that defines the formation of vowel allophones in Malayalam. The following section describes the process of finding rule set for the formation of Malayalam vowel allophones in detail.

3.3.1 Rule set for the formation of Malayalam Vowel Allophones based on the Position and Neighbourhood Information

A rule set for the formation of vowel allophones in Malayalam based on the position and neighbouring information is created. In the case of vowel ഇ /i/, the three allophones [I],[^{y}i] and [i^{y}] are formed correspond to medieval, initial and final position of occurrence of the phone ഇ /i/ . The vowel phone, ഉ /u/ has 7 allophones in Malayalam. The allophone [^{w}u] is characterized by the initial position and the second allophone [u^{w}] is characterized by the presence of off-glide in the final position. The third allophone [ɯ] occurs in the middle syllable. The fourth allophone [ɯ^{v}] occurs in Sanskrit originated words after certain phonemes. The fifth allophone [U] happens if preceded by an initial consonant in a word and finally the last two allophones correspond to the semivowel [ə]. In general, this allophonic categorizations in Malayalam is sufficient to capture the variations due to place and co-articulation effects. Table 3.3 represents these newly framed rule set to denote the formation of Malayalam vowel allophones.

Table 3.3 Position and neighbourhood based rule set to denote the formation of Malayalam vowel allophones

Sl.No.	Phoneme	Allohone	Position	Rule
1	ഇ[i]	[i]	Middle	**Metadata**: Low high front unrounded short vocoid. **Neighbourhood**: Any

Continue on the next page

Table 3.3 Position and neighbourhood based rule set to denote the formation of Malayalam vowel allophones(cont.).

Sl.No.	Phoneme	Allohone	Position	Rule
		[^{y}i]	Initial	**Metadata**: High front unrounded long tense vocoid with onglide. **Neighbourhood**: Any
		[y^{i}]	Final	**Metadata**: High front unrounded short tense vocoid with offglide. **Neighbourhood**: Any
2	ഇൗ[i:]	[^{y}i:]	Initial	**Metadata**: High front unrounded long tense vocoid with onglide. **Neighbourhood**: Any
		[i:]	Middle	**Metadata**: High front unrounded tense long vocoid; medially. **Neighbourhood**: Any
3	ഏ[e]	[^{y}e]	Initial	**Metadata**: Higher mid front unrounded short tense vocoid with onglide. **Neighbourhood**: Any

Continue on the next page

Table 3.3 Position and neighbourhood based rule set to denote the formation of Malayalam vowel allophones(cont.).

Sl.No.	Phoneme	Allohone	Position	Rule
		[^{e}y]	Initial	**Metadata**: Higher mid front unrounded short tense vocoid with offglide [j]. **Neighbourhood**: Any
		[E]	Middle	**Metadata**: Mean mid front unrounded short vocoid. **Neighbourhood**: Any
4	ഏ[e:]	[^{y}e:]	Initial	**Metadata**: Higher mid front unrounded long tense vocoid with onglide. **Neighbourhood**: Any
		[e^{r}:]	Final	**Metadata**: Higher mid front unrounded long tense vocoid with offglide [j]. **Neighbourhood**: Any

Continue on the next page

Table 3.3 Position and neighbourhood based rule set to denote the formation of Malayalam vowel allophones(cont.).

Sl.No.	Phoneme	Allohone	Position	Rule
		[e:]	Middle	**Metadata**: Higher mid front unrounded long tense vocoid. **Neighbourhood**: Any
5	അ[a]	[ʌ]	Inital& Final	**Metadata**: Low mid back vocoid in the initial syllable and word. **Neighbourhood**: Any
		[A]	Middle	**Metadata**: Low mid central vocoid in the medial syllable. **Neighbourhood**: Any
6	ആ[a:]	[a:]	-	**Metadata**: Low back long tense vocoid after velar consonants. **Neighbourhood**: Left : Velar Consonants

Continue on the next page

Table 3.3 Position and neighbourhood based rule set to denote the formation of Malayalam vowel allophones(cont.).

Sl.No.	Phoneme	Allohone	Position	Rule
		[a:]	-	**Metadata**: Low central long vocoid after all non-velaric consonants. **Neighbourhood**: Left : Non-velar Consonants
7	ഉ[u]	[^{w}u]	Initial	**Metadata**: High back rounded tense short vocoid with onglide [w]. **Neighbourhood**: Any
		[u^{w}]	Final	**Metadata**: Higherback rounded tense short vocoid with offglide [w]. **Neighbourhood**: Any
		[ɯ]	Middle	**Metadata**: High back unrounded short vociod in the medial syllable. **Neighbourhood**: Any

Continue on the next page

Table 3.3 Position and neighbourhood based rule set to denote the formation of Malayalam vowel allophones(cont.).

Sl.No.	Phoneme	Allohone	Position	Rule
		[ə]	Final	**Metadata**: Higher,mid central unrounded open vocoid, between consonant and vowel inthe word boundary with open juncture. **Neighbourhood**: Any Other:Open juncture.
		[ə*]	-	**Metadata**: In open juncture 'ə' is in free variation after the following,some phonemes. **Neighbourhood**: [ɳ̩] [l̩] [r̩] [l̩]

Continue on the next page

Table 3.3 Position and neighbourhood based rule set to denote the formation of Malayalam vowel allophones(cont.).

Sl.No.	Phoneme	Allohone	Position	Rule
		[ɯ^{v}]	-	**Metadata**: Low high back unrounded vocoid after some phonemes. **Neighbourhood**: Right :Labial, dental, palatal, and velar plosives and labial, anddental nasals and non-retroflex fricatives and labio-dental,continuants.Other:In-Sanskrit words
		[U]	-	**Metadata**: Low high back rounded tense vocoid, after word initial consonant. **Neighbourhood**: Right :Consonant as the first letter in a word

Continue on the next page

Table 3.3 Position and neighbourhood based rule set to denote the formation of Malayalam vowel allophones(cont.).

Sl.No.	Phoneme	Allohone	Position	Rule
8	ഊ/u:/	[^{w}u:]	Initial	**Metadata**: High back rounded long tense vocoid with onglide [w]. **Neighbourhood**: Any
		[u]	Other than Initial	**Metadata**: High back rounded tense vocoid, elsewhere. **Neighbourhood**: Any
9	ഒ/o/	[^{w}O]	Initial	**Metadata**: Higher,mid back rounded tense short vocoid with onglide [w] in the,initial position. **Neighbourhood**: Any
		[O]	Middle	**Metadata**: Mean,mid back tense rounded short vocoid. **Neighbourhood**: Any

Continue on the next page

Table 3.3 Position and neighbourhood based rule set to denote the formation of Malayalam vowel allophones(cont.).

Sl.No.	Phoneme	Allohone	Position	Rule
10	ഓ/o:/	[^{w}O:]	Initial	**Metadata**: Higher mid back rounded tense long vocoid with onglide. **Neighbourhood**: Any
		[O]	Medial and Final	**Metadata**: Higher mid back tense long vocoid. **Neighbourhood**: Any

These ten rules, accountable for the formation of Malayalam vowel allophones shown in Table 3.2, are then effectively used in the implementation of Grapheme to Phoneme (G2P) transcription algorithm as discussed in Chapter 4. The following section describes the formation of rule set for Malayalam consonant allophones.

3.3.2 Rule Set for the formation of Malayalam Consonant Allophones

A rule set for the formation of consonant allophones in Malayalam are also constructed based on the position and neighbouring information. The list of Malayalam consonants with its allophones is shown in table 3.4. It can be seen that there are nineteen consonants with only one allophone and few allophones have more than one rule for its formation. For example, the allophone [v] of the vowel വ /v/ has three rules for the formation based on its position and neighbourhood. In this case the

allophones may occur in (a) word initial position (b) medial position (in clusters where [w] does not appear) (c) mostly short and rarely long intervocalic positions. In another case, the consonant phoneമ [m] has 4 allophones in Malayalam. The allophone [m̪^{h}] occurs before velar fricative. The next allophone [M] occurs in consonant clusters when preceded by alveolar flap. The third allophone [m] is characterized by the presence of labio-dental continuant before it and the final allophone [m] (bilabial nasal) occurs elsewhere other than the above three allophones.

Table 3.4 List of Malayalam consonant allophones

Sl.No.	Phoneme	Allophone
1	പ [P]	[p] [β] [b] [P]
2	ഫ[p^{h}]	[p^{h}]
3	ബ[b]	[B] [b]
4	ഭ [b^{h}]	[b^{h}]
5	മ [m]	[m̪^{h}] [M] [m] [m]
6	വ [v]	[w] [v]
7	ത[t]	[t] [t'] [ð]

Continue on the next page

Table 3.4 List of Malayalam consonant allophones(cont.).

Sl.No.	Phoneme	Allophone
		[d]
8	ഥ [t^h]	[t^h]
9	ദ [d]	[d̪] [d]
10	ധ [d^h]	[d^h]
11	ന [n]	[n] [n]
12	ഩ [r̲]	[d] [t]
13	ന[n]	[n^h] [n]
14	സ [s]	[s]
15.a	ര [r]	[r]
15.b	റ [ṛ]	[ṛ]
16	ല/ൽ [l]	[l]
17	ട [ʈ]	[ɖ] [ɾ] [ʈ] [T]
18	ഠ [$ʈ^h$]	[$ʈ^h$] [T^h] [h]
19	ഡ [ɖ]	[ɖ]
20	ഢ[$ɖ^h$]	[$ɖ^h$]
21	ണ [ɳ]	[ɳ]

Continue on the next page

Table 3.4 List of Malayalam consonant allophones(cont.).

Sl.No.	Phoneme	Allophone
22	ഷ [ʂ]	[ʂ]
23	ള/ശ [ɭ]	[ɭ]
24	ഴ [ɻ]	[ɻ]
25	ച [c]	[c] [c̣] [ɟ] [C]
26	ഛ [c^h]	[c^h] [C^h]
27	ജ [ɟ]	[J] [j]
28	ഝ [ɟ ^{h}]	[ɟ ^{h}]
29	ഞ [ɲ]	[ɲ]
30	ശ [ʃ]	[ʃ]
31	യ [y]	[y]
32	ക [k]	[k] [kj] [Y̠] [g̠] [t̠] [K]
33	ഖ [k^h]	[k^h] [K^h] [K^h]
34	ഗ [g]	[G]

Continue on the next page

Table 3.4 List of Malayalam consonant allophones(cont.).

Sl.No.	Phoneme	Allophone
		[g]
35	ഘ [g^h]	[g^h]
36	ങ [ŋ]	[ŋ] [ŋj] [ŋ$^<$] [ŋ$^>$] [ŋ']
37	ഹ [h]	[H]

3.3.3 TEMU Malayalam Phonetic Dataset

In this work, an inclusive Malayalam phonetic dataset which is being designed and developed in collaboration with the Malayalam phonetic archive project owned by Thunchath Ezhuthachan Malayalam University (TEMU), Kerala, India which is used for the experimental purposes [171]. It is a fairly comprehensive dataset created by taking into consideration of a carefully compiled inventory of phones which are currently employed in the Malayalam language. The author of this work is also involved directly in the aforementioned project by providing proper directions for content listing and technical support for digitization and web publishing. Malayalam phoneme segments are recorded in its standardized orthography followed by a number of examples of its occurrence in phonologically relevant different positions. Allophones are listed together and pronunciation of each example recorded from the natural speech is demonstrated in both male and female voices. The dataset comprises of 11 vowels, 2 diphthongs and 38 consonants, and its allophonic variation with 900 spoken words as examples. It is presently available as a web portal in

public domain under creative common license [171]. The following section describes the durational properties of Malayalam vowel allophones derived on the basis of the detailed analysis performed on the TEMU dataset.

3.4 Durational Properties of Malayalam Vowel Allophones

This section describes a detailed analysis conducted to obtain the durational properties of Malayalam vowel allophones. These findings could be the basis for the preliminary works that are conducted to incorporate duration-specific knowledge to build speech recognition or speech synthesis systems. Phoneme segmentation is the most important preprocessing task to be performed in the phoneme level speech recognition. In phoneme segmentation algorithms, mostly the phonemes are assumed to be of the same length and segmented using a fixed size window. Durational analysis of phonemes performed in many languages reveals the variability in the duration of individual phonemes [28]. So the phone segmentation algorithms must consider the variability in phoneme duration for an improved performance. The phoneme level duration variability is language specific. Considering these facts a detailed analysis is conducted to establish durational phoneme models for Malayalam language.

The duration of each individual vowel phones and its allophonic examples are computed from the utterances of the same phone by different male and female speakers. Figure 3.2 shows the spectral representation indicating the duration of the allophone [^{y}e] of the vowel എ /e/ obtained from the word എലി [eli]. It is observed that, this duration computed for the allophone [^{y}e] varies from the duration of the other allophones [e^{y}] and [E] of the same vowelഎ /e/.

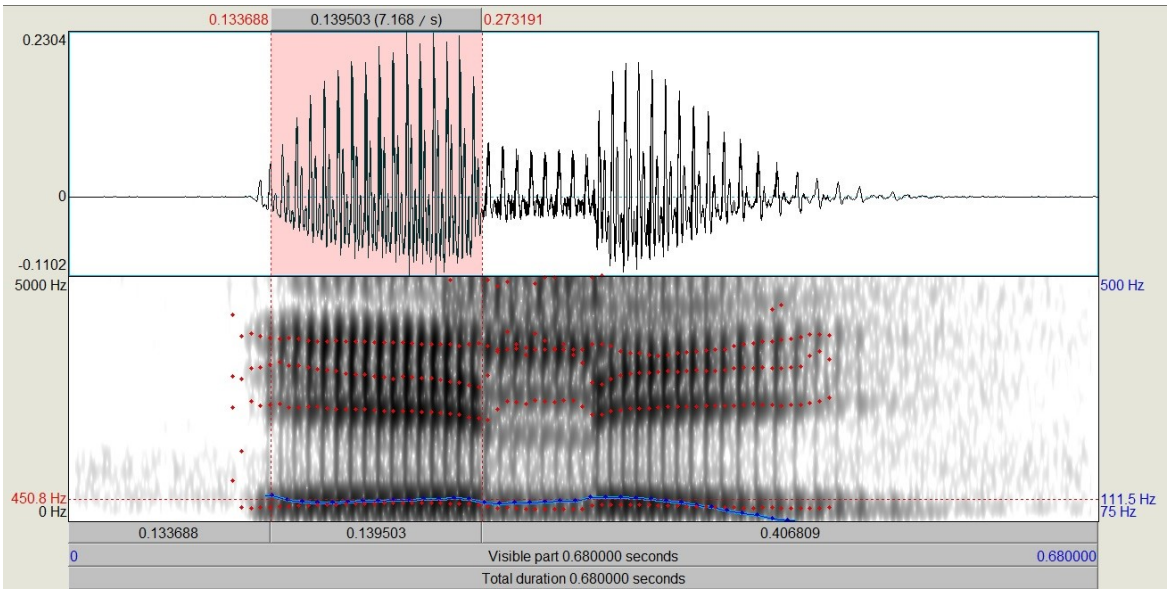


Fig. 3.2 Duration of the allophone [ʸe] of the word [eli]

We have selectively used 238 Malayalam words from TEMU dataset accommodating all vowel allophonic variability for the conduct of the experiments. The average durations of all allophonic variations of ten Malayalam vowels are computed distinctly from the selected set of words comprising that allophone taken from the TEMU dataset. Figure 3.3 shows the average durations of seven different allophones of the vowel ഉ /u/ computed distinctly from the selected words comprising these allophones. From the figure, it is evident that there exist significant variabilities in the duration obtained for different allophones of the vowel phoneme ഉ /u/.

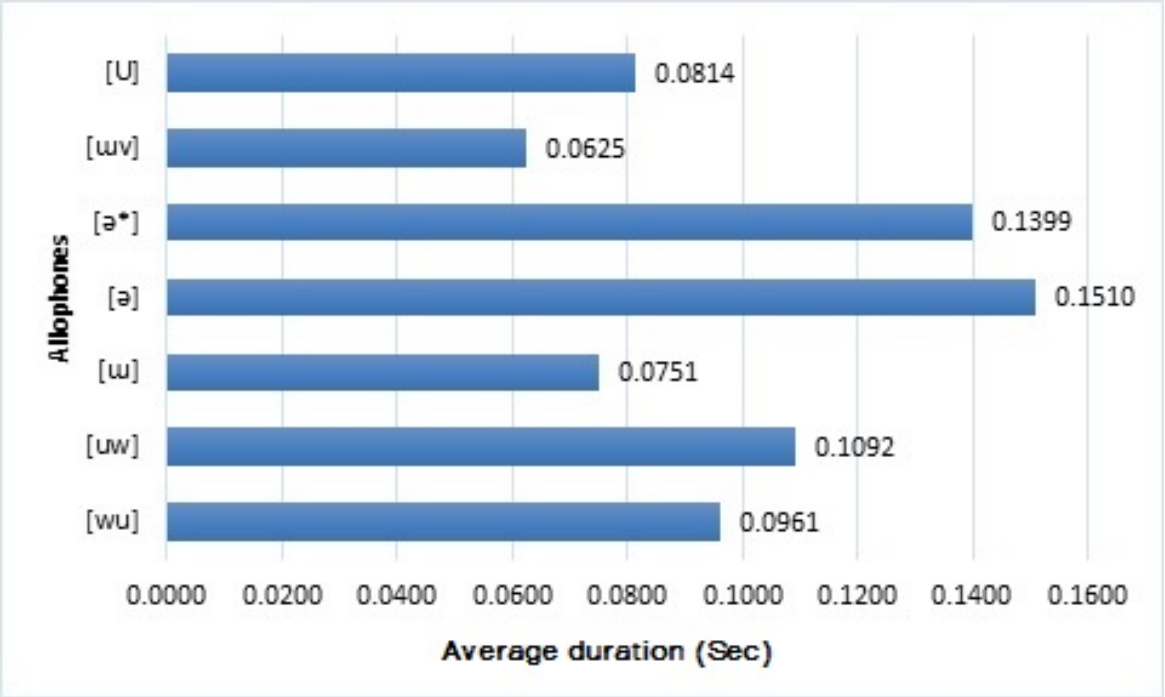


Fig. 3.3 Average Durations of seven allophones of the Malayalam vowel ഉ /u/

The mean duration obtained for the of ten Malayalam vowels including all the allophonic variations is 0.14284 sec for the male speaker and 0.15563 sec for the female speaker. The duration range is from 0.04155 sec to 0.28936 sec for the male, and 0.04316 sec to 0.33001 sec for the female speaker. The detailed statistics including the average duration of allophones corresponding to each Malayalam vowel together with the Mean and Standard Deviation (SD) are computed and listed in table 3.5. The speech samples from the TEMU dataset is used to conduct this analysis.

Table 3.5 Durational statistics of Malayalam vowel allophones

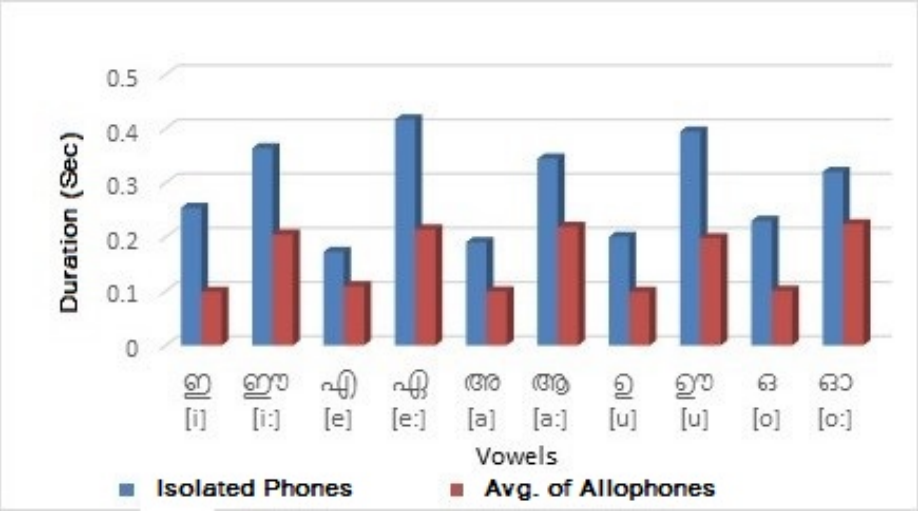
Sl.No	Vowel	Allophones	Average Duration (sec)	
			Male	Female
1	ഇ[i]	[i]	0.08871	0.12866
		[^{y}i]	0.11778	0.14847
		[^{y}i]	0.09356	0.10311
		Mean	0.09871	0.12617
		SD	0.01758	0.02210
2	ഈ [i:]	[^{y}i:]	0.20628	0.22192
		[i:]	0.20229	0.22763
		Mean	0.20429	0.22478
		SD	0.03948	0.04673
3	എ[e]	[^{y}e]	0.12792	0.14015
		[E]	0.08989	0.10110
		Mean	0.10890	0.12062
		SD	0.02454	0.02778
4	ഏ[e:]	[^{y}e:]	0.24340	0.25133
		[^{r}e:]	0.20149	0.11787
		[e:]	0.20317	0.24064
		Mean	0.21391	0.20435
		SD	0.03674	0.06900
5	അ[a]	[∧]	0.11452	0.15527
		[A]	0.08414	0.08588
		Mean	0.09933	0.12057
		SD	0.01819	0.03656
6	ആ[a:]	[a:]	0.22137	0.26522
		[a]	0.21536	0.27507

Continue on the next page

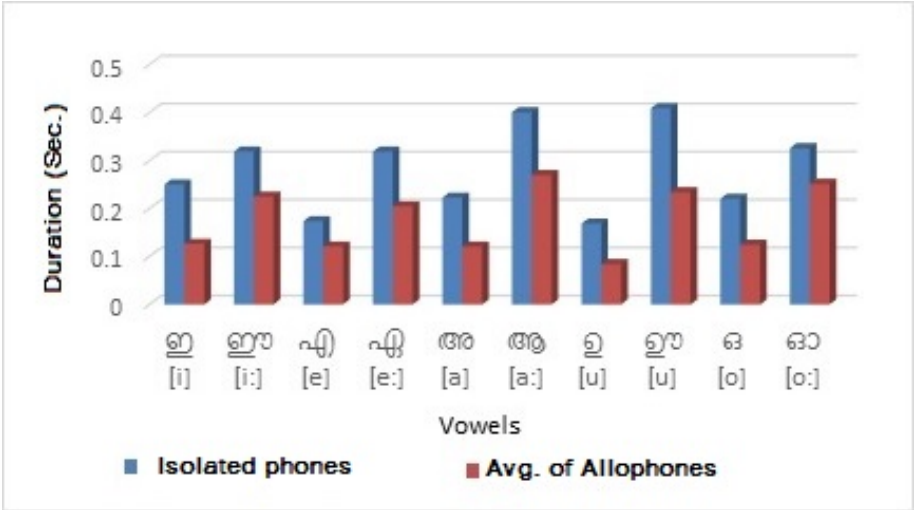
Table 3.5 Durational statistics of Malayalam vowel allophones(cont.).

Sl.No	Vowel	Allophones	Average Duration (sec)	
			Male	Female
		Mean	0.21870	0.26960
		SD	0.02341	0.02772
7	ഉ[u]	[ʷu]	0.09610	0.10346
		[uʷ]	0.10917	0.08431
		[ɯ]	0.07510	0.07979
		[ə]	0.15100	0.08847
		[ə*]	0.13990	0.08253
		[ɯv]	0.06245	0.07106
		[U]	0.08139	0.07943
		Mean	0.09856	0.08409
		SD	0.03565	0.01455
8	ഊ/u:/	[ʷu:]	0.19784	0.24334
		[u]	0.19789	0.21948
		Mean	0.19786	0.23380
		SD	0.02819	0.04677
9	ഒ/o/	[ʷO]	0.10789	0.11639
		[O]	0.09402	0.13223
		Mean	0.10096	0.12431
		SD	0.01459	0.02296
10	ഓ/o:/	[ʷO:]	0.24083	0.26733
		[O]	0.20402	0.23338
		Mean	0.22351	0.25135
		SD	0.03539	0.04161

A comparison is performed on the durations computed for the isolated vowel phones and the average durations obtained for the vowel allophones extracted from the example word set present in the TEMU dataset. Figure 3.4a - 3.4b shows the duration of the isolated vowel phones together with the average duration for each vowel phoneme extracted from the allophonic variations obtained from the TEMU dataset for both male and female speakers.



(a) **Male Speaker**



(b) **Female Speaker**

Fig. 3.4 Duration of isolated vowel phones and the average duration of its allophonic variations

The following section describes the analysis of the spectral characteristics of the Malayalam vowel allophones particularly based on their formant frequencies.

3.5 Formant Frequency Analysis of Malayalam Vowel Allophones

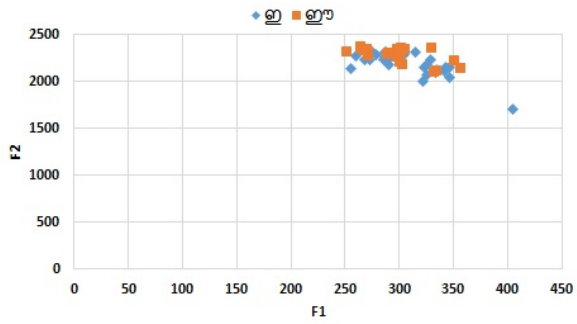
Formant frequency is identified as acoustic resonance of vocal tract reflected as a spectral peak in the sound spectrum. The lowest frequency formant is called F1 followed by F2 and F3. It is obvious from the literature that the first two formants F1 and F2 are sufficient to recognize a vowel in most of the languages [169]. In this work the first formant (F1) and second formant (F2) are computed from the isolated utterances of each vowel phone and also from all its allophonic variations. The Malayalam short and long vowels are grouped into five pairs including (ഇ /i/, ഈ /i:/), (എ /e/, ഏ /e:/), (അ /a/, ആ /a:/), (ഉ /u/, ഊ /u:/), (ഒ /o/, ഓ /o:/). A scatter-plot is then constructed by plotting F1 against F2 computed from each allophonic examples corresponding to a vowel pair. Similarly scatter-plots corresponding to all such five vowel pairs are constructed separately for both male and female speakers.

Figure (3.5a - 3.5e) and figure (3.6a - 3.6e) show the scatter-plot of F1 and F2 values for the short and long vowel pairs extracted from the speech samples corresponding to male and female speaker respectively. From the scatter-plots, it is evident that the formant frequencies of the short and long vowel pairs corresponding to each Malayalam vowel fall in same range.

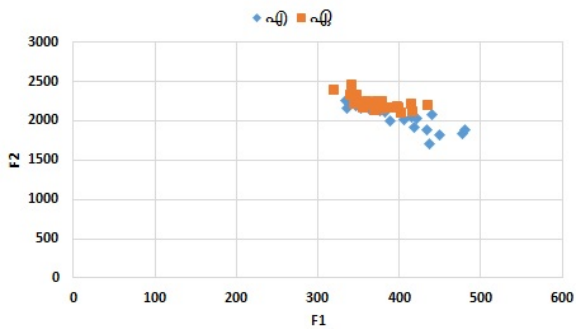
As another experiment, the scatter-plot of F1 and F2 values, computed from each allophonic examples,corresponding to all the five short vowels

are constructed. Figure (3.7a - 3.7b) show the scatter plot of F1 and F2 values obtained for the five Malayalam short vowels. From the plots it can be seen that the allophones of the same short vowel tend to form a cluster based on their formant frequencies. Five such clusters corresponding to each vowels can be distinguished from these plots.

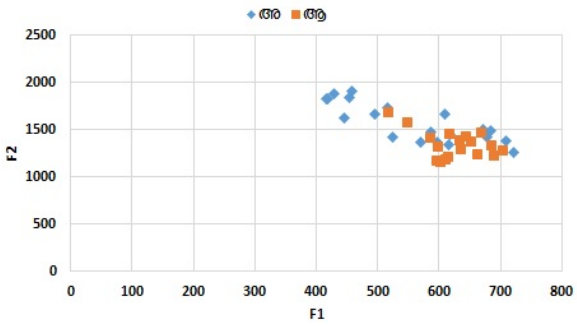
From these analysis it can be concluded that the formant frequencies show the capability to distinguish Malayalam short vowels whereas cannot be considered as efficient parameter for the classification of short and long vowels. Finally, an experiment is conducted to analyse the formant frequency distribution among the allophones of the Malayalam vowel. For further analysis F1 and F2 of the allophonic examples of the Malayalam vowels അ /a/, ഇ /i/ and ഉ /u/ are computed separately for male and female speakers. Figure (3.8a - 3.8c) and figure (3.9a -3.9c) show the scatter plots of F1 and F2 values obtained from the allophones of the aforementioned three Malayalam vowels corresponding to both male and female speakers respectively. From these F1-F2 plots it is observed that separate clusters are formed corresponding to two allophones for vowel അ /a/, three allophones for vowel ഇ /i/ and seven allophones for vowel ഉ /u/. That is the allophones of the same vowel form clusters among themselves. The result of the formant frequency analysis performed on Malayalam vowel allophones substantiate the major authentic theoretical findings about the allophonic variations of Malayalam vowels reported by Asher and V.R. Prabodhachandran Nair [165, 9].



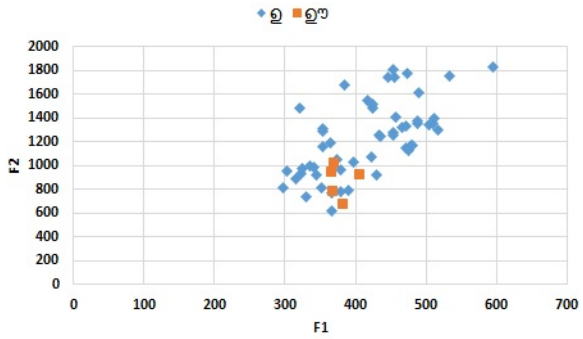
(a) (ഇ /i/, ഈ /i:/)



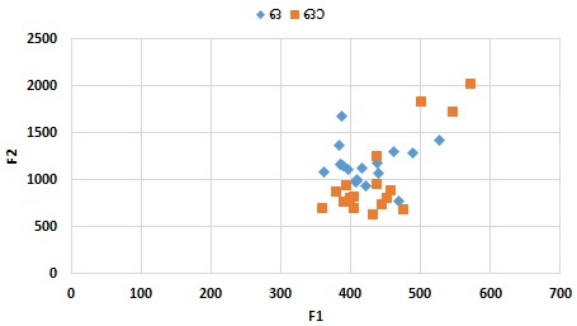
(b) (എ /e/, ഏ /e:/)



(c) (അ /a/, ആ /a:/)

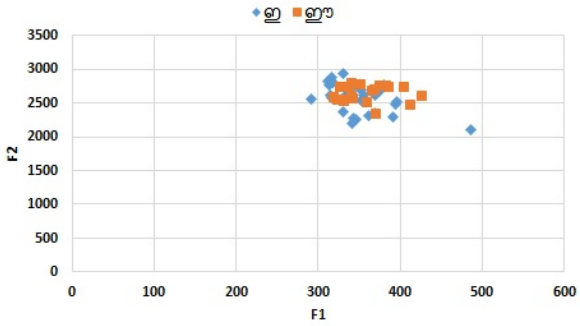


(d) (ഉ /u/, ഊ /u:/)

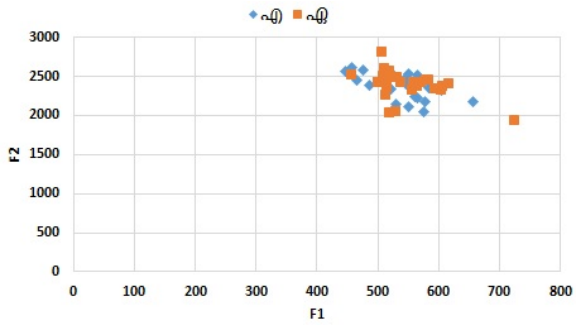


(e) (ഒ /o/, ഓ /o:/)

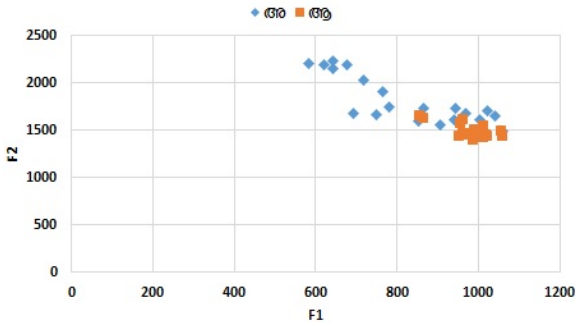
Fig. 3.5 (a-e): F1-F2 scatter plot for short and long vowel pairs of male speakers



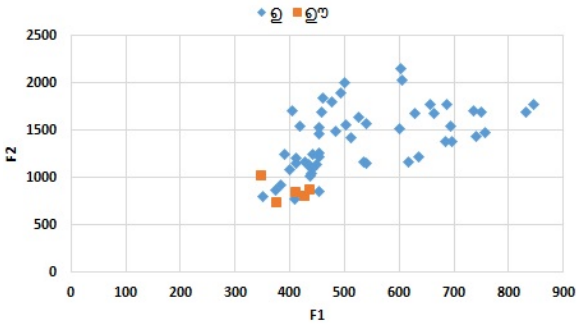
(a) (ഇ /i/, ഈ /i:/)



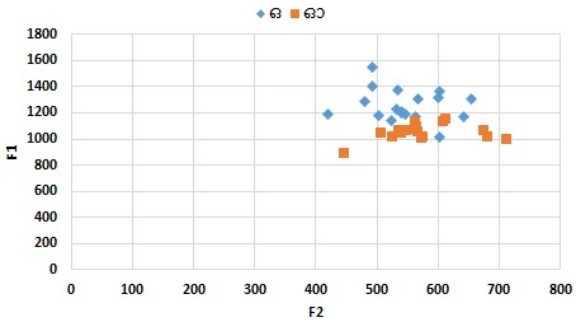
(b) (എ /e/, ഏ /e:/)



(c) (അ /a/, ആ /a:/)

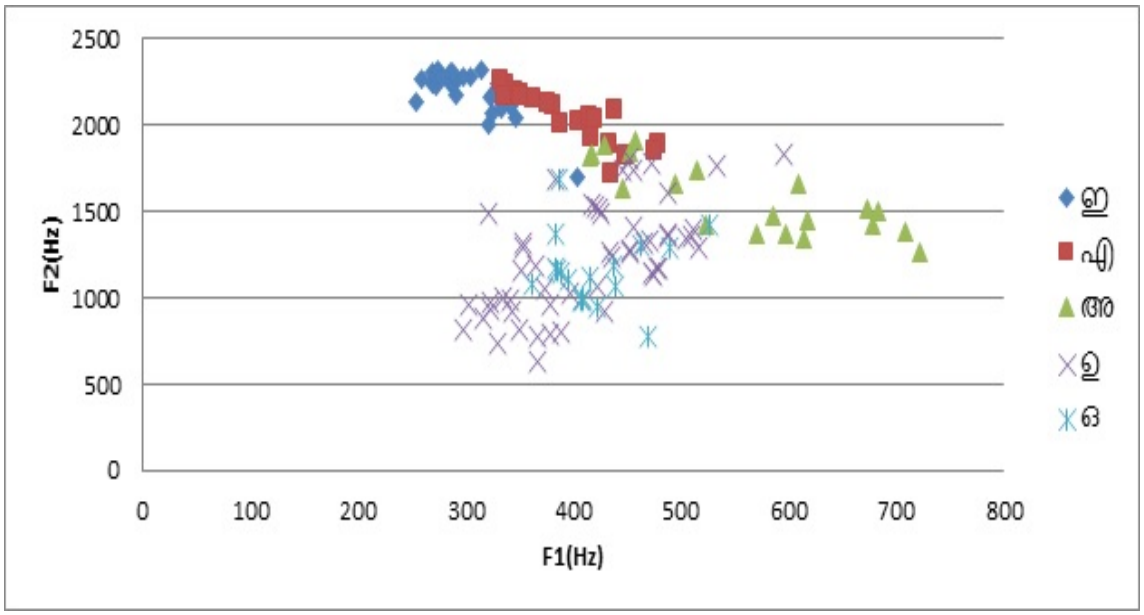


(d) (ഉ /u/, ഊ /u:/)

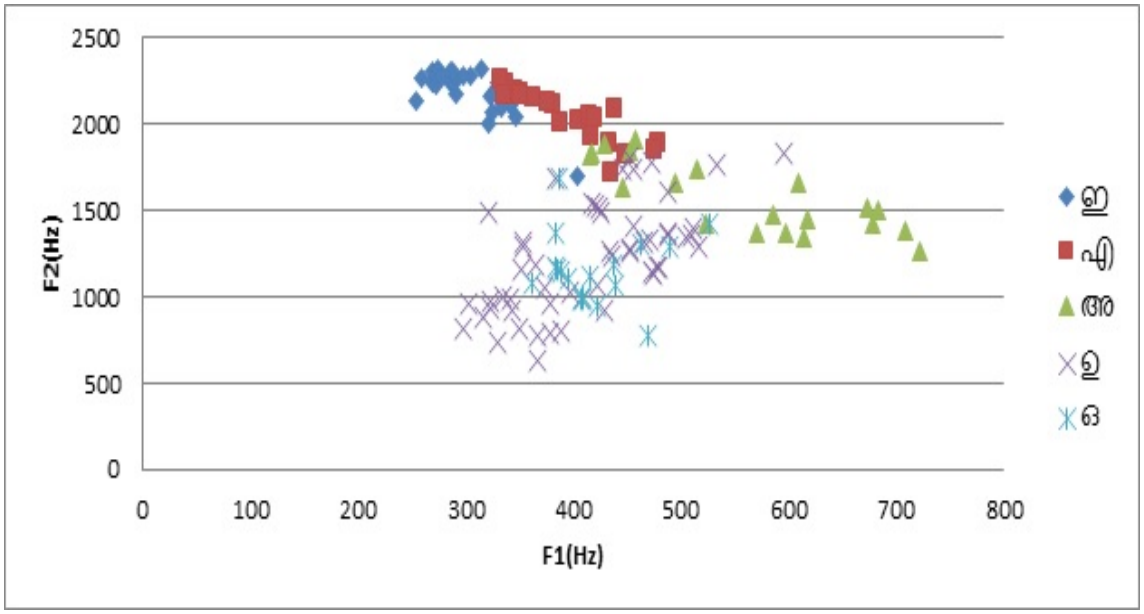


(e) (ഒ /o/, ഓ /o:/)

Fig. 3.6 (a-e): F1-F2 scatter plot for short and long vowel pairs of female speakers

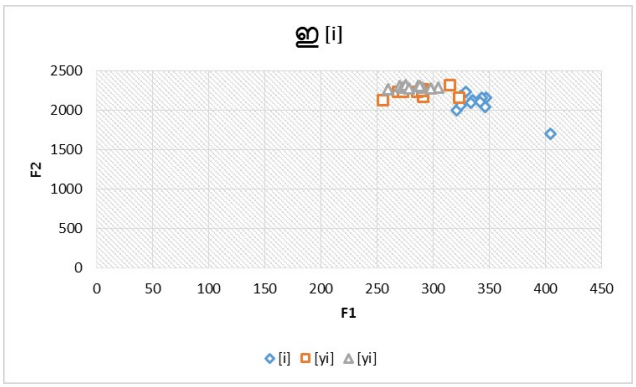


(a) Male Speaker

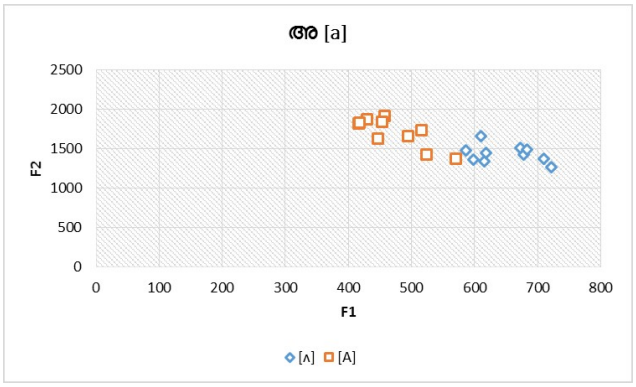


(b) Female Speaker

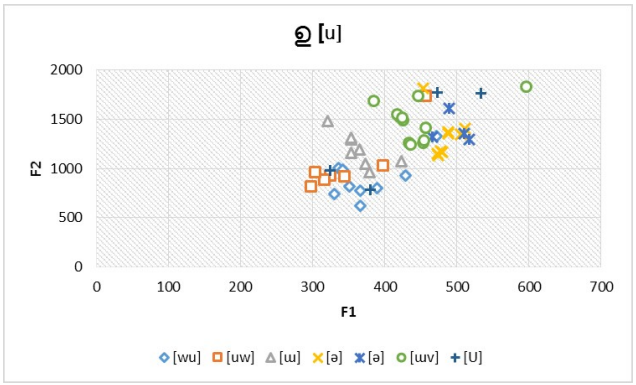
Fig. 3.7 (a-b): F1 - F2 scatter plot of five Malayalam short vowels



(a) Vowel ഇ /i/

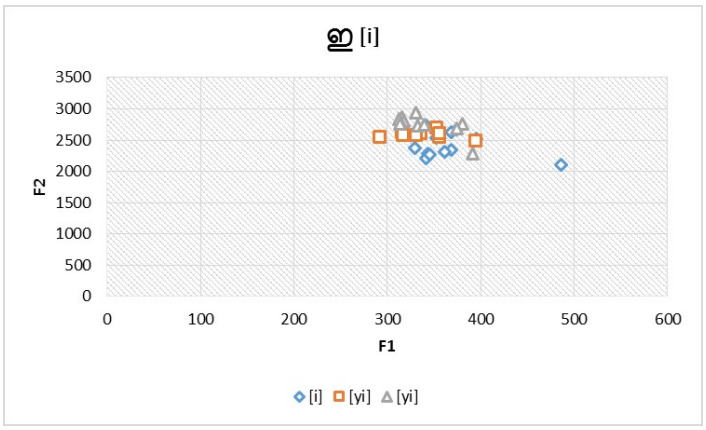


(b) Vowel അ /a/

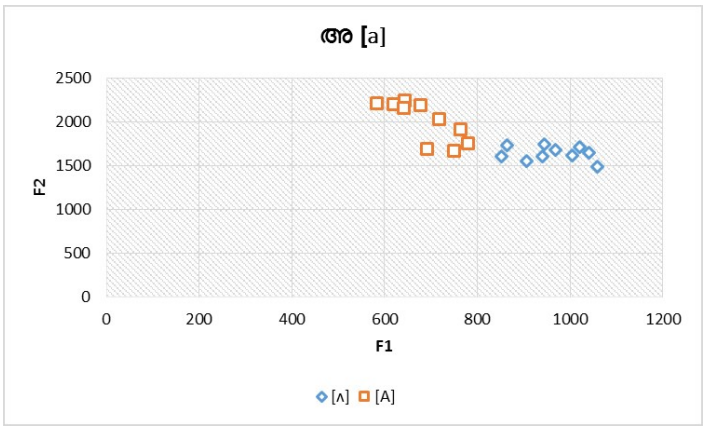


(c) Vowel ഉ /u/

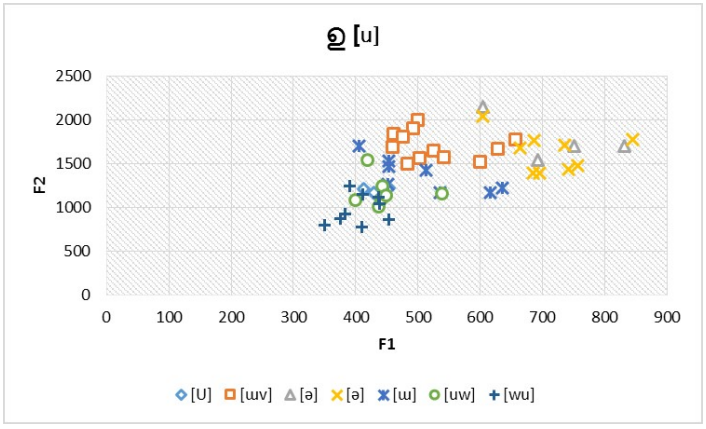
Fig. 3.8 (a-c): F1 - F2 scatter plot of allophones of the Malayalam vowels ഇ /i/, അ /a/, and ഉ /u/ for male speakers



(a) Vowel ഇ /i/



(b) Vowel അ /a/



(c) Vowel ഉ /u/

Fig. 3.9 (a-c): F1 - F2 scatter plot of allophones of the Malayalam vowels ഇ/i/, അ /a/, and ഉ /u/ for female speaker

The mean and standard deviation of first and second formant frequencies of each vowel phones are computed using the TEMU dataset Table 3.6 shows the spectral statistics including the average formant frequencies F1 and F2 computed for each allophones and its mean and standard deviation obtained from 10 Malayalam vowels. Considering all the vowels, average F1 and F2 values obtained for the male speakers are 415.62 Hz and 1642.06 Hz respectively and that of female speakers are 557.45 Hz and 1859.42 Hz respectively. It is also observed that for the male speakers, the range of F1 obtained is from 251.60 Hz to 721.46 Hz and the range for F2 is from 624.80 Hz to 2453.28 Hz. The range of F1 obtained for the female speakers is from 291.30 Hz to 1060.96 Hz and that of F2 is from 731.71 Hz to 2945.71 Hz.

Table 3.6 Spectral statistics of Malayalam vowel allophones based on F1 and F2

Sl.No	Vowel	Allophones	Average formants			
			F1 (Hz)		F2 (Hz)	
			Male	Female	Male	Female
1	ഇ[i]	[i]	336.61	367.27	2094.12	2416.08
		[^{y}i]	288.12	338.66	2221.14	2590.26
		[y^{i}]	282.40	340.88	2293.25	2741.87
		Mean	305.05	350.45	2195.23	2571.74
		SD	33.14	36.45	127.11	210.49
2	ഈ[i:]	[^{y}i:]	282.78	341.06	2288.09	2608.66
		[i:]	316.45	379.53	2256.12	2651.03
		Mean	299.61	360.29	2272.10	2629.84
		SD	29.36	31.04	85.16	122.81
3	എ[e]	[^{y}e]	372.10	511.95	2108.67	2496.38
		[E]	423.11	566.94	1968.79	2226.60

Continue on the next page

Table 3.6 Spectral statistics of Malayalam vowel allophones based on F1 and F2 (cont.).

Sl.No	Vowel	Allophone	Average formants			
			F1(Hz)		F2(Hz)	
			Male	Female	Male	Female
		Mean	397.60	539.44	2038.73	2361.49
		SD	46.70	52.80	151.84	174.11
4	ഏ[e:]	[^{y}e:]	368.92	534.82	2167.02	2444.70
		[e^{r}:]	357.75	559.85	2295.09	2322.66
		[e:]	390.33	550.18	2181.84	2397.72
		Mean	373.91	548.97	2213.93	2386.86
		SD	28.05	54.34	85.24	183.321
5	അ[a]	[∧]	649.19	960.24	1438.16	1636.41
		[A]	472.56	687.03	1710.52	2002.61
		Mean	560.87	823.63	1574.34	1819.51
		SD	102.99	155.31	204.10	252.20
6	ആ[a:]	[a:]	616.44	976.50	1396.32	1506.93
		[a]	640.13	987.34	1266.34	1474.89
		Mean	626.97	981.32	1338.55	1492.69
		SD	48.60	53.74	145.29	77.103
7	ഉ[u]	[^{w}u]	375.73	405.41	887.91	975.19
		[u^{w}]	348.31	446.86	1039.61	1179.78
		[ɯ]	365.06	508.20	1189.76	1367.61
		[ə]	483.41	712.61	1321.23	1629.12
		[ə*]	495.69	719.61	1393.10	1769.99
		[ɯ^{v}]	445.64	527.15	1481.77	1704.11
		[U]	427.52	432.02	1323.50	1183.08
		Mean	417.57	531.85	1232.55	1413.21

Continue on the next page

Table 3.6 Spectral statistics of Malayalam vowel allophones based on F1 and F2 (cont.).

Sl.No	Vowel	Allophone	Average formants			
			F1(Hz)		F2(Hz)	
			Male	Female	Male	Female
		SD	69.08	128.39	312.620	334.88
8	ഊ /u:/	[^{w}u:]	380.84	378.48	961.63	863.45
		[u]	374.92	431.73	724.13	834.61
		Mean	378.48	399.78	866.63	851.92
		SD	17.07	37.27	140.44	104.79
9	ഒ/o/	[^{w}O]	430.02	559.49	1107.66	1221.04
		[O]	411.79	532.99	1204.7	1290.43
		Mean	420.91	546.24	1156.18	1255.74
		SD	43.03	59.32	206.82	121.55
10	ഓ/o:/	[^{w}O:]	431.61	550.4	782.06	1016.43
		[O]	451.04	606.09	1254.96	1087.03
		Mean	440.76	576.6	1004.6	1049.65
		SD	57.74	66.33	433.4	62.34

3.6 Clustering of Vowel Allophones using K-means clustering

A detailed study on the durational properties and formant frequencies of Malayalam vowel allophones are presented in the previous sections. From the study it is observed that the average duration of the allophonic variations of each vowel varies significantly (refer Figure 3.3). It is also observed that the formant frequencies are capable of classifying five Malayalam short vowels, but on the same time it fails to distinguish

between short and long vowels. In this section the credibility of the combined features including average duration, F1 and F2 in vowel classification are verified using k-means clustering technique. For this purpose the proposed features are extracted from the allophonic examples of the vowel phones taken from the TEMU dataset. The clustering experiments are then conducted based on the combined feature vectors extracted from the 238 allophonic examples of ten Malayalam vowels.

K-means is an unsupervised algorithm that clusters the attributes or features into the k number of groups [172, 173]. The clustering is performed based on the idea of minimizing the distance (usually Euclidian distance) between the feature vectors representing the data samples and the corresponding cluster centroids. Initially the cluster centroids are assigned randomly and these centroids are updated in each iteration. In this work, the clustering experiments are conducted on the combined feature vectors (average duration, F1 and F2) which are extracted from the 238 allophonic examples. The cluster analysis is performed on ten Malayalam vowels ഇ [i], ഈ [i:], എ [e], ഏ [e:], അ [a], ആ [a:], ഉ [u], ഊ [u:], ഒ [o] and ഓ [o:] to understand their grouping pattern. Implementation of the k-means clustering is carried out using the process detailed in Agorithm1.

Algorithm 1 k- means clustering

Input: Vowel allophone feature vectors $(x_1, x_2, \ldots x_n)$ representing average duration, L1, L2, assign k=10 corresponding to the number of vowel classes.

Output: Vowel Allophone instances partitioned in to 10 clusters

1: Initialize 10 Cluster Centroids $(m_1, m_2, \ldots m_{10})$ by random selection from the input feature set
2: Partition the input data points in to 10 clusters by assigning each data instance x_j to the closest cluster Centroid mi using the Euclidian distance measure
3: Re-estimate each m_i as $m_i = \frac{\sum_i^{N_k} X_i}{N_k}$, where N_k is the number of current instances in i$^{\text{th}}$ cluster
4: Repeat the steps 2 to 4 until cluster centroids no longer change significantly

The result of the clustering performed on the combined feature vector using K means clustering technique is depicted in figure 3.10. The '*' indicates the sample which are native to the cluster and '□' indicates the misclassified samples.The recognition accuracy of each vowel is computed based on the clustering results with the help of majority voting scheme. The average recognition accuracy obtained for the ten Malayalam vowels is 65.6388%. From the experiment results it is evident that the combined durational and spectral features of the vowel allophones can be effectively used to support the classification of Malayalam vowel phones.

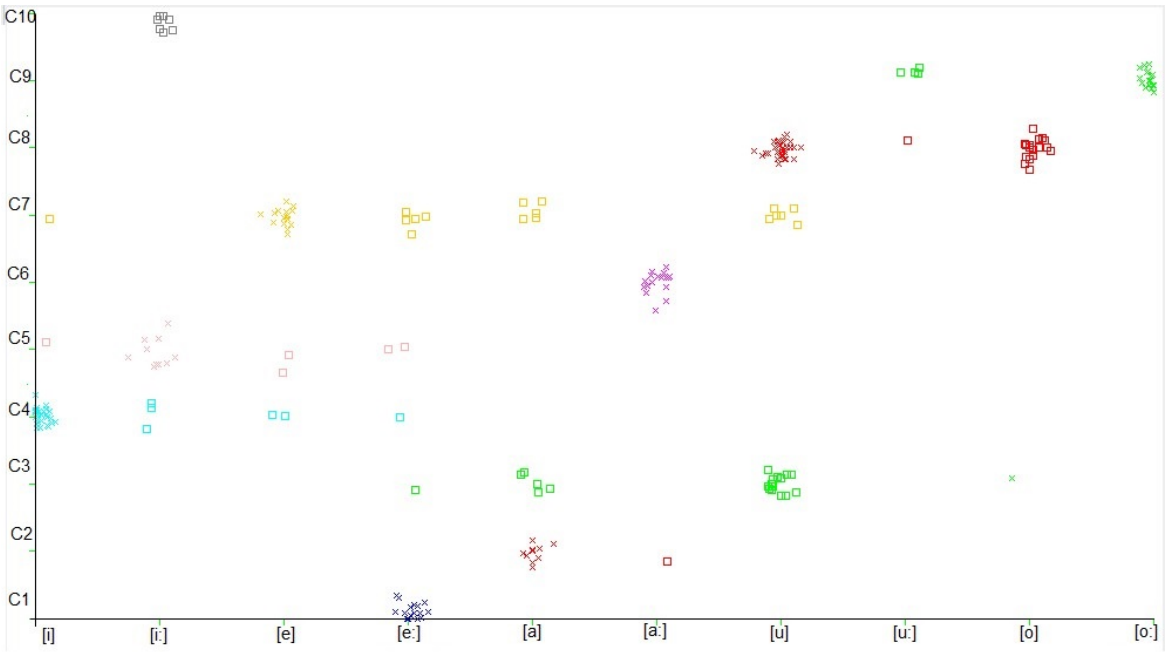


Fig. 3.10 Clustering of Malayalam vowels based on average duration, F1 and F2 using K-means clustering techniques

3.7 Conclusion

Every phone in any spoken language is pronounced as one of its allophone. For the very reason, the properties of allophonic variations of each phone are very vital in continuous speech recognition and speech synthesis studies. In this chapter, the phonological structure of Malayalam language is discussed in detail. Malayalam vowel allophones are identified, classified and analyzed based on their durational and spectral properties. For the experimental studies the TEMU Malayalam phonetic archive which consists of 238 selected words to interpret the allophonic variability of Malayalam phoneme is used. The statistical properties of duration, first and second formant frequencies of the vowel allophones are analyzed thoroughly. A clustering analysis is also performed to verify the significance of the duration and spectral properties of the Malay-

alam vowels and its allophonic variations. The contextual variations in phonemes are found to be encoded in the linguistic understanding of the allophones. From the experimental results, it is evident that the durational and spectral features computed from the vowel allophones can be effectively used to improve the performance of the Malayalam ASR and speech synthesis systems designed for continuous speech. This work can be considered as a first step towards a paradigm shift to allophone based Malayalam speech processing. As a future work, duration and spectral properties of consonant and diphthong allophones have to be explored in detail.

Acknowledgement

We would like to thank The Centre for Malayalam Language Technology, Thunjath Ezhuthachan Malayalam University, Kerala, India for providing necessary support to work with the TEMU Malayalam phonetic archive project and for granting permission to use same dataset for experimental purposes.

Chapter 4

A Comprehensive Grapheme-to-Phoneme Transcription Algorithm for Malayalam with Application to Speech Processing

4.1 Introduction

The idea of interaction between computers and human in natural languages has reached the realm of reality. Language specific pre-processing applications are essential for developing language computing tools to support speech processing applications. Grapheme to Phoneme (G2P) converter is one of the most vital computational tools and has been used in many important speech recognition tasks including keyword spotting, spoken document retrieval, speech synthesis *etc.* Grapheme-to-Phoneme transcription is the act of converting text into a sequence of phonemes [174]. Many implementation of G2P convertors for various languages using rule based method, dictionary lookup approach and data driven method have been reported [175] [176]. Dictionary based methods use

an electronic dictionary with phonetic transcription entries for each word. This approach is flexible but the enormous amount of manual labour required to maintain a large vocabulary dictionary makes it less attractive. Framing linguistic rules for G2P conversion is the underlying approach in rule based systems. It is more suitable for phonetically perfect languages such as Sanskrit where written text and pronunciation are directly connected [176]. Data driven approaches use Hidden Markov Model (HMM), Artificial Neural Network (ANN) or some other statistical machine learning techniques to learn the underlying patterns in the dataset. Multilingual systems mainly use data driven approaches due to the complexity in processing such systems. Many implementation of successful G2P conversion tools make use of the combination of above mentioned three approaches.

In this work, the possibility of implementing rule based G2P convertor for Malayalam is investigated to obtain the necessary phonetic transcription which in turn can be used further in speech analytic applications. The set of Malayalam graphemes is divided into six sub sets and distinct rule based processing routines are employed for each to support transcription. The proposed system is implemented using Python based framework. The phoneme statistics derived out of the day to day use of language is considered as a salient factor in designing language processing tools. Finally, as part of performance evaluation of the system, the phoneme frequencies based on the word corpus and the sentence corpus are estimated separately using the proposed G2P conversion tool. The rest of the paper is arranged as follows. Section 4.2 describes Malayalam grapheme categorization. Session 4.3 discusses rule based Malayalam grapheme to phoneme transcription algorithm. Session 4.4 describes the frequency analysis of Malayalam phones based

on the proposed G2P transcription tool and section 4.5 concludes the work.

4.2 Malayalam Orthography and Categorisation of Grapheme Units

Malayalam script is a Brahmic script used commonly to write the Malayalam language. Like many other Indic scripts, it is alphasyllabary and the writing system is partially alphabetic and partially syllable-based [177]. This section briefly describes the categorization of Malayalam graphemes based on its orthography and its general phonetic values which can be applied to the frame the rules to build the transcription tool. The following section presents a detailed description of Malayalam orthography comprising of vowels, diphthongs, consonants, special characters and compound letters.

4.2.1 Malayalam Vowels

There are 6 short vowels and 5 long vowels in Malayalam. In general, an independent vowel letter usually occurs as the first letter of a word that begins with a vowel. The dependent vowel modifies the consonant symbols to form valid Consonant–Vowel (CV) units. Table 4.1 shows the list of independent vowel graphemes and the corresponding dependent vowel signs (diacritics) of the Malayalam script.

Table 4.1 List of Malayalam vowels and dependent vowel signs

Vowel list		Independent vowel (Grapheme)	Dependent vowel (Grapheme)	
			Sign	Example
Short	a	അ <a>	-	ക /ka/
	i	ഇ <i>	ി	കി /ki/
	u	ഉ <u>	ു	കു /ku/
	e	എ <e>	െ	കെ /ke/
	o	ഒ <o>	ൊ	കൊ /ko/
	ṛ	ഋ <ṛ>	ൃ	കൃ /kṛ/
Long	a:	ആ <a:>	ാ	കാ /ka:/
	i:	ഈ <i:>	ീ	കീ /ki:/
	u:	ഊ <u:>	ൂ	കൂ /ku:/
	e:	ഏ <e:>	േ	കേ /ke:/
	o:	ഓ <o:>	ോ	കോ /ko:/

From the table 4.1, it can be observed that,

I The dependent vowel signs of <e> and <e:> are placed to the left of a consonant letter.

II The vowel signs <i>, <a:>, <i:> are placed to the right of a consonant letter to which it is attached.

III The vowel signs <o> and <o:> consist of two parts: the first part goes to the left of a consonant letter and the second part goes to the right of it.

IV In the reformed orthography, the vowel signs <u>, <u:>, <ṛ> are simply placed to the right of the consonant letter.

The consonant–vowel unit formed by combining dependent vowel with consonant creates valid syllable unit in Malayalam.

4.2.2 Malayalam Diphthongs

Diphthongs are the combination of two adjacent vowels within the same syllable. Malayalam has diphthongs as a separate category that is phonologically distinct from monophthongs [178]. The two distinguished diphthongal articulations in Malayalam are ഐ <ai>and ഔ <au>[177]. The diphthong /ai/ is named as *falling diphthong* as it starts with a central vowel and ends in a semivowel with less prominence. The diphthong /au/ is a *closing diphthong* as it starts with an open element and ends in a more close element and it usually occurs in Sanskrit loans only. Table 4.2 shows Malayalam diphthongs and its corresponding signs.

Table 4.2 Malayalam diphthongs and its signs

Diphthong	Independent	Dependent signs	
		Sign	Example
ai	ഐ <ai>	ൈ	കൈ /kai/
au	ഔ <au>	ൗ	കൗ /kau/

4.2.3 *Anusvaram* and *Chandrakkala*

An *anusvaram*, denotes the nasalization where the preceding vowel was changed into a nasalized vowel. In Malayalam, *anusaram* is represented as അം /am/ and is treated as a sort of vowel sign. It simply represents a consonant /m/ after a vowel, though this /m/ may be assimilated to another nasal consonant. *Chandrakala* (്/ə/) is a diacritic restricted to the final position of a word. It is attached to a consonant letter which is not followed by an inherent vowel. It is treated as a semi vowel that is an allophone of the vowel /u/. Table 4.3 lists the signs and the usage of special graphemes *Anusvaram* and *Chandrakala* with examples.

Table 4.3 Sign and usage of *anusvaram* and *chandrakala*

Special Graphemes	Independent	Dependent	
		Sign	Example
Anusvaram	അം <am>	ം	കം /kam/
Chandrakkala	-	് /ə/	ക് /kə/

4.2.4 Malayalam Consonant Classes

Malayalam consonants are grouped based on their articulation classes. Malayalam has 5 sets of *varga* consonants and 11 other consonants. The *varga* set comprises of velar, palatal, retroflex, dental, labial and holds 5 elements in each set. The *varga* set is again classified as voiceless and voiced, which are then subdivided into aspirated, unaspirated, and nasal based on its phonetic property. Malayalam varga consonants and its classification are given in table 4.4. List of Malayalam consonants other than *varga* are given in table 4.5.

Table 4.4 Malayalam *varga* consonant classification

Consonant classes	Voiceless		Voiced		
	Unaspirated	Aspirated	Unaspirated	Aspirated	Nasal
Velar	ക <ka>	ഖ <kha>	ഗ <ga>	ഘ <gha>	ങ <a>
Palatal	ച <ca>	ഛ <cha>	ജ <ja>	ഝ <jha>	ഞ <na>
Retroflex	ട <ṭa>	ഠ <ṭha>	ഡ <ḍa>	ഢ <ḍha>	ണ <a>
Dental	ത <ta>	ഥ <tha>	ദ <da>	ധ <dha>	ന <na>
Labial	പ <pa>	ഫ <pha>	ബ <ba>	ഭ <bha>	മ <ma>

Table 4.5 List of Malayalam consonants other than *varga*

Sl.No.	1	2	3	4	5	6	7	8	9	10	11
Consonant	യ <ya>	ര <ra>	ല <la>	വ <va>	ശ <śa>	ഷ <ṣa>	സ <sa>	ഹ <ha>	ള <ḷa>	ഴ <l̲a>	റ <r̲a>

4.2.5 Formation of Malayalam Compound Letters

Compound letters are formed by the combinations of more than one consonants in Malayalam. They are usually formed either by doubling up of the same consonant or by combining different consonants. All the consonant combinations are not valid compounds in Malayalam. For example aspirated consonants usually do not double in Malayalam. The formation of compound letters in Malayalam is broadly classified based on the following two factors.

1. Position and order of the component letters
2. Presence of the special symbols

In the first category, based on the position and order of component letters, 7 rules are framed for the formation of compound letters in Malayalam as listed in table 4.6.

Table 4.6 Rule set for the formation of compound letters based on the position and order of component letters

Sl.No.	Rules	Example
1	Component letters written completely on side by side.	ത്ത (ത + ത <t>+ <t>), മ്മ (മ + മ <m>+ <ma>) *etc.*
2	Component letters written completely one top of the other.	പ്പ (പ + പ <pa>+ <pa>), ഷ്ട (ഷ + ട <sha>+ <ta>) *etc.*
3	First letter written completely on left side and second component partially attached to it.	ന്ന (ന + ന <na>+ <na>), ഞ്ച (ഞ + ച <ɲa>+<ca>) *etc.*

Continue on the next page

Table 4.6 Rule set for the formation of compound letters based on the position and order of component letters (cont.).

Sl.No.	Rules	Example
4	First letter partially on left side and the second one completely on right side.	ണ്ട (ണ + ട <ṭa>+ <ṭa>), ന്ത (ന + ത <na>+ <ta>) *etc.*
5	Second component written below to the first one.	ട്ട (ട + ട <ṭa>+ <ṭa>), ദ്ദ (ദ + ദ <da>+ <da>) *etc.*
6	Component letters written opposite to their pronunciation order.	The compound letter ഹ്മ , ഹ്ന are written like the combinations ഹ + മ (<ha>+ <ma>) and ഹ + ന (<ha>+ <na>) but it is pronounced as മ + ഹ (<ma>+ <ha>), ന + ഹ (<na>+ <ha>) respectively.
7	Compound letter as a component. That is more than two components joined together to form a compound letter.	ക്ത്ര (ക + ത + ര <ka>+ <ta>+<ṛa>), ദ്ധ്യ (ദ + ധ + യ <da>+ <dha>+ <ya>) *etc.*

The second category of compound letters are formed in 3 distinct ways as detailed below.

1. Special symbols are used to represent approximant consonants (യ <ya>, ര <ra>, ല <la>, വ <va>) in compound letters. These symbols are usually called as consonant diacritics. Compound letters

formed by combining different approximants with the consonant ക /k/ are listed in table 4.7.

2. To show doubling of some letters, the symbol '് ' is used with the consonant component. Example: ച്ച (ച <ca>+ ച <ca>), വ്വ (വ <va>+ വ <va>) *etc.*

3. Compound letters formed by arbitery joining of the component letters. Example: ക്ക (ക <ka> + ക <ka>), ങ്ക (ങ <ŋ> + ക <ka>), മ്പ (മ <ma>+ പ <pa>) *etc.*

Table 4.7 Formation of compound letters by combining 4 different approximants with the Malayalam consonant ക <ka>

Sl.No.	Consonants	Compound unit
1	ക + യ <ka>+ <ya>	ക്യ <kya>
2	ക + ര <ka>+ <ra>	ക്ര <kra>
4	ക + ല <ka>+ <la>	ക്ല <kla>
5	ക + വ <ka>+ <va>	ക്വ <kva>

In early times there were more than 500 character representations in Malayalam. In the year 1971 the Government of Kerala constituted a committee to conduct a detailed study on Malayalam script with an objective of encoding Malayalam character and to frame a standard keyboard layout [179]. As a result, scientific restructuring of graphemes were proposed by which the size of the character set has reduced to 80.

4.2.6 *Chillukal*

A *chillu*, (*chillaksharam*), represents a pure consonant that exists independently without the help of a *virama*. *Chillu* is considered as special consonant letter which is never followed by an inherent vowel. In UNICODE representation *chillu* letters are treated as independent characters

and encoded automatically [3]. There are five *chillu* letters in Malayalam which are lsited in table 4.8.

Table 4.8 *Chillu* letters in Malayalam

Sl.No.	*chillu*	Base letter
1	ൺ	ണ <ṇa >
2	ൻ	ന <na>
3	ർ	ര <ra>
4	ൽ	ല <la>
5	ൾ	ള <ḷa >

The following section describes the algorithm and implementation strategies proposed for developing a comprehensive rule based automatic G2P transcriptor for Malayalam in detail.

4.3 A Complete Rule based Automatic G2P Transcriptor for Malayalam

An inclusive rule based algorithm is designed and implemented based on the properties of Malayalam graphemes discussed in section 4.2. In this section, the pre-processing steps and the implementation details of the proposed G2P transcriptior are discussed. A complete rule based algorithm is designed for converting Malayalam graphemes units to phonemes. The following section describes the the pre-processing steps followed by the implimentation of the grapheme to phoneme conversion algorithm.

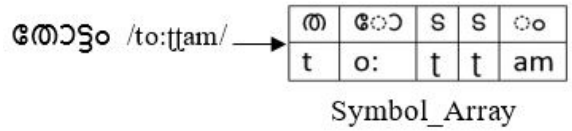
4.3.1 Pre-Processing Stages

Language specific pre-processing steps are essential for the implementation of the proposed rule based G2P conversion algorithm. Malayalam

isolated word texts are given as the input to the proposed transcription system. Initially as a pre-processing, the input array of characters are converted into a *Symbol_Array* of graphemes based on the following rules;

1. Vowel and consonant symbols are considered as independent characters.
2. Component consonants are converted into individual consonants
3. ഩ /r̲r̲a/ is considered as a consonant compound, even though it is actually known as an individual character.
4. All vowel and consonant symbols are stored to the right of the effected consonant in the *Symbol_Array*.

The characters in the input word are entered into the ***Symbol_ Array*** by applying all possible preprocessing rules mentioned above. The 4 rules described above are considered while performing this pre-processing step. The simulation of the pre-processing steps on a sample Malayalam word തോട്ടം /to:ṯṯam/ is given below.



The detailed description of the algorithm used for implementing Malayalam G2P conversion tool is given in the following section.

4.3.2 Implementation of the Proposed Malayalam G2P Transcription Algorithm

A complete rule based G2P conversion algorithm is proposed for Malayalam. Implementation of the algorithm is described in such a way that,

the transcription of the various grapheme classes are carried out based on separate subroutines. Hence, the grapheme classes of plosives, long vowels and vowel symbols, short vowels and *Chillu* are transcripted using respective conversion rule set. Since the graphemes ര <ra>, ം <am> and ന <na> possess some special characteristics, separate subroutines are implemented to support transcription. The special cases that are to be considered while implementing G2P conversion algorithm for Malayalam are discussed below.

I The character ര <ra> – is considered as an independent consonant unit. It also comes as a component of the grapheme ട്ര <tta>. In the case of grapheme ന്റ <nte>, it has to be subscripted as ര <ra> placed under ന് <n>. Considering as a special case this transcription is symbolically represented with a separate phoneme denoted as **φ**.

II The grapheme ം *Anuswaram* has a different transcription behaviour when it comes together with vowels other than consonants.

III The grapheme ന <na> can have two different transripted form in Malayalam. In some situations it is transcripted to dental ന /n̪/ and in other conditions it is transcripted to the alveolar ന /n/ and is represented in the algorithm as ന /n/ and ന1 /n̪/ respectively.

The pseudocode for the proposed rule based Malayalam G2P transcription algorithm is given below.

Algorithm 2 Automatic G2P conversion for Malayalam

Input: *Symbol_Array*, set of separated symbols corresponding to each input word obtained from the pre-processing stage

Output: *Phonemic_Out_Array*, Sequence of transcripted phoneme symbols.

while Not end of the *Symbol_Array* **do**
Read the next character from *Symbol_Array* to *Cur_Symbol*
if *Cur_Symbol* is equal to <ം> **then**
goto SUB_ROUTINE1 (Index of the *Cur_Symbol*)
else if *Cur_Symbol* is equal to /ം/ *Anuswaram* **then**
goto SUB_ROUTINE2
else if *Cur_Symbol* is equal to <ം> **then**
goto SUB_ROUTINE3
else if *Cur_Symbol* is in the list of *Plosives* [1] **then**
goto SUB_ROUTINE4
else if *Cur_Symbol* is in the list of Long Vowels OR Vowel Symbols **then**
Identify the corresponding symbol from the Look_Up_Table_1 and insert into *Phonemic_Out_Array*
else if *Cur_Symbol* is in the list of Short Vowels OR *Chillu* **then**
Insert *Cur_Symbol* to *Phonemic_Out_Array*
else
goto SUB_ROUTINE5
end if
end while

[1] List of plosives are given in section 1.2 of Chapter 3

The following subroutine is initiated when the *Cur_Symbol* is <ഠ>. Three distinct rules for the graphene <ഠ> based on their context is framed in SUB_ROUTINE1.

---

procedure SUB_ROUTINE1(Index of the *Cur_Symbol*)
if (*Symbol_Array*[Index of the *Cur_Symbol*-1] is <ൺ>) AND (*Symbol_Array*[Index of the *Cur_Symbol* + 1] is <ര്>) **then**
Insert /ഫ/ to *Phonemic_Out_Array*
else if (*Symbol_Array*[Index of the *Cur_Symbol* + 1] is <ു>) AND (*Symbol_Array*[Index of the *Cur_Symbol* + 2] is <ഠ>) **then**
Insert /ൺ / to *Phonemic_Out_Array* AND advance *Symbol_Array* by two positions
else
goto SUB_ROUTINE5
end if
end procedure

SUB_ROUTINE2 is initiated when *Cur_Symbol* is *Anuswaram*. Transcription decision of *Anuswaram* depends upon whether the left neighbourhood has a vowel or not.

---

procedure SUB_ROUTINE2(Index of the *Cur_Symbol*)
if *Symbol_Array*[Index of the *Cur_Symbol* - 1] is a vowel or vowel symbol **then**
Insert /ം/ to *Phonemic_Out_Array*
else
Insert /അ + ം/ to *Phonemic_Out_Array*
end if
end procedure

SUB_ROUTINE3 deals with the grapheme <ന>. In Malayalam, two phones textitviz. alveolar nasal and dental nasal, are represented by the same grapheme <ന>.

---

procedure SUB_ROUTINE3(Index of the *Cur_Symbol*)
if Index of the *Cur_Symbol* is in initial position **then**
if *Symbol_Array*[Index of the *Cur_Symbol* + 1] is <ഩ> **then**
then Insert /ന̆/[2] to *Phonemic_Out_Array*
else
Insert /ന̆1/[3] to *Phonemic_Out_Array*
end if
end if
if (*Symbol_Array*[Index of the *Cur_Symbol* - 1] is in ര <r>,ഠ / ഺ <ṛ>) OR (*Symbol_Array* [Index of the *Cur_Symbol* + 1] is in the list of /dental/[4]) **then**
Insert /ന̆1/ to *Phonemic_Out_Array*
else if (*Symbol_Array*[Index of the *Cur_Symbol* - 1] is in the list of vowels) AND (*Symbol_Array*[Index of the *Cur_Symbol* + 1] is <ന>) AND (*Symbol_Array*[Index of the *Cur_Symbol* + 2] is in the list of vowels) **then**
Insert two /ന̆1/ to *Phonemic_Out_Array* and advance *Symbol_Array* by one position
else
Insert /ന̆/ to *Phonemic_Out_Array*
end if
end procedure

---

[2]Alveolar Nasal
[3]Dental Nasal
[4]List of dental is ത /t/, ഥ /th/, ദ /d/, ധ /dh/, ന /n̪/

The following subroutine differentiates aspirated plosive consonants from unaspirated classes.

---

procedure SUB_ROUTINE4(Index of the *Cur_Symbol*)
  if Symbol is an aspirated plosive **then**
    Insert corresponding entry from Look_Up_Table_2 in to *Phonemic_Out_Array*
  else if symbol is an unaspirated plosive **then**
    goto SUB_ROUTINE5
  end if
end procedure

Consonants other than the graphemes discussed in the above mentioned subroutines are transcripted using the SUB_ROUTINE5.

---

procedure SUB_ROUTINE5(Index of the *Cur_Symbol*)
  if *Symbol_Array*[Index of the *Cur_Symbol* + 1] is a /vowel symbol/ **then**
    Insert (symbol + /◌്/ /*chandrakala*) / to *Phonemic_Out_Array*
  else
    Insert (symbol + /◌്/ + /അ/) to *Phonemic_Out_Array*
  end if
end procedure

The lookup table for short vowel signs, long vowels and long vowel signs are shown in table 4.9. The algorithm also performs the G2P conversion of plosive aspirated with the help of a separate lookup table. The look up table used in the algorithm for plosive aspirated with their phoneme representation is shown in table 4.10.

Table 4.9 Lookup table-1 for vowel signs and long vowels

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

Table 4.10 Lookup table-2 for plosive aspirated

Sl.No:	Grapheme	Phoneme
1	ഫ് /p^{h}/	പ് ഹ് /P h/
2	ഥ് /t^{h}/	ത് ഹ് /t h/
3	ഠ് /ʈ^{h}/	ട് ഹ് /ʈ h/
4	ഛ് /c^{h}/	ച് ഹ് /t h/
5	ഖ് /k^{h}/	ക് ഹ് /k h/
6	ഭ് /b^{h}/	ബ് ഹ് /b h/
7	ധ് /d^{h}/	ദ് ഹ് /d h/
8	ഢ് /ɖ^{h}/	ഡ് ഹ് /ɖ h/
9	ഝ് /ɟ ^{h}/	ജ് ഹ് /ɟ h/
10	ഘ് /g^{h}/	ഗ് ഹ് /g h/

After converting the input text to phoneme set, the IPA (International Phonetic Alphabet) symbols are used to represent it to enable for the

processing. This following section describes the mapping process in detail.

4.3.3 Malayalam Phoneme to IPA Mapping

IPA (International Phonetic Alphabet) is used to represent the set of sounds in a spoken languages. It tries to accommodate phonemes of all important languages in the world. IPA clarifies the pronunciation issues of a language to a universal audience. The conversion of Malayalam phonemes obtained as an output of the G2P convertor to its corresponding IPA symbol is performed by applying one to one mapping with the help of the lookup table given in table 4.11.

Table 4.11 Representation of Malayalam Phonemes

Sl.No:	Phone	IPA	Sl.No:	Phone	IPA
1	ഇ	i	2	ഉ	u
3	ഈ	i:	4	ഊ	u:
5	എ	e	6	ഒ	o
7	ഏ	e:	8	ഓ	o:
9	അ	a	10	ആ	a:
11	്	ə	12	ഐ	ai
13	ഔ	au	14	പ്	p
15	ത്	t	16	ഫ	t̲
17	ട്	ʈ	18	ച്	c
19	ക്	k	20	ഫ്	p^h
21	ഥ്	t^h	22	ഠ്	$ʈ^h$
23	ഛ്	c^h	24	ഖ്	k^h
25	ബ്	b	26	ദ്	d
27	ഡ്	ɖ	28	ജ്	ɟ

Continue on the next page

Table 4.11 Representation of Malayalam Phonemes(cont.).

Sl.No:	Phone	IPA	Sl.No:	Phone	IPA
29	ഗ്	g	30	ഭ്	b^h
31	ധ്	dh	32	ഢ്	$ɖ^h$
33	ഝ്	$ɟ^h$	34	ഘ്	g^h
35	മ്	m	36	ന്	n
37	ന്1	n̪	38	ൻ	n
39	ണ്/ൺ	ɖ	40	ഞ്	ɲ
41	ങ്	ŋ	42	സ്	s
43	ഷ്	ʂ	44	ശ്	ʃ
45	ഹ്	h	46	ര്	r
47	റ്/ർ	r̲	48	ഋ	ɻ̩
49	ല്/ൽ	l	50	ള്/ൾ	ɭ
51	ഴ്	ɻ	52	വ്	v
53	യ്	y	54	റ്റ്	r̲r̲

4.4 Statistical Analysis of Malayalam Phonemes

Statistical analysis of phonemes is a prime segment of general language modelling framework. Language modelling is a prerequisite for building applications including automatic speech recognition, automatic translation, speech synthesis, parsing and dictionary tools. The probabilistic distribution of phoneme n-gram sequence is used in sub word language modelling. Phoneme statistics information is widely used to improve the performance of language model based speech recognition systems [58, 180, 181]. Phonotactic observations about the permissible phoneme combinations in a language can also be derived from the phoneme level statistical information. Phonotactics based information is of great assis-

tance in designing language identification systems [182–184]. Phoneme probability understanding can also be used to improve the performance of spell checking systems, especially in phonemic languages such as Malayalam [185, 186]. This work can be considered as the first step towards the devolopment a phoneme n-gram language model for speech processing applications in Malayalam.

The Malayalam Grapheme to Phoneme Converter, described in section 3, is used to estimate the phoneme and diphone probabilities of Malayalam language. The analysis is performed using 50 Malayalam phones discussed in chapter 3 considering /n/ as a single phone. Another objective of performing statistical analysis is to evaluate performance the of proposed G2P transcriptor implemented as part of this study. A detailed demonstration of the word and sentence corpora used to perform the statistical analysis of Malayalam phonemes is given in the following section.

4.4.1 Malayalam Word and News Sentence Text Corpora

A growing crowd-sourced English-Malayalam electronics dictionary dataset Olam[5] is used as word corpus to perform statistical analysis. The Datuk Corpus consisting of 83,000 Malayalam words are extracted from Olam dataset.

A sufficiently large Malayalam News Sentence Text Corpus (MNSTC) is also created for this purpose. Initially, news text corpus is generated from the online news portals of the popular Malayalam dailies. These news text sentences are then classified into five categories including state news, national news, international news, sports news and news related to cultural importance. The first three categories *viz.* state news, national

[5]The Datuk Corpus: A free and open MAlayalam Dictionary dataset with over 106,000 definitions for more than 83,000 Malayalam words

news and international news represent news related to the current affairs excluding sports and culture categories. A total of 5250, news text sentences of various such categories are collected for dataset creation. Numerical proportion of the samples representing each category included in the dataset is shown in figure 4.1.

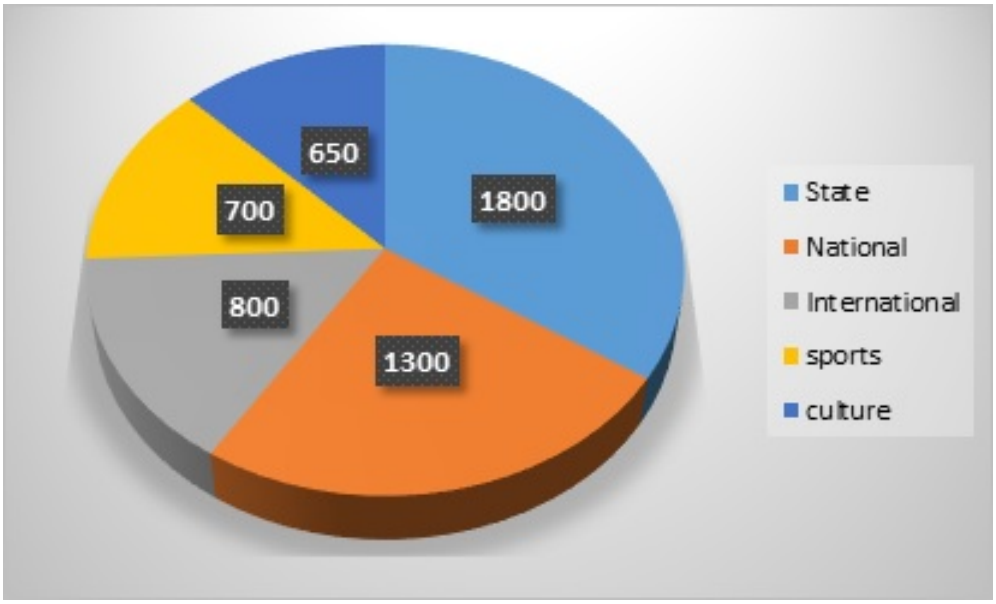


Fig. 4.1 Category wise proportion of news audio data

4.4.2 Phoneme Statistical Analysis Results

The general properties of word and the sentence corpora used in the work are given in table 4.12. Initially, the frequency of occurrence of each phone in the dataset is computed. Then the percentage of occurrence of each phone in the word and sentence corpus are computed separately. These phoneme probabilities obtained from both word and sentence corpora are given in table 4.13. The graphical representations of the percentage of occurrence of the most frequently occuring Malayalam phones present in word corpus and sentence corpus are shown in figure 4.2a and figure 4.2b respectively. From the experiment results, it can be

seen that the phone /a/ is the most frequently occuring phone in both the corpora with the frequency of occurrence 1,52,364 (word corpus), 56,332 (sentence corpus) and percentage of occurrence 19.96% (word corpus), 13.74% (sentence corpus)

Table 4.12 Details of word and sentence corpus

	Source	**Total number of sentences**	**Total number of words**	**Total number of phonemes**
Word Corpora	Olam	-	82,324	7,63,392
Sentence Corpora	MNSTC	5,250	28,943	3,56,808

Table 4.13 Malayalam phoneme statistics based word and sentence corpuses

Sl.No.	Phones	Word corpus		Sentence corpus	
		Frequency	Percentage	Frequency	Percentage
1	ഇ	40578	5.3154	27299	6.6603
2	ഉ	26746	3.5035	21806	5.3202
3	ഈ	6008	0.787	2053	0.5008
4	ഊ	4549	0.5958	1668	0.4069
5	എ	3267	0.4279	9665	2.358
6	ഒ	2046	0.268	1773	0.4325
7	ഏ	5951	0.7795	3889	0.9488
8	ഓ	6127	0.8026	3873	0.94492
9	അ	152364	19.958	56332	13.7438
10	ആ	34088	4.4653	18470	4.5062

Continue on the next page

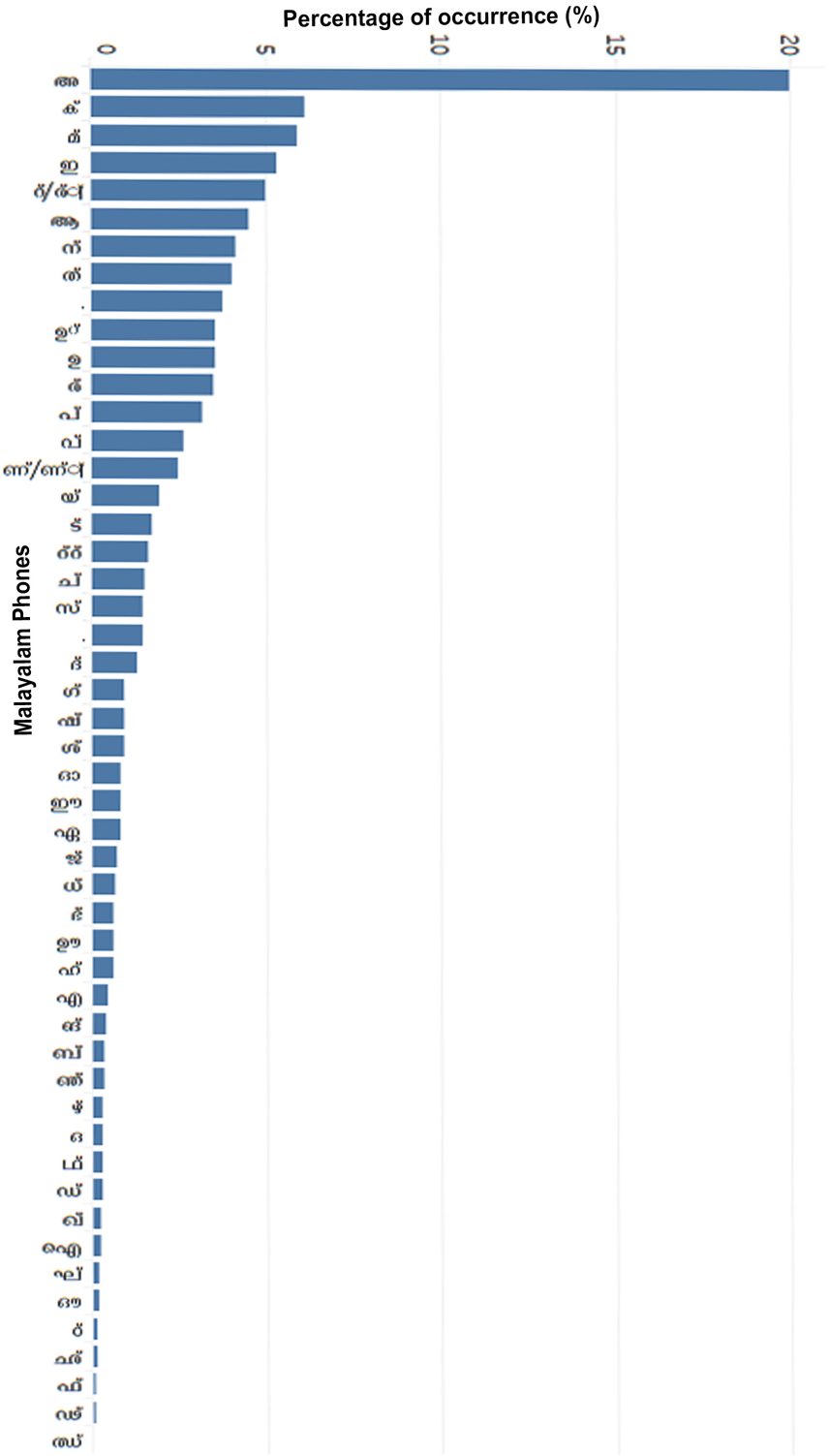
Table 4.13 Malayalam phoneme statistics based word and sentence corpuses(cont.).

Sl.No.	Phones	Word corpus		Sentence corpus	
		Frequency	Percentage	Frequency	Percentage
11	ഉ്	26748	3.5038	21824	5.3245
12	ഐ	1476	0.1933	499	0.1217
13	ഔ	1021	0.1337	222	0.0541
14	പ്	23895	3.1301	10982	2.6793
15	ത്	30493	3.9944	19354	4.7219
16	റ്	12338	1.6162	6675	1.6285
17	ട്	13077	1.713	11793	2.8772
18	ച്	11238	1.4721	5599	1.366
19	ക്	46509	6.0924	24486	5.974
20	ഫ്	531	0.0695	393	0.0958
21	ഥ്	2021	0.2647	805	0.1964
22	ഠ്	759	0.0994	116	0.0283
23	ഛ്	726	0.0951	55	0.0134
24	ഖ്	1505	0.1971	412	0.1005
25	ബ്	2615	0.3425	1014	0.2473
26	ദ്	9932	1.301	2215	0.5404
27	ഡ്	1910	0.2501	658	0.1605
28	ജ്	5386	0.7055	1390	0.3391
29	ഗ്	7140	0.9352	1610	0.3928
30	ഭ്	4554	0.5965	1290	0.3147
31	ഢ്	5061	0.6629	1771	0.432

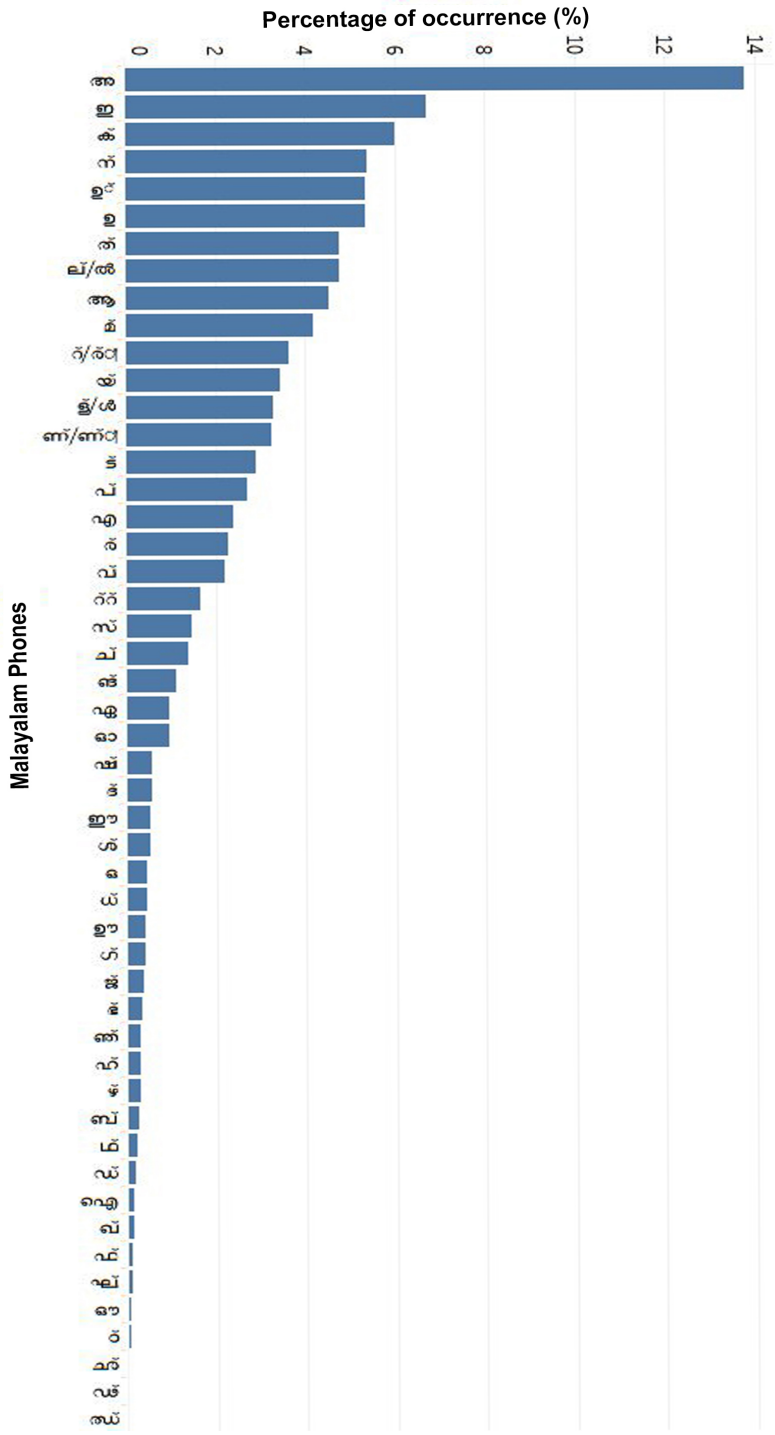
Continue on the next page

Table 4.13 Malayalam phoneme statistics based word and sentence corpuses(cont.).

Sl.No.	Phones	Word corpus		Sentence corpus	
		Frequency	Percentage	Frequency	Percentage
32	ഢ്	250	0.0327	28	0.0068
33	ഝ്	85	0.0111	9	0.0022
34	ഘ്	1271	0.1664	286	0.0697
35	മ്	45003	5.8951	16952	4.1359
36	ന്	31533	4.1306	21987	5.3643
37	ണ്/ൺ	18900	2.4757	13174	3.2141
38	ഞ്	2449	0.3208	1095	0.2671
39	ങ്	2850	0.3733	4497	1.0971
40	സ്	11081	1.4515	5908	1.4414
41	ഷ്	7062	0.925	2218	0.5411
42	ശ്	6935	0.9084	1997	0.4872
43	ഹ്	4509	0.5906	1087	0.2652
44	ര്	26550	3.4778	9179	2.2394
45	റ്/ർ	37872	4.961	14808	3.6128
46	ല്/ൽ	28454	3.7273	19325	4.7148
47	ള്/ൾ	11064	1.4493	13425	3.2754
48	ഴ്	2062	0.2701	1068	0.2605
49	വ്	20075	2.6297	8892	2.1694
50	യ്	14728	1.9292	13946	3.4025



(a) Phone probability obtained from word corpus (Olam)



(b) Phone probability obtained from sentence corpus (MNSTC)

Fig. 4.2 Phoneme level statistical analysis of Malayalam text corpus

A diphone is an adjacent pair of phones, which is usually referred as a recording of the transition between two phones. In many speech processing applications diphone is treated as an important unit of processing like speech synthesis. Hence understanding the probabilistic nature of these combinations is a vital stage in this direction. The relative frequency of the diphone computed based on the word corpus using the G2P conversion tool is given in Appendix 2. There are 50 phones in Malayalam (considering alveolar nasal /nന്/ and dental nasal / n̪ന്/ as single phone) and 2500 possible diphone combinations. These diphones are compared with the diphones occurring in the word corpus. It is found that 1243 diphone (49.72%) do not occur in the word corpus. The 25 most frequently occurring diphones in word corpus are given in Table 4.14. A similar list of 25 most frequently occurring diphones is constructed from sentence corpus and compared with those obtained from the word corpus. It is observed that 18 of such diphones are common in both the corpuses. The new seven diphone combinations generated out of sentence corpus are (ഉ, മ്) (ന്, ന്) (ഇ, ല് /ൽ) (ത്,ത്) (ആ,യ്) (അ, ള് /ൾ) (ഉ,ന്). Some combinations of diphones are very rare in both corpuses. Table 4.15 gives the number of diphones which occur very rarely (frequency of occurrence listed up to 5) in the investigated set of Malayalam word corpus.

Table 4.14 Relative frequencies of the most frequent 25 diphones obtained from Malayalam word corpus

Rank	Phone pairs	Relative frequency	Rank	Phone pairs	Relative frequency
1	(അ /a/, മ്/m/)	0.0533	14	(ക്/k/, ക്/k/)	0.0116
2	(ക്/k/, അ/a/)	0.0329	15	(ഉ/u/,ക്/k/)	0.0112

Continue on the next page

Table 4.14 Relative frequencies of the most frequent 25 phoneme pairs in Malayalam word corpus(cont.).

Rank	Phone pairs	Relative frequency	Rank	Phone pairs	Relative frequency
3	(അ/a/, ന്/n/)	0.0315	16	(അ/a/, ത്/t/)	0.0111
4	(അ/a/, ര്/r/)	0.0209	17	(ഇ/i/, ക്/k/)	0.0102
5	(ത്/t/, അ/a/)	0.0181	18	(അ/a/, വ്/v/)	0.0157
6	(ര്/r/, അ/a/)	0.0165	19	(ല്‍/l̩/,അ/a/)	0.0101
7	(അ/a/, ക്/k/)	0.0161	20	(അ/a/, ല്/ൽ/l/)	0.0099
8	(യ്/y/, അ/a/)	0.0157	21	(ക്/k/, ഉ /u/)	0.0088
9	(പ്/P/, അ/a/)	0.0152	22	(ആ/a:/, ര്/r/)	0.0086
10	(വ്/v/, (അ/a/)	0.0139	23	(ത്/t/, ഇ/i/)	0.0078
11	(ന്/n/, അ/a/)	0.0134	24	(ട്/(ṭ/, അ/a/)	0.0078
12	(മ്/m/, അ/a/)	0.0121	25	(അ/a/, പ്/P/)	0.0076
13	(ര്‍ /ṛ/, അ/a/)	0.0119			

Table 4.15 Number of infrequent diphones (with the fequency of occurence limited to five)

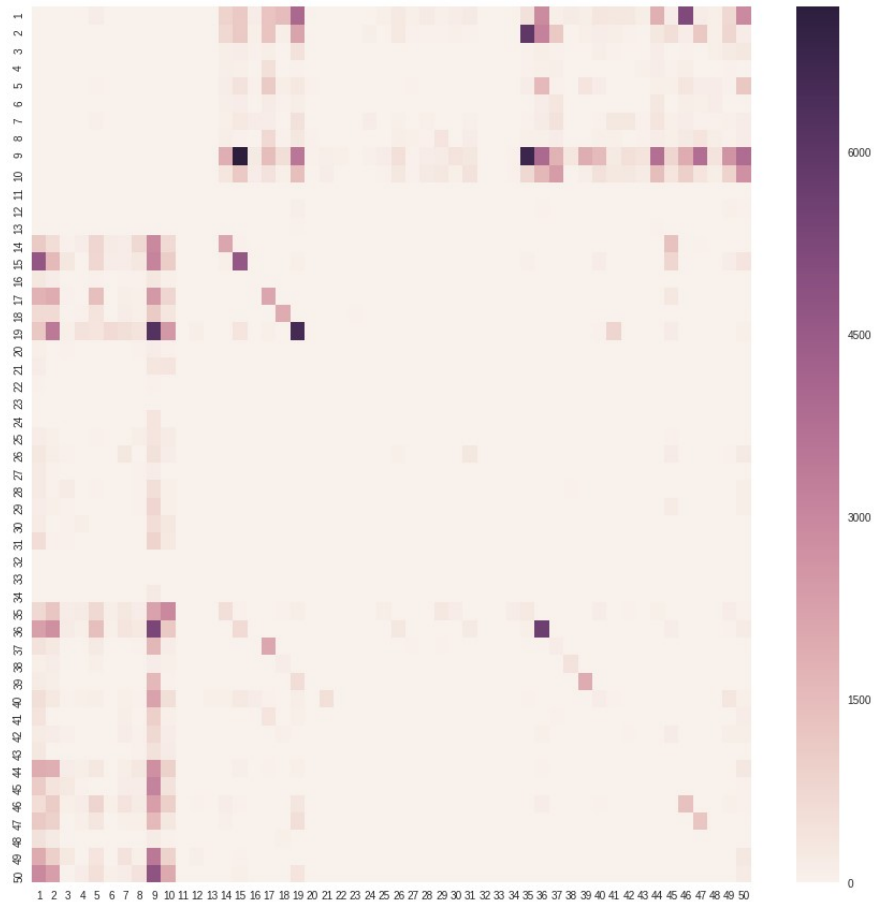
Frequency of occurence	Number of diphones	
	Word corpus (Olam)	Sentence corpus (MNSTC)
1	109	125
2	63	74
3	38	36
4	42	33
5	22	27

The result of phonotatic analysis which is performed with the help of G2P convertor, including relative frequency of occurrence of diphones obtained as part of the study is highly beneficial for the successful

implementation of various speech processing applications. The statistical analysis of phones and diphones presented in this study is significant as there is no such exhaustive study reported for Malayalam. The frequency of occurrences of diphones can be graphically represented with the help of a heat map. Figure 4.3a-4.3b illustrates the heat map representing the frequency of occurrence probability distribution of the diphone in Malayalam obtained from word and sentence corpora.

This investigation can also be extended to triphones and higher combinations of phones identify the most frequent and infrequent phone combinations in Malayalam which inturn can be used as an input for developing the language models that can support automatic speech recognition applications.

(a) Diphone probability - Word corpus (Olam)



(b) Diphone probability - Sentence corpus (MNSTC)

Fig. 4.3 Heat map of probability of occurence of diphones in Malayalam

4.5 Conclusion

Malayalam orthography comprising of vowels, diphthongs, consonants, special characters and compound letters are discussed in detail in this chapter. A comprehensive automatic rule-based Grapheme-to-Phoneme transcriptor for Malayalam is designed and implemented for the first time. In the implementation of the algorithm, the transcription of the grapheme classes are performed based on separate subroutines. The research findings derived out of the study can be effectively used in the

devolopment of automatic speech processing applications including ASR and text-to-speech converter in Malayalam. This can also be used as an effective tool for phoneme based statistical analysis in various language computing set-ups. The information derived out of this analysis can also be incorporated in the simulation experiments conducted using the language processing tool such as Sphinx, Kaldi *etc.* for better perfomance

As the next step, statistical analysis of individual phoneme as well as diphone occurrences in Malayalam is performed. The analysis is performed on both the word corpus (Olam) and an own developed news text sentence corpus (MNSTC). To the best of our knowledge this is the first fully automated G2P transcription based phoneme and diphone analysis conducted so far using a large dataset in Malayalam. We expect many advances in Malayalam speech processing will evolve using the preposed G2P convertor and the phoneme statistics derived out of the study.

Chapter 5

Implementation of Keyword Spotting in Malayalam Speech using Continuous Hidden Markov Modelling

5.1 Introduction

Generally, the news audio archives consist of a large number of recorded speech data. Implementation of audio mining, audio searching, forensic analysis and such speech analytic applications do not normally rely on the transcription of the entire speech. Only certain keywords of relevance are significant for building such applications. These keywords could either relate to a particular person or belong to certain topic of interest. The process of locating the occurrences of a given list of keywords W in a speech utterance O is entitled as keyword spotting (KWS). This process is also termed as spoken keyword spotting, or spoken query detection or Query by Example (QbE) or Spoken Term Detection (STD).

The most widely used speech recognition algorithms emerged in the past two decades are based on Hidden Markov Models (HMM) [187]. In

HMM statistical framework, a set of elementary probabilistic models of basic linguistic units (e.g., phonemes) is used to build speech representation. Speech modelling is a kind of representation of the speech signal to make some aspect explicit and hence to improve the efficiency and flexibility of speech processing applications. Both acoustic modelling and language modelling are important parts of modern statistical speech recognition algorithms. The proposed keyword spotting approach in Malayalam speech is implemented with standard keyword spotting strategies based on HMMs. The main emphasis of this work is on the speech modelling which is effectively used for speech recognition and keyword spotting based speech analytics.

In this chapter, a keyword spotting system and its implementation details are proposed for Malayalam news audio using continues Hidden Markov Modelling. The architecture of the proposed KWS system is shown in figure 5.1. Basically, speech signals are non-stationary in nature and dynamically change over time. But it can be viewed as a piecewise stationary signal or a short-time stationary signal which has some common acoustic properties for a short time interval (e.g., 10 ms). Hidden Markov Models (HMMs) are widely used statistical recognizer which considers speech as a piece-wise stationary signal [188] .

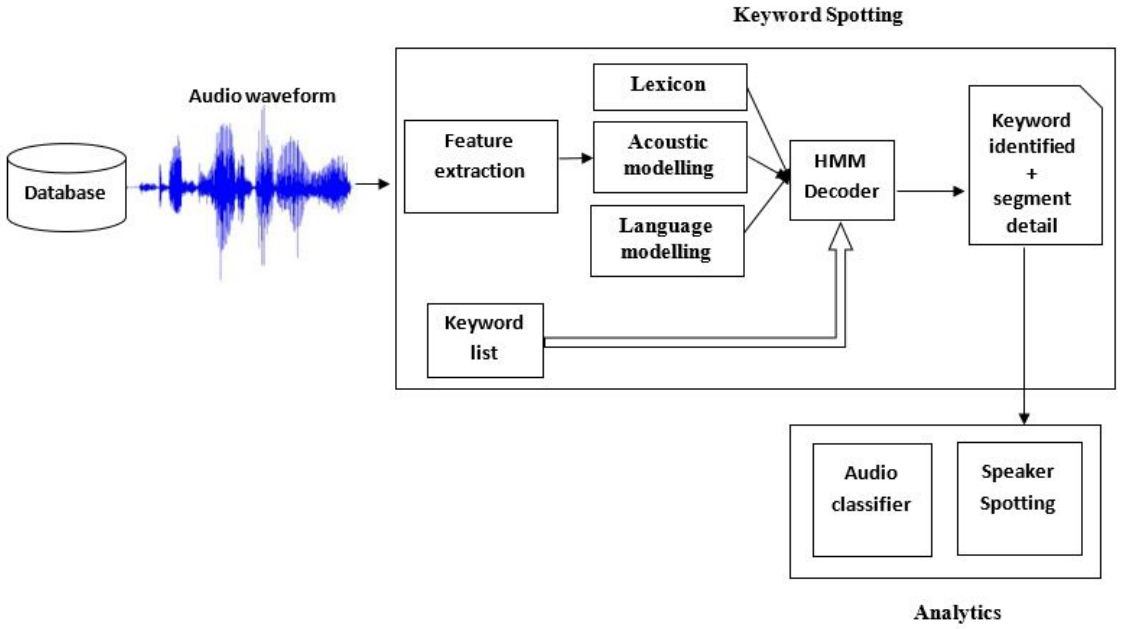


Fig. 5.1 The proposed KWS system architecture

HMM training algorithm creates statistical representation model of the speech signal. HMM decoder determines the best sequence of hidden states, the one that most likely predicts the spoken term. Preparation of knowledge base for HMM decoder is an important pre-processing step towards speech modelling and recognition. The knowledge base facilitate the construction and testing of statistical speech models. A Knowledge Base Preparation Tool for Malayalam (KBPT-M) that can be used for the implementation of HMM decoder is introduced as part of this work. This tool generates language models and related components required for acoustic modelling of the Malayalam language.

The HMM decoder recognizes the spoken keyword and present its segment details in the audio signals. The performance evaluation of the proposed system is carried out using two methods namely Exact Matching Method (EMM) and Relaxed Matching Method (RMM). In EMM, keywords are recognized only if the system encounters a precise

match. Whereas in relaxed method, inflected forms of the given keyword are also be considered for evaluation. The rest of this chapter is arranged as follows. Section 5.2 describes the dataset preparation for the implementation of HMM decoder. Section 5.3 discusses the process of knowledge base resource generation using the proposed KBPT-M. Section 5.4 describes the architecture of the HMM based ASR system. Section 5.5 presents the implementation details of the proposed keyword spotting system. Section 5.6 presents the experimental results and section 5.7 concludes the work.

5.2 Data Preparation

In this section, the dataset preparation procedure adapted for the implementation of the proposed keyword spotting system is discussed in detail. The dataset is prepared in such a way that information content is best exposed for the HMM decoder. The following sections discus the data preparation done for the conduct of training and performance evaluation of the HMM decoder implemented for keyword spotting from Malayalam continous speech.

5.2.1 Data Preparation for KWS System Training

In this section, various inter-related tasks performed in the data preparation phase are discussed. Speech dataset needs to be prepared separately for the purpose of training, testing and evaluating the models implemented as part of the proposed system.

The news text corpus (MNSTC) as described in Chapter 4 is converted to Malayalam News audio by recording the speech corresponding to text data, spoken by both male and female speaker in the normal

acoustic environment. This own developed Malayalam News Audio Corpus (MNAC) consists of news audio samples spoken by 35 speakers (both male and female) from different age group. The distribution of the number of speakers based on the age group contributed to the creation of this News audio corpus is shown in figure 5.2. Each speaker uttered 150 sentences taken randomly from any of the five news categories of the text corpus (MNSTC) created for this purpose. All these 5,250 spoken sentences are labelled with the News category id, sample id, speaker id and gender/age of the speaker that they belongs to. The average length of the news audio samples present in the dataset is 5.35 seconds. This news audio dataset is further used to generate the Acoustic Model and the HMM topology for the conduct of KWS experiments. The size of the dataset is comparable to the moderate size benchmark datasets designed for the development of speech processing applications in other languages.

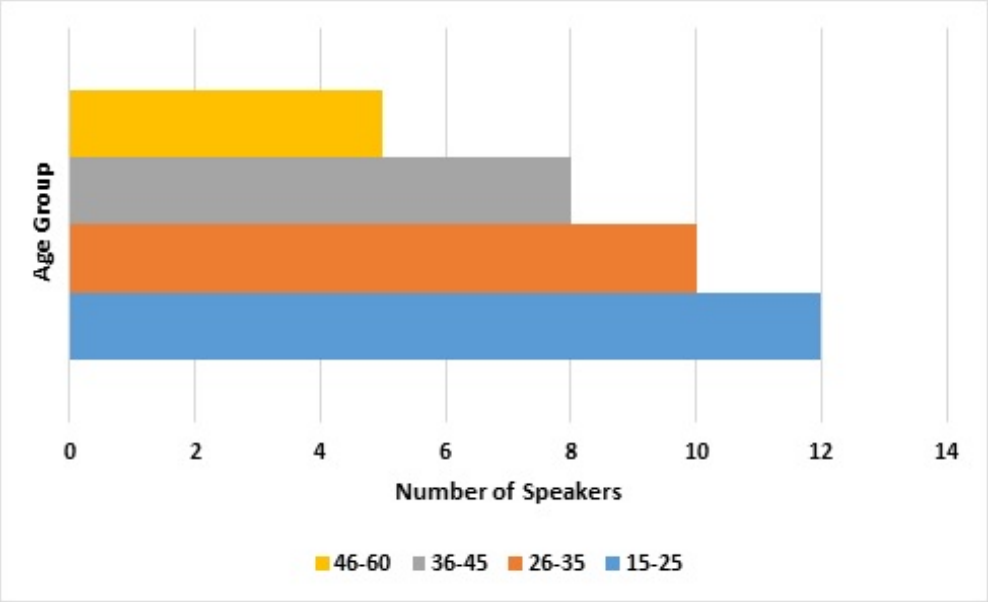


Fig. 5.2 Distribution of number of speakers, based on the age groups, contributed to the news audio dataset preparation.

5.2.2 Data Preparation for KWS System Evaluation

This section describes the preparation of dataset used for the performance evaluation of the proposed KWS system. A set of ninety one keyword texts selected from various news categories is created for the conduct of performance evaluation experiments. Inflected forms of the selected words are also included in the set. For example in Malayalam, the word കേരളം /ke:raḷam/ has different inflected forms like കേരളത്തിൽ /ke:raḷattil/ കേരളത്തെ /ke:raḷatte/ കേരളത്തിന്റെ /ke:raḷattin̲te/ *etc.* The selected keyword set includes names of persons, places, events, incidences, organization *etc.* which are also present in the news dataset MNSTC. The average length of words present in the keyword set is 5.6875. The statistical parameters including the number of characters in each keyword (Word length) and the number of occurence of the keywords in the news audio dataset are computed. Figure 5.3 shows the histogram of the keyword length, where the keyword count is plotted against the number of characters. Each column bar represents the total count of keywords with a particular number of characters in it. Figure 5.4 shows the histogram of the number of occurence of the keywords where the keyword count are plotted against the number of occurence of the keyword present in the MNAC news audio corpus.

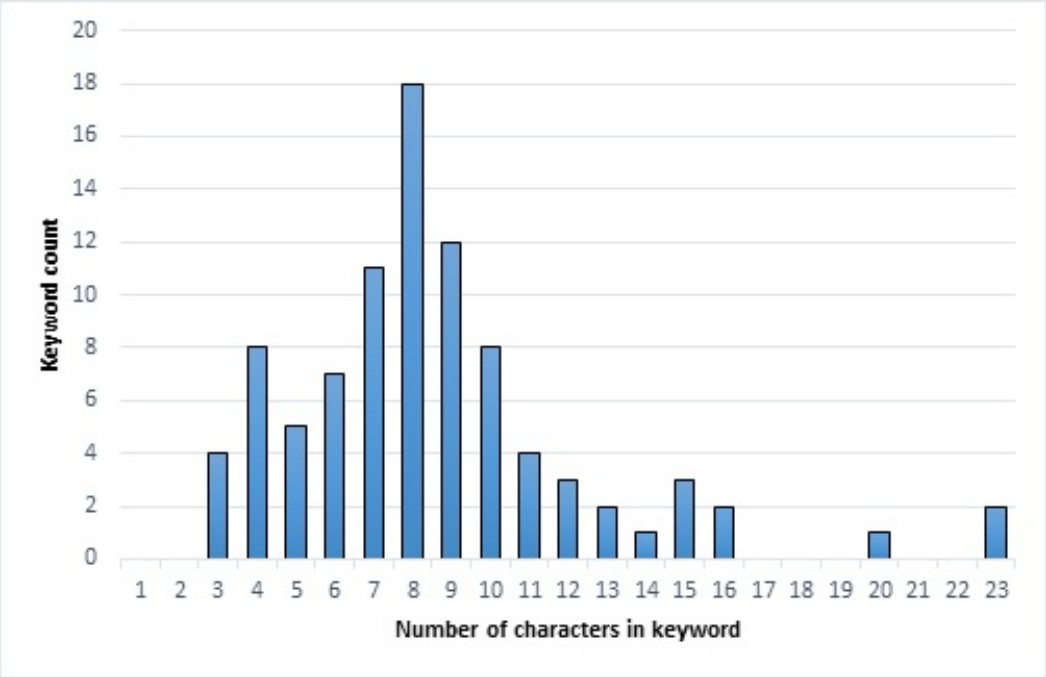


Fig. 5.3 Histogram of the number of characters present in keyword set

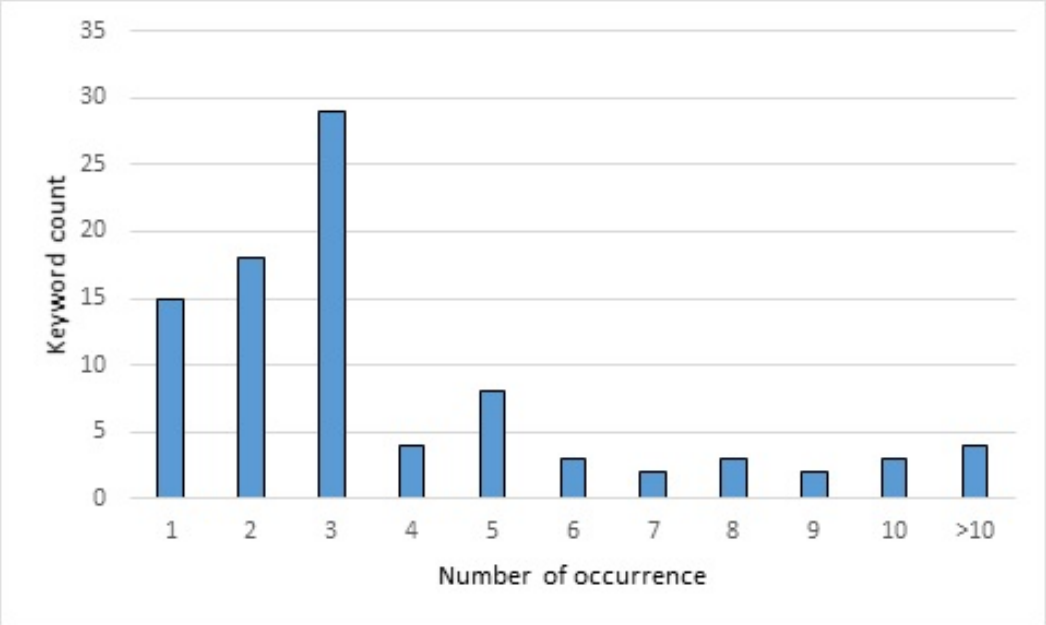


Fig. 5.4 Histogram of the number of occurrence of the keywords belongs to MNAC news audio corpus

The KWS experiments are conducted on the MNAC news audio dataset, consisting of 5,250 Malayalam news audio samples. The performance evaluation of the system is conducted by a list of ninety one keywords chosen from various news categories. The following section describes the process of knowledge base generation adopted for the implementation of the HMM decoder.

5.3 Knowledge Base Generation for the Implementation of HMM Decoder

Speech is produced with pulmonary pressure provided by the lungs that generates sound by phonation through the glottis in the larynx that is then modified by the vocal tract into different vowels and consonants [189]. In this work, acoustic and language models are generated while constructing the HMM decoder to simulate the KWS system. Acoustic modelling provides acoustic observations of the signal. These observations are coefficient vectors that represent the speech signal characteristics. The knowledge based generation aims at extraction of suitable information from the news text dataset to support the implementation of proposed KWS system. Language models can be considered as the probability distribution over sequences of words. It also provides contextual information to distinguish between words and phrases that sound similar. A Knowledge Base Preparation Tool for Malayalam (KBPT-M) is proposed with the aim of generating language models and resources for the preparation of acoustic models to facilitate the implementation of the HMM decoder. The following section describes the implementation of the proposed KBPT-M in detail.

5.3.1 Knowledge Base Preparation Tool (KBPT-M)

A knowledge base preparation tool for English language has been developed by CMU (Carnegie Mellon University) [190]. A knowledge base preparation tool for Malayalam is proposed as part of this study. The Knowledge Base Preparation Tool for Malayalam (KBPT-M) compiles three text-based components to support acoustic models *viz.* sentence corpus file, word list and lexical model (phonetic dictionary). In addition to this KBPT-M also compiles to generate language model. The proposed tool automatically generates four text files as output corresponding to above mentioned components from the news sentence dataset. A block diagram representing the different components of the KBPT-M is shown in figure 5.5. The following sections describe the construction of the sentence corpus file, word list, phonetic dictionary and language model in detail.

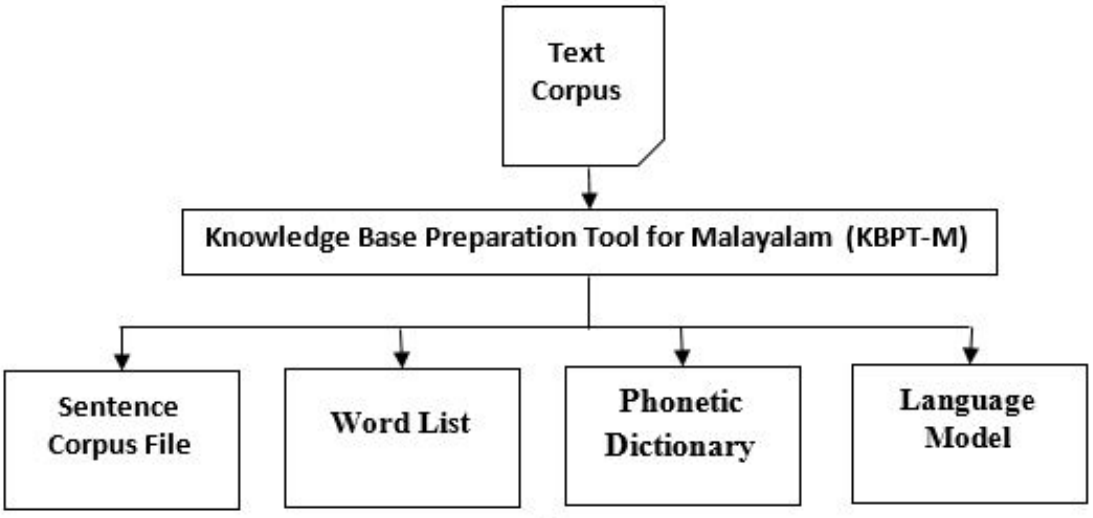


Fig. 5.5 Components of the proposed KBPT-M

a. **Sentence Corpus File**

A well-structured sentence corpus is required for the generation of acoustic and language models used in HMM decoder. A refer-

ence text file is generated for this purpose with each sentence text delimited by <s> and </s> tags. This sentence corpus file lists out all the sentences taken from the Malayalam news text corpus (MNSTC) bounded by start and end sentence markers: <s> and </s>. A simple text corpus file generated using the sentences that come under state news category selected from the MNSTC news text corpus is given as follows.

```
#Name: DCS-KBPT-M/Sentence corpus file-015
#Language: Malayalam
#Dataset: Malayalam news text dataset/state_news
#Format: <s> <Sentence> </s>
0001<s>സംസ്ഥാനത്ത് മൂന്നുലക്ഷം ഹെക്ടറിൽ നെൽകൃഷി ഇറക്കുന്നതിനുള്ള
        നടപടികളുമായി മുന്നോട്ടുപോകുമെന്നു മന്ത്രി സുനിൽ കുമാർ</s>
0002<s>ആറന്മുള വിമാനത്താവള പദ്ധതി പ്രദേശത്തു കൃഷി ഇറക്കുന്നതിന്റെ മുന്നോടിയായി
        നിലമൊരുക്കൽ ഉദ്ഘാടനം ചെയ്തു പ്രസംഗിക്കുകയായിരുന്നു അദ്ദേഹം</s>
0003<s>തരിശുകിടക്കുന്ന ഭൂമിയിലെല്ലാം കൃഷി ഇറക്കാനാണ് പദ്ധതിയെന്നു മന്ത്രി പറഞ്ഞു</s>
0004<s>കുറെക്കാലം തരിശിട്ടാൽ കൃഷി ഇല്ലാതാകുമെന്നു കരുതി ഏതെങ്കിലും ഭൂമാഫിയ ഭൂമി
        തരിശായി കൈവശം വച്ചിട്ടുണ്ടെങ്കിൽ അതവരുടെ തെറ്റായധാരണ മാത്രമാണ്</s>
0005<s>അവരുടെ കൈവശത്തിലുള്ള ഭൂമിയും കൃഷിക്ക് ഉപയുക്തമാക്കണമെന്നാണ്
        തനിക്ക് അഭ്യർഥിക്കാനുള്ളതെന്ന് സുനിൽ കുമാർ പറഞ്ഞു</s>
0006<s>തരിശു ഭൂമിയിൽ കൃഷി ഇറക്കുന്ന ഹരിതകേരളം പദ്ധതിയുടെ ഭാഗമായി
        ആദ്യം വിത്തിറക്കുന്നത് ആറന്മുളയിലായിരിക്കും</s>
0007<s>അതിനായി കേരളപ്പിറവിക്കു മുമ്പായി നിലം പൂർണമായി കൃഷി സജ്ജമാക്കും</s>
0008<s>തരിശുഭൂമിയിൽ കൃഷി ഇറക്കുന്നതിനുള്ള ഊർജസ്രോതസു കൂടിയാണ്
        ആറന്മുളയെന്നും സുനിൽ കുമാർ പറഞ്ഞു</s>
0009<s>വീണാ ജോർജ് എംഎൽഎ അധ്യക്ഷത വഹിച്ചു</s>
0010<s>രാഷ്ട്രീയ നേതാക്കളെ ആവശ്യമില്ലാതെ ഗുണ്ടാ ആക്ടിൽ ഉൾപ്പെടുത്തരുതെന്നു
        മുഖ്യമന്ത്രി പിണറായി വിജയന്റെ നിർദേശം</s>
0011<s>ആഭ്യന്തര വകുപ്പിലേയും പോലീസിലേയും ഉന്നത ഉദ്യോഗസ്ഥരുടെ
        യോഗത്തിലാണു മുഖ്യമന്ത്രി നിർദേശം നൽകിയത്</s>
0012<s>രാഷ്ട്രീയ നേതാക്കളെയും വ്യക്തിവൈരാഗ്യമുള്ളവരേയും ഉൾപ്പെടുത്തുന്ന
        പോലീസ് സമീപനം അവസാനിപ്പിക്കണം</s>
0013<s>പൊതുജനങ്ങളോടുള്ള ചില പോലീസ് ഉദ്യോഗസ്ഥരുടെയെങ്കിലും ഇപ്പോഴത്തെ
        പെരുമാറ്റം ശരിയായ രീതിയിൽ അല്ല</s>
0014<s>സേനയുമായി ബന്ധപ്പെട്ടുയരുന്ന ആക്ഷേപങ്ങൾ അടിക്കടി വർധിക്കുകയാണെന്നും
        ഇക്കാര്യം ഗൗരവമായി കണ്ടു തിരുത്തൽ നടപടികൾ കൈക്കൊള്ളണം</s>
0015<s>സംസ്ഥാനത്തു ലഹരിവിരുദ്ധ പ്രവർത്തനങ്ങൾ കൂടുതൽ കാര്യക്ഷമമാകണം</s>
0016<s>പോലീസ് സ്റ്റേഷനുകളിൽ കെട്ടിക്കിടക്കുന്ന വാഹനങ്ങൾ ഒഴിവാക്കാനും നിർദേശിച്ചു</s>
#EOF
```

b. **Word List**

A file consisting of complete vocabulary list is generated from the MNSTC news text corpus. In this file, all the words extracted from

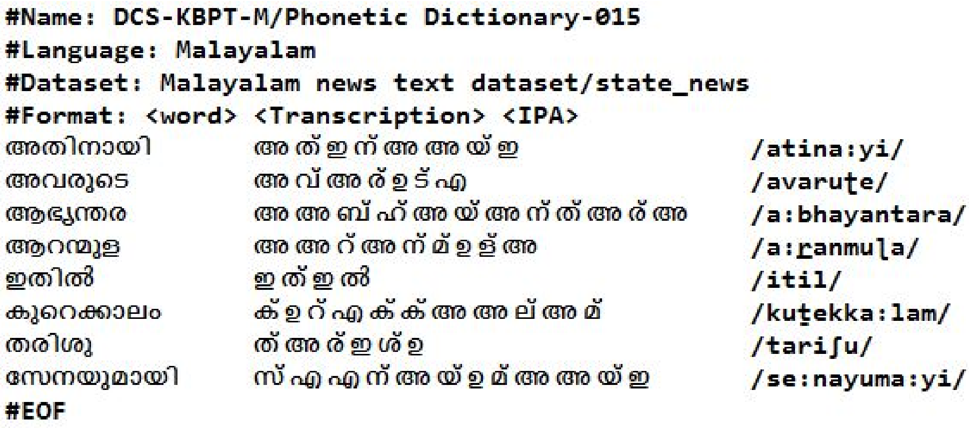
the text corpus are listed in ascending order. A list of all Malayalam graphemes in its ascending order is provided in Appendix 1. Special characters and symbols present in the sentences are removed in the pre-processing stage. A sample word list is as follows.

```
#Name: DCS-KBPT-M/Word lsit-015
#Language: Malayalam
#Dataset: Malayalam news text dataset/state_news
#Format: <word> <IPA>
അതിനായി          /atina:yi/
അവരുടെ           /avaruṭe/
ആഭ്യന്തര         /a:bhayantara/
ആറന്മുള          /a:ṟanmuḷa/
ഇതിൽ             /itil/
കുറെക്കാലം       /kuṭekka:lam/
തരിശു            /tariʃu/
തരിശുകിടക്കുന്ന  /tariʃukiṭakkunna/
തരിശുഭൂമിയിൽ    /tariʃubhu:miyil/
പൊതുജനങ്ങളോടുള്ള /potuʒanaŋŋaḷo:ṭuḷḷa/
പോലീസ്           /po:li:s/
രാഷ്ട്രീയ        ra:ʂṭṟi:ya/
സംസ്ഥാനത്ത്      samstha:nattu/
സേനയുമായി        se:nayuma:yi
#EOF
```

c. Phonetic Dictionary

Development of an automatic Grapheme to Phoneme (G2P) automatic transcription tool for Malayalam is already presented in chapter 4. The proposed G2P transcriptor is enhanced to map the vocabulary word list entries to its corresponding phone sequences. Possibility of alternative pronunciations is rare in Malayalam language as it has almost one to one correspondence between graphemes and phonemes. Hence the provision for the alternative pronunciation for the words are not included in the phonetic dictionary. All the entries in the word list are also included in the phonetic dictionary. The phonetic transcription is performed based on the 50 Malayalam phones introduced in chapter 3. The structure of the

phonetic dictionary generated by the help of KBPT-M is given as follows.



```
#Name: DCS-KBPT-M/Phonetic Dictionary-015
#Language: Malayalam
#Dataset: Malayalam news text dataset/state_news
#Format: <word> <Transcription> <IPA>
അതിനായി        അ ത് ഇ ന് അ അ യ് ഇ                /atina:yi/
അവരുടെ         അ വ് അ ര് ഉ ട് എ                  /avaruṭe/
ആഭ്യന്തര       അ അ ബ് ഹ് അ യ് അ ന് ത് അ ര് അ      /a:bhayantara/
ആറന്മുള        അ അ റ് അ ന് മ് ഉ ള് അ              /a:r̲anmuɭa/
ഇതിൽ           ഇ ത് ഇ ൽ                          /itil/
കുറെക്കാലം      ക് ഉ റ് എ ക് ക് അ അ ല് അ മ്          /kuṯekka:lam/
തരിശു          ത് അ ര് ഇ ശ് ഉ                    /tariʃu/
സേനയുമായി      സ് എ എ ന് അ യ് ഉ മ് അ അ യ് ഇ          /se:nayuma:yi/
#EOF
```

d. **Language Model**

A language model for the purpose of building HMM models are generated based on the IRST Language Modeling (IRSTLM) Toolkit. IRSTLM is widely used to estimate, store, and access very large n-gram language models [191]. It is an experimental language model compiler that produces a conventional backed-off trigram language model based on the given corpus of text data in which n-gram probabilities are also shown together. In this work, the IRSTLM toolkit is used to generate the language model based on sentence corpus file. A sample language model generated using the IRSTLM compiler on a MNSTC news audio corpus is given as follows.

```
#Name: DCS-KBPT-M/Language Model-015
#Language: Malayalam
#Dataset: Malayalam news text dataset/state_news
#Format:#\1-grams: <probability> <word> <probability>
        #\2-grams: <probability> <word> <word> <probability>
        #\3-grams: <probability> <word> <word> <word>
\data\
ngram 1= 62
ngram 2= 61
ngram 3= 2

\1-grams:
-1.989005 <s>കശാപ്പ് -0.041393
-1.989005 നിയന്ത്രണം -0.041393
-1.989005 കേന്ദ്ര -0.041393
-1.989005 വിജ്ഞാപനത്തിന് -0.041393
-1.989005 താൽക്കാലിക -0.041393
-1.989005 സ്റ്റേ -0.041393
-1.335792 </s> -0.000000
-1.989005 <s>ബാബറി -0.217484
-0.497643 <unk>
#...
\2-grams:
-0.439333 <s>ബാബറി മസ്ജിദ് -0.332064
-1.518514 <s>ബാബറി കേസ് 0.000000
#...
\3-grams:
-0.272066 <s>ബാബറി മസ്ജിദ് കേസ്
-0.272066 മസ്ജിദ് കേസ് ബിജെപി
#...
#EOF
```

This knowledge base, including the sentence corpus file, word list, phonetic dictionary and language model generated using KBPT-M is further used to implement the HMM decoder which is an integral part of the proposed KWS system. The following section describes the architecture of the HMM based speech recognition system which assists the implementation of the proposed KWS system.

5.4 Architecture of the HMM based Automatic Speech Recognition (ASR) System

Input to an ASR system can generally be an audio wave either prior recorded or taken live from the microphone. The HMM encoder first converts the audio waveform into a sequence of fixed size acoustic feature vectors. The proposed ASR uses Mel Frequency Cepstral Coefficients (MFCC) features for speech encoding. Later in speech decoding, these feature vectors are used for recognising the input speech. The speech decoding problem can be defined as finding the sequence of words $W_{1:m} = w_1, w_2, \ldots w_m$ which is most likely correspond to the feature vector sequence $F_{1:n} = f_1, f_2, \ldots f_n$. Using the Bayes' Rule, the speech decoding problem can be defined as

$$\hat{w} = \underset{w}{Argmax}\{p(F \vee W)p(W)\} \tag{5.1}$$

Where $p(F \vee W)$ is the likelihood of the word w_m with respect to F and $p(W)$ is the prior probability of the word sequence W. The product $p(F \vee W)p(W)$ is computed for every w_m. Error probability is presume to be the minimum for which the product term is maximized. HMM computes the product term and underlying probabilities using statistical models. The likelihood $p(F \vee W)$ is determined by its acoustic model and the prior probability $p(W)$ by the language model. Acoustic model of a given word w_m is synthesised by concatenating phone models defined in pronunciation dictionary. The language model used in this work is based on n-gram model. In n-gram model each word is conditioned based on its n-1 predecessors. Its parameters are estimated by counting n-tuples in the text corpus. The proposed ASR system is implemented using an open source framework Sphinx-4 [192]. Sphinx-4 is a flexible, modular and

pluggable framework that helps to build Hidden Markov model (HMM) based speech recognition systems. The following sections describe the MFCC feature extraction process, acoustic modelling, language modelling and the speech decoding algorithm used to build the HMM based ASR system.

5.4.1 MFCC Feature Extraction Process

The proposed HMM based ASR system uses Mel Frequency Cepstral Coefficients (MFCC) parameters for speech encoding. Psycho-physical studies have shown that Mel scale, a nonlinear frequency scale which approximates the response of the human ear, is approximately linear below 1 kHz and increases logarithmically above it. MFCC feature extraction method uses Mel filter banks to extract subjective information from the spectrum of speech signal [193]. MFCC algorithm focuses on the numerical analysis of the signal which has to be converted to fixed point arithmetic. It is the simplest and most widely used encoding scheme. The block diagram of the MFCC feature extraction method is shown in figure 5.6. The feature vectors are generated by applying a truncated Discrete Cosine Transformation (DCT) to a log spectral estimate computed by smoothing a Fast Furrier Transform (FFT) with 20 frequency bins distributed non-linearly across the speech spectrum.

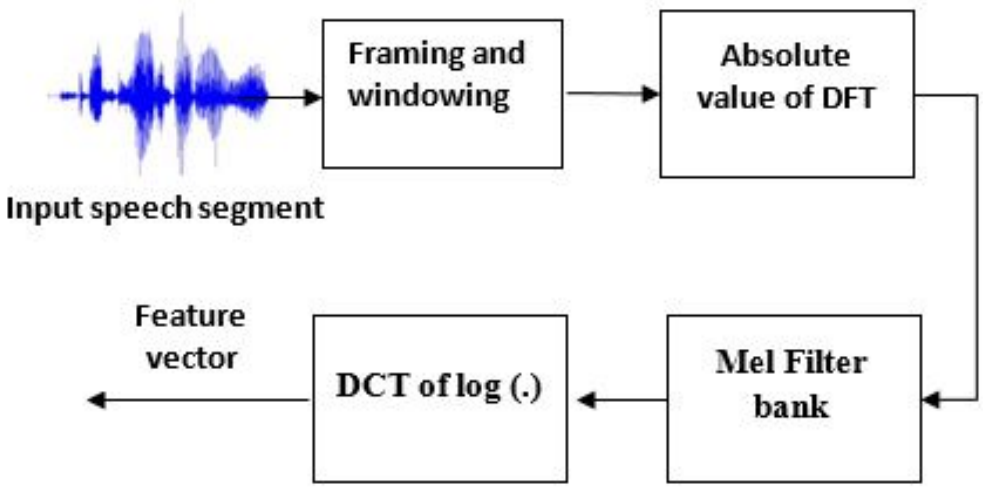


Fig. 5.6 Block diagram of MFCC feature extraction process

In addition to the spectral coefficients, first order (delta) and second-order (delta–delta) regression coefficients are often appended in a heuristic attempt to compensate for the conditional independence assumption made by the HMM-based acoustic models [**?**]. Let the original feature vector be y_t^s, the delta parameter Δy_t^s is given by,

$$\Delta y_t^s = \frac{\sum_{i=1}^{n} w_i (y_{t+i}^s - y_{t-i}^s)}{2\sum_{i=1}^{n} w_i^2} \tag{5.2}$$

Where w_i is the regression coefficient and n is the window width. The second derivative $\Delta^2 y_t^s$ is computed using the same method with different delta parameters. Finally the feature vector y_t is formed by concatenating these values,

$$y_t = [y_t^{sT} \Delta y_t^{sT} \Delta^2 y_t^{sT}]^T \tag{5.3}$$

The dimensionality of the final MFCC feature vector used in the study is 39. This feature vector is further used to develop acoustic model of the speech signal.

5.4.2 Development of Acoustic Models for HMM

This section describes the method adopted for the preparation of acoustic models for HMM based speech recognition system. A word w is considered as combination of phone sequence. This phone sequence is called as *pronunciation* and is represented as $q^w_{1:k} = q_1, q2, \ldots q_k$. Each q_i can be represented as a continuous density HMM with transition probability parameter A and output observation distribution B. The general representation of a HMM based phone model is illustrated in figure 5.7. It can be seen that on each step HMM makes a transition from its current state to one of its neighbouring state. The probability of transition from a state s_i to state s_j is given by the state transition probability a_{ij}. On each state, a feature vector is generated using the distribution associated with the corresponding state, $b_j()$.

This type of operations yield the following two standard conditional independent assumptions for an HMM [191]

1. States are conditionally independent of all other states given the previous state;
2. Observations are conditionally independent of all other observations given the state that generated it.

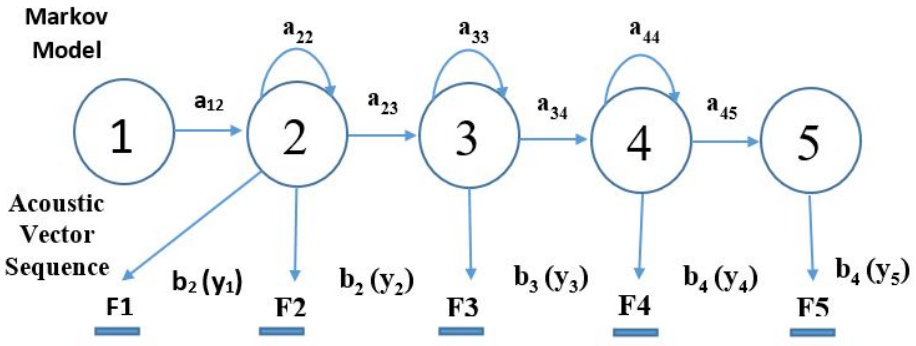


Fig. 5.7 HMM based phone model

The core acoustic model of an HMM based speech recognizer often consists of tied three-state HMMs with Gaussian output distributions [194]. The procedure used for the generation of core acoustic model is given in algorithm 3.

Algorithm 3 Acoustic model generation using HMM

This algorithm takes the phone sequence from the speech sample as input and produces acoustic model as output. Initially a prototype model is defined for HMM training. A set of identical monophone HMMs in which every mean and variance is identical are created. These are then re-estimated using tri-phones for creating acoustic model set.

Input: Phone sequence of the speech sample

Output: Tied-state context-dependent acoustic model set

1: Create a flat start monophone set. Each base phone in the set is modelled as a single-Gaussian HMM.
2: Re-estimate Gaussian monophone parameters using expectation maximization (EM) method.
3: Perform cloning of each Gaussian monophone q based on each distinct triphone $x-q+y$ that appears in the training data.
4: Re-estimate training data tri-phones using EM.
5: Create decision tree for each state in each base phone. The triphones are mapped into a smaller set of tied-state triphones and iteratively re-estimated using EM.

The detailed description of the implementation of the acoustic model generation algorithm is detailed below.

In the first step, single multivariate Gaussian is considered for the output distribution as,

$$b_j(f) = N(y; \mu^j, \Sigma^j) \tag{5.4}$$

where, μ^j is the mean of state s_j and Σ^j is the covariance of state s_j

The acoustic vector f has relatively high dimensionality. Hence its covariances are often considered to be diagonal. The composite HMM Q is formed by concatenating all the constituent base phone

$q_{(w1)}, q_{(w2)}, \ldots q_{(wL)}$. The acoustic likelihood is given in equation number 5.5.

$$p(F|Q) = \sum_{\theta} p(\theta, F|Q) \tag{5.5}$$

where, θ is a state sequence $\theta_0, \theta_1, \theta_{T+1}$ through the composite model Q and $p(\theta, F|Q)$ is,

$$p(\theta, F|Q) = a\theta_0\theta_1 \prod_{t=1}^{T} b_{\theta_t}(f_t) a\theta_t\theta_{t+1} \tag{5.6}$$

Here θ_0 and θ_{T+1} are non-emitting states for entry and exit. Hence for the modelling purpose, the focus will be on the remaining states as shown in figure 5.7.

The acoustic model parameters $\lambda = [a_{ij}, b_j()]$ are estimated using forward–backward algorithm from the training dataset [195]. The proposed estimation method is an example of Expectation Maximisation (EM) [196]. The HMMs that correspond to the word sequence in the utterance F^{γ} of length T^{γ} is constructed and the corresponding composite HMM is also built using the forward–backward algorithm. In the first phase, the forward probability $\alpha_t^{rj} = p(F_{1:t}^r, \theta_t = s_j; \lambda)$ and the backward probability $\beta_t^{ri} = p(F_{1:t}^r, \theta_t = s_j; \lambda)$ are computed via the following recursion equations 5.7 and 5.8

$$\alpha_t^{(rj)} = \left[\sum_i \alpha_{t-1}^{(ri)} a_{ij}\right] b_j\left(f_t^{(r)}\right) \tag{5.7}$$

$$\beta_t^{(ri)} = \left[\sum_j a_{ij} b_j\left(f_{t+1}^{(r)}\right) \beta_{t+1}^{(rj)}\right] \tag{5.8}$$

where i and j are added up for all states. For long speech segments usually *log* arithmetics are used in recursion for reducing the compu-

tational cost [197]. Using the above computed forward and backward probabilities, the probability of the model occupying state s_j at time t for the utterance r is computed using the equation 5.9,

$$\gamma_t^{(rj)} = P(\theta_t = s_j | F^{(r)}; \lambda) = \frac{1}{P^{(r)}} \alpha_t^{(rj)} \beta_t^{(rj)} \tag{5.9}$$

where $P^{(r)} = p(F^{(r)}; \lambda)$.

These state occupation probabilities γ represent a soft alignment of the model states to the data [198]. Hence the new set of Gaussian parameters defined to maximise the likelihood of the data based on these alignments, are given in equation 5.10 and 5.11,

$$\hat{\mu}^{(j)} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \gamma_t^{(rj)} f_t^{(r)}}{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \gamma_t^{(rj)}} \tag{5.10}$$

$$\hat{\Sigma}^{(j)} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \gamma_t^{(rj)} (f_t^{(r)} - \hat{\mu}^{(j)})(f_t^{(r)} - \hat{\mu}^{(j)})^T}{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \gamma_t^{(rj)}} \tag{5.11}$$

Similarly, a re-estimation equation for the transition probabilities is given as,

$$\hat{a}^{(j)} = \frac{\sum_{r=1} \frac{1}{P^{(r)}}^{R} \sum_{t=1}^{T^{(r)}} \alpha_t^{(ri)} a_{ij} b_j(f_{t+1}^{(r)}) \beta_{t+1}^{(rj)} f_t^{(r)}}{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \gamma_t^{(ri)}} \tag{5.12}$$

In the second step of the algorithm, based on the initial estimate of the parameters $\lambda^{(0)}$, EM algorithm yield parameter sets $\lambda^{(1)}, \lambda^{(2)}, \ldots$ on successive iterations. These estimation improves the likelihood up to some local maximum. In flat start model, the initial value for the $\lambda^{(0)}$ is assigned to global mean and covariance of the data to the Gaussian output distributions and set all transition probabilities to be equal. Thus each word is decomposed into a sequence of context independent base

phones in this model. It is also observed that these models fails to capture the large degree of context dependent variability existing in real speech.

In order to incorporate the context dependent variation of the phones unique phone model for every possible pair of left and right neighbours are used, named as *tri-phones*. $x-q+y$ denotes the triphone corresponding to phone q spoken in the context of a preceding phone x and a following phone y. These model has N base phones and N^3 potential triphones. To avoid the resulting data scarcity problems, the complete set of logical triphones L can be mapped to a reduced set of physical models P by clustering and tying together the parameters in each cluster. This mapping process is illustrated in Figure 5.8. The parameter is said to be tied in the HMMs of two sound units if it is identical for both of them. The process of parameter tying is depicted in figure 5.9 with an example.

Pronunciation q of each phone is obtained from the pronunciation dictionary generated as part of the knowledge base generation process as explained in section 5.3. These pronunciations are then mapped on to the logical phones based on their context. Logical phones are then mapped on to the physical models. Logical to physical model clustering usually operates at the state level rather than the model level.

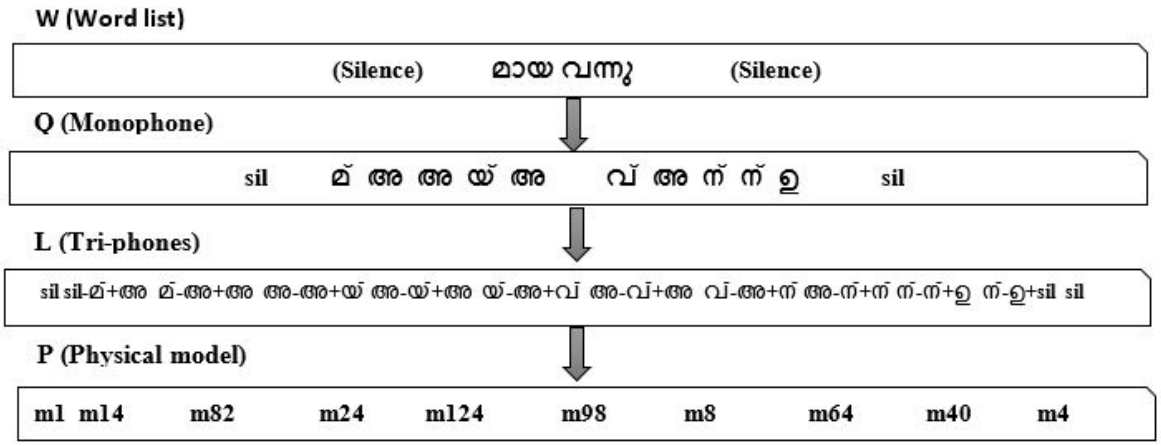


Fig. 5.8 Context dependent phone modelling applied on Malayalam text മായ വന്നു /ma:ya vannu/

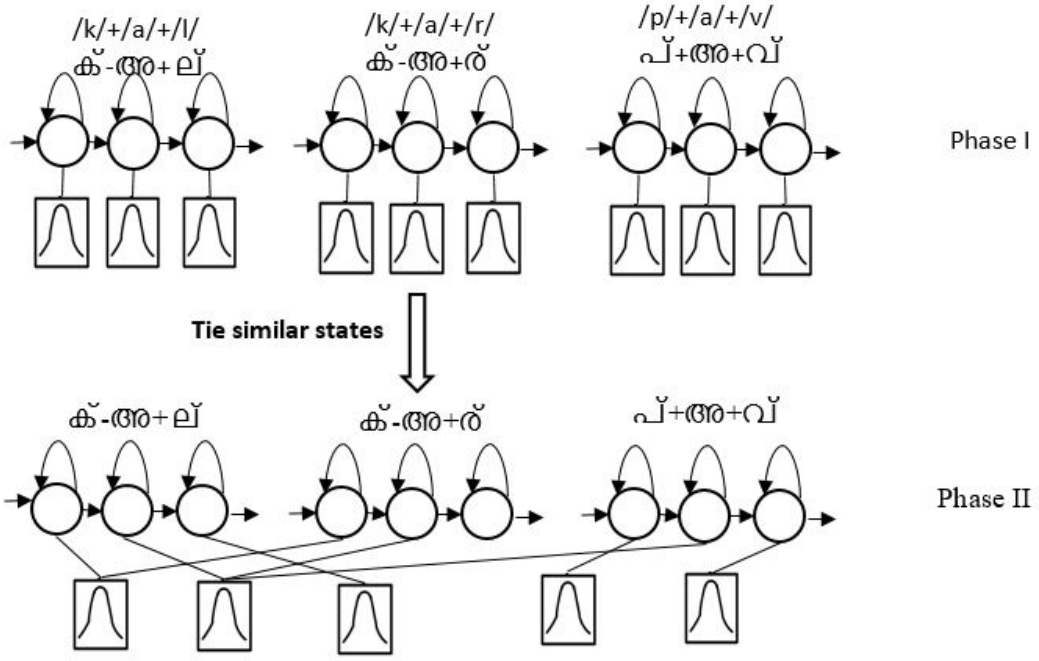


Fig. 5.9 Tied-state phone model for the formation of physical model

Finally as described in step 5 of the Algorithm 3, decision trees are implemented to make the choice of which states to tie. State position

of every phone q_i has a binary tree associated with it. Consider the state i of phone q. All states i of all the logical models derived from q are collected into a single pool at the root node of the tree. The splitting of state pools depends on the answer to the conditions at each node. The process is repeated until all the states are trickled down to leaf nodes. The generic example of decision tree clustering is shown in figure 5.10. Here R represents right nodes and L indicates left nodes. Transition in decision tree will be based on left and right neighbouring values. All states in each leaf node are then tied to form a physical model. The questions corresponding to each node are fixed based on the predetermined set to maximize the likelihood of the training data. The decision tree expansion is very effectively performed using a greedy iterative node splitting algorithm [198]. As a post processing, single Gaussian models are converted to mixture Gaussian models [191]. Thus the final result is the required tied-state context-dependent acoustic model set.

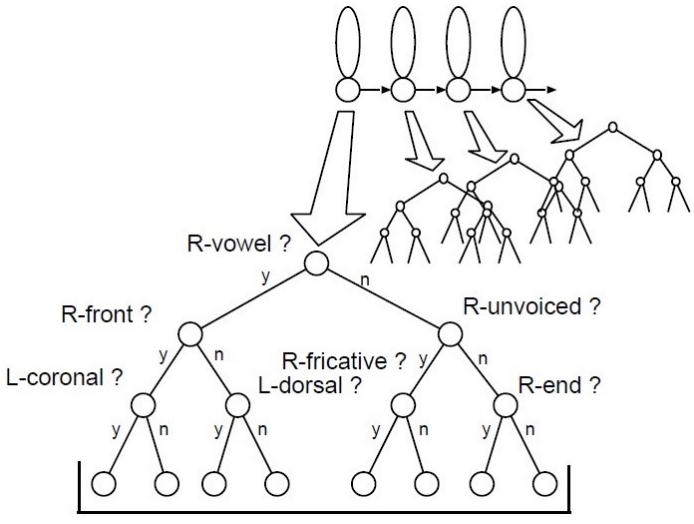


Fig. 5.10 Decision tree clustering model

5.4.3 Generation of *n-gram* Language Models

In speech recognition applications,language model $P(w)$ has critical influence on the recognition accuracy. In most of the cases, it is generated from the given text corpus. This section discusses the process adopted for the construction of language models using *n-gram* modelling for building HMM decoder.

In an n-gram model, the probability of observing the sentence $w = w_1, w_2, \ldots w_k$ defined in 5.1 is approximated as,

$$P(w) = p(w_1).p(w_2|w_1)\ldots p(w_k|w_1^{k-1}) \tag{5.13}$$

where, $w_1^{k-1} = w = w_1, w_2, \ldots w_{k-1}$ is the sequence of occurrence of words. The value of n is fixed as 3, as we have used *tri-gram* models for the generation of language models.

The performance of proposed language models is assessed in the context of speech recognition, based on the *perplexity*. The *perplexity* is defined as the degree of difficulty that the recognizer encounters, on an average, when it has to determine a word from the same source [188]. This difficulty depends on the actual probability $P(w_1, w_2, \ldots w_k)$ which is not known before and thus has to be estimated. Hence the *perplexity* H can be defined as

$$H = -\lim_{K\to\infty}\frac{1}{K}\log_2(P(w_1,\ldots w_k)) \approx -\frac{1}{K}\sum_{k=1}^{K}\log_2(P(w_k|w_{k-1}, w_{k-2}, \ldots w_{k-N+1})) \tag{5.14}$$

The language model with lower H is considered as better than any of the other language model which leads to a higher *perplexity*. The *n-gram* probabilities are estimated from training texts by counting *n-gram*

occurrences to form Maximum Likelihood (ML) parameter estimates. For example, let $C(w_{k-2}, w_{k-1}, w_k)$ represent the number of occurrences of the three words w_{k-2}, w_{k-1} and w_k and similarly $C(w_{k-2}, w_{k-1})$ represents the number of occurrence of w_{k-2}, w_{k-1}, then

$$P(w_k|w_{k-1}, w_{k-2}) \approx \frac{C(w_{k-2}w_{k-1}w_k)}{C(w_{k-2}w_{k-1})} \tag{5.15}$$

This simple ML estimation shows sparsity. To overcome this sparsity problem the scheme is integrated by a combination of discounting and backing off [199].

$$P = (w_k|w_{k-1}, w_{k-2}) = \begin{cases} d\frac{C(w_{k-2}w_{k-1}w_k)}{C(w_{k-2}w_{k-1})} & if \quad 0 < C < C' \\ \frac{C(w_{k-2}w_{k-1}w_k)}{C(w_{k-2}w_{k-1})} & if \quad C > C' \\ \alpha(w_{k-1}, w_{k-2}) \quad P(w_k|w_{k-1}) & otherwise, \end{cases} \tag{5.16}$$

where C is a count threshold, C is short-hand for $C(w_{k-2}w_{k-1}w_k)$, d is a discount coefficient and α is a normalisation constant. Thus, when the *n-gram* count exceeds the threshold, the ML estimate is used. When the count is small, the same ML estimate is used but discounted slightly. The following section describes the implementation of HMM decoding algorithm and its enhancement based on word lattice generation.

5.4.4 HMM Decoding and Word Lattice Generation

The most likely word $\hat{w}$ given a sequence of feature vectors $F_{1:T}$ is found by searching all possible state sequences for the sequence which was most likely to have generated the observed features. The generated sequence that shows maximum likelihood to the observed feature data $F_{1:T}$ is to

be selected. Dynamic programing concepts are used to efficiently solve the decoding problem. Let,$\phi_t^{(j)} = max_\theta p(F_{1:t,\theta_t})s_j;\lambda$ i.e., the maximum probability of observing the partial sequence $F_{1:t}$ and then being in state s_j at time t, given the model parameter λ. This probability can be suitably computed using the Viterbi Algorithm [200] as follows,

$$\phi_t^{(j)} = max_i \phi_{t-1}^{(j)} a_{ij} b_j(f_t) \tag{5.17}$$

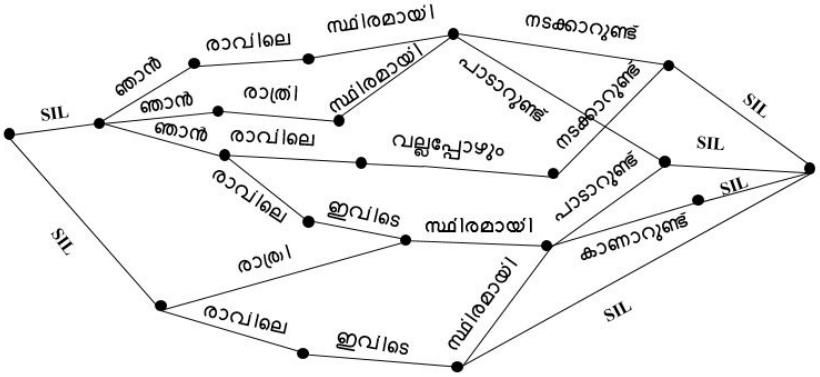
The non-emitting, entry state $\phi_t^{(j)}$ is initialized to 1 and all other state to 0. The probability of the most likely word sequence is then given by $max_i \phi_t^{(j)}$.After recording all the decisions, a traceback is performed to obtain the best matching state sequence.

HMM decoder primarily finds the solution for the equation (17). Rather than the most likely hypothesis, the N-best set of hypothesis generation is optimal for speech processing applications. It allows multiple passes over the data without the computational expense of repeatedly solving the equation (17) from scratch. A compact and efficient structure for storing these hypotheses is the word lattice [68].

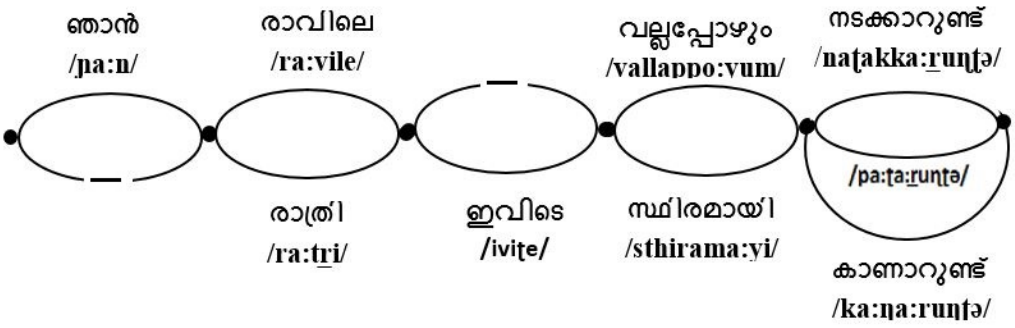
A word lattice consists of a set of nodes representing points in time and a set of spanning arcs representing word hypotheses. An example of word lattice structure created using nine Malayalam words is shown in Figure 5.11a. In addition to the words assigned to each arc as shown in the figure, every arc also carry acoustic and language model score information. In lattice structure *SIL* stands for silence detected in the beginning and end of each file.

Word lattices are flexible in nature and can be expanded to allow rescoring by a higher order language model. This lattice structure is then compacted into a very efficient representation called a confusion network [68]. A confusion network generated out of the given word lattice

structure is shown in Figure 5.11b where the "-" arc label indicates NULL transitions. The confusion network has the property that for every path through the original lattice, there exists a corresponding path through the confusion network. Each arc in the confusion network carries the posterior probability of the corresponding word w. Confusion networks are then used to construct minimum word-error decoding [201]. The list of sample words used for the creation of word lattice and confusion network is given in table 5.1.



(a) A word lattice structure



(b) Confusion network formed out of the given word lattice structure

Fig. 5.11 Word Lattice Structure and the corresponding Confusion Network

Table 5.1 List of Malayalam words used to generate the word lattice and confusion network

Sl.No.	Word	IPA
1	ഞാൻ	/ɲa:n/
2	രാവിലെ	/ra:vile/
3	രാത്രി	/ra:t̪r̄i/
4	വല്ലപ്പോഴും	/vallappo:ẏum/
5	സ്ഥിരമായി	/sthirama:yi/
6	നടക്കാറുണ്ട്	/naṭakka:r̄uɳṭə/
7	പാടാറുണ്ട്	/pa:ṭa:r̄uɳṭə/
8	കാണാറുണ്ട്	/ka:ɳa:r̄uɳṭə/
9	ഇവിടെ	/iviṭe /

5.5 Proposed Keyword Spotting System Architecture for Malayalam

Two different Keyword Spotting (KWS) techniques are resulted as part of this study. The first method is based on Automatic Speech Recognition (ASR) approach and the second method is a Filler Model based Acoustic (FMA) approach. The implementation details of these techniques are discussed in the following sections. The experimental results obtained out of keyword spotting experiments are also compared and discussed in detail.

5.5.1 ASR based Keyword Spotting Technique

In this method, an ASR based Keyword Spotting system (ASR-KWS) is implemented based on automatic continuous speech recognition technique.

There are two phases in the proposed technique. First one is the speech to text conversion and the second phase is text-based search to locate the keyword. During the first phase, most probable sequence of words are listed based on the knowledge base information using Viterbi search algorithm. In the second phase, KWS engine uses the ASR output for text-based search to locate keywords.

Word lattices are created to impliment an efficient keyword searching in ASR based KWS with word lattice search (ASR–LS-KWS). A word lattice comprise of a set of nodes representing the points in time and spanning arcs representing word hypotheses are then computed [68] as discussed in section 5.4. The ASR engine finds the acoustic and language model likelihood of each word. Forward, backward and posterior scores are computed using forward-backward inference algorithm [202]. The mathematical equation used for the computation of these confidence scores is given below.

$$\begin{aligned} C(N) &= L_{\alpha}(N) + L_{\beta}(N) + L(N) - L_{best} \\ L_{\alpha}(N) &= L_{\alpha accoustic}(N) + L_{\alpha language}(N) + \min_{N_P} L_{\alpha}(N_P) \\ L_{\beta}(N) &= L_{\alpha accoustic}(N) + L_{\alpha language}(N) + \min_{N_F} L_{\alpha}(N_F) \end{aligned} \tag{5.18}$$

where,

$L_{\alpha}(N)$ is the forward likelyhood of the best path from the begining of the latice to the keyword.

$L_{\beta}(N)$ is the backword likelyhood of the best path from the end of the latice to the keyword.

$L(N)$is the word posterior.

L_{best} is the best path through lattice.

$L_{\alpha accoustic}(N)$ is the acoustic model score.

$L_{\alpha language}(N)$ is the language model score.

The keywords to be spotted may be a single word or a combination of words. In such cases the largest confidence score obtained for any word constituting the keyword is chosen as the whole keyword confidence score. Lattice is represented as a directed graph that contains nodes representing words spoken over a particular period of time and edges that correspond to the score information such as the acoustic and language model scores. The branching of lattices usually increases processing time, but it improves the spotting results.

5.5.2 Filler Model based Acoustic Approach for Keyword Spotting(FMA-KWS)

The basic idea of filler model based acoustic approach is to create the HMM for the keywords and a separate HMM for the filler (i.e., non-keyword) regions. These two models are joined to form the composite keyword-filler HMM. Here the acoustic keyword spotting system implementation is conducted based on Audio Aligner Implementation Algorithm [203]. The aligner identifies the time when each word in the transcription was spoken in the speech utterance. The system uses acoustic model and lexicon dictionary which contains the pronunciation of the searched keywords and then non-keywords are modelled using filler models. The cost of the alignment C can be defined as

$$cost(C) = \sum_{(x_i y_j) \in C} \alpha_{x_i y_j} + \sum_{i:x_i} \delta + \sum_{j:y_j} \delta \tag{5.19}$$

where X and Y are transcription and pronunciation series. The parameters δ and α are gap and mismatch penalties respectively.

The aligner system contains keyword grammar that includes all the keywords that are to be spotted. Phoneme insertion, word insertion and out of grammar probability are the three main parameters used to tune the performance of the system. Out of grammar words are modelled with fillers and the multiword keywords were simply concatenated. Even though the system delivers fast result, the processing cost increases linearly on adding a new keyword to the list of keywords.

5.6 Experimental Results

Keyword spotting experiments are conducted based on two keyword spotting methods *viz.* ASR-KWS and FMA-KWS as discussed in section 5.5. The KWS experiments are also conducted with the improved ASR-KWS method based on word lattice search ASR-LS-KWS. The experiments are conducted on the MNAC news audio corpus familiarised in section 5.2, containing 5,250 Malayalam news audio samples. The performance evaluation of the system is conducted by a set of ninety one keywords chosen from various news categories. There are 570 mentions of these keywords in the MNAC news audio corpus. Three basic numeric scores, true positive, false positive and false negative are computed to evaluate the performance of the proposed KWS system.

- **True positive**: Number of keywords correctly recognized
- **False positive**: Number of keywords falsely recognized
- **False negative**: Number of missed keywords

The performance scores *viz.* precision, recall and F1-score of the KWS system is computed based on the above mentioned parameters. The precision is computed to measure how many of recognized keywords

are identified correctly. Recall measures how many of keywords in test data are not missed in recognition. F1 score is the weighted average of Precision and Recall and it is also referred as harmonic mean which represents the overall accuracy of the system.

$$Precision = \frac{Truepositive}{Truepositive + Falsepositive} \tag{5.20}$$

$$Recall = \frac{Truepositive}{Truepositive + Falsenegative} \tag{5.21}$$

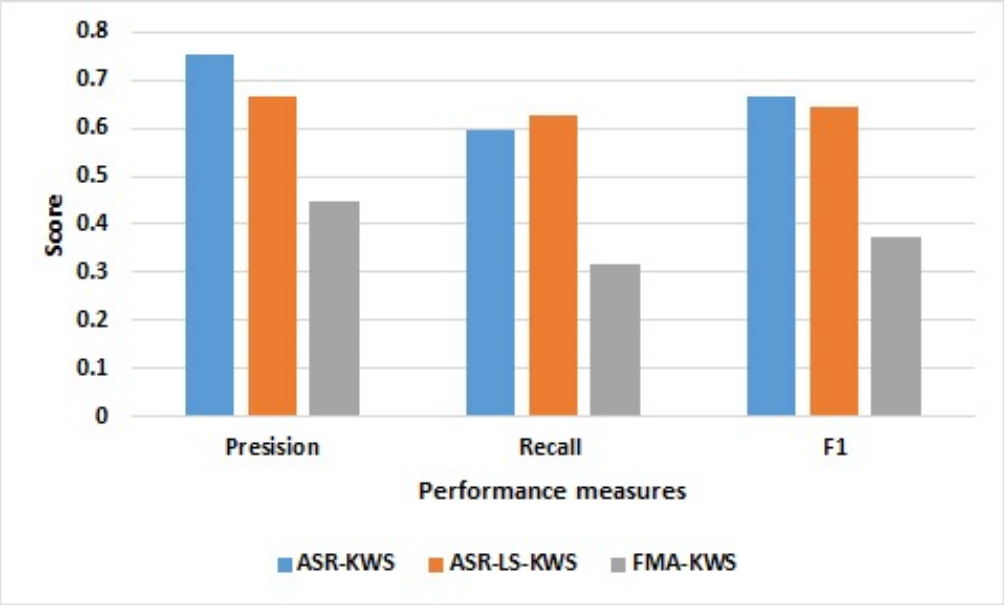
$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5.22}$$

The Exact Matching Method (EMM) uses exact keyword targeting criteria. Whereas in Relaxed Matching Method (RMM) the inflected form is allowed in the keyword targeting criteria. The performance score of the proposed KWS system is given in table 5.2.

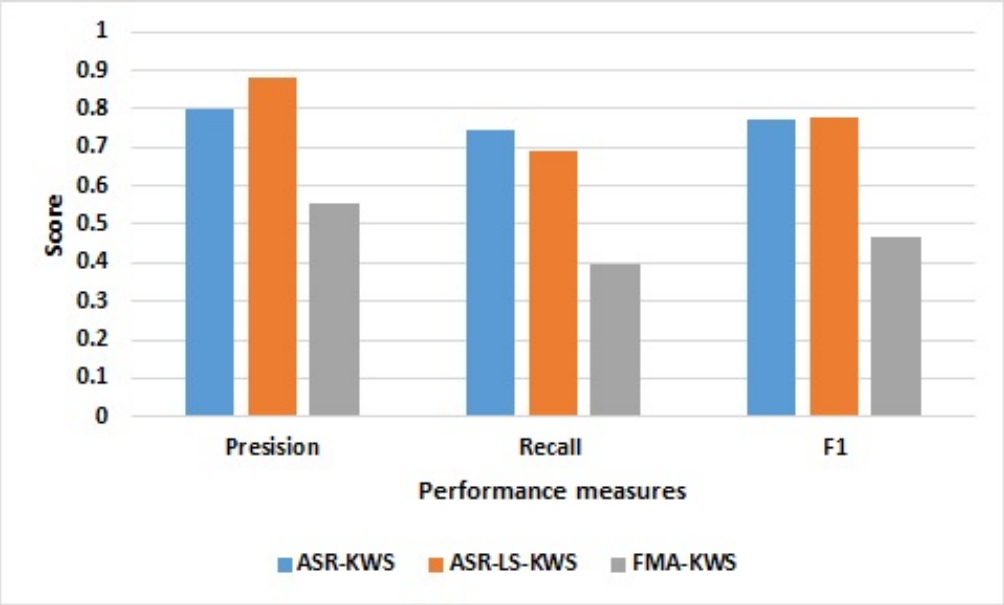
Table 5.2 Experimental result – Keyword spotting

Methods used		True Positive	False Positive	False Negative	Precision	Recall	F1 Score
Exact Matching Method (EMM)	**ASR-KWS**	285	143	171	0.6658	0.625	0.6447
	ASR-LS-KWS	298	98	202	0.7525	0.596	0.6651
	FMA-KWS	137	169	293	0.4477	0.3186	0.3722
Relaxed Matching Method (RMM)	**ASR-KWS**	376	95	128	0.7983	0.746	0.7712
	ASR-LS-KWS	379	50	168	0.8834	0.6928	0.7766
	FMA-KWS	181	144	273	0.5569	0.3986	0.4646

The graphical representation of the performance scores viz. precision, recall, and F1 of the proposed KWS method obtained using the Exact Matching Method (EMM) and Relaxed Matching Method (RMM) are shown in Figure 5.12a - 5.12b respectively. The results indicates that the FMA-KWS approach shows higher false alarm rate and is hard to tune it to reach a better result. One of the major reason is that, this approach does not use language model information. It is also evident that in ASR-LS-KWS approach, the searching for keywords in the lattice graph, precision rate is significantly increased but the recall value decreases simultaneously. The best F1-score of 0.7766 is obtained for ASR-LS-KWS in RMM.



(a) Exact Matching Method (EMM)



(b) Relaxed Matching Method (RMM)

Fig. 5.12 Performance Evaluation of Proposed KWS Systems

5.7 Conclusion

This chapter discussed the implementation details of proposed keyword spotting systems designed for Malayalam based on continuous Hidden Markov Modelling. The proposed system is trained using Malayalam news audio dataset with 5250 speech signals. A separate set of keywords containg 570 keyword samples, are also created to perform KWS experiments. A Knowledge Base Generation Tool (KBPT-M) is implemented for Malayalam language. The proposed KBPT-M is used to generate resources required for the construction of acoustic and language models using the training data.

Keyword spotting systems are implemented using an open source framework Sphinx-4. The Automatic Speech Recognition (ASR) based keyword spotting approach ASR-KWS, ASR based KWS with word lattice search (ASR-LS-KWS) and a Filler Model based acoustic keyword spotting FMA-KWS are implemented. The evaluation of the system is performed on two methods *viz.* Exact Matching Method (EMM) and Relaxed Matching Method (RMM). Precision, Recall and F1 scores are measured for verifying the effectiveness of both the systems. The experimental results show that the ASR-LS-KWS method gives better results, compared to other methods, with precision rate of 0.79 and recall 0.75. Acoustic KWS gives higher false alarm rate and low F1 score. The experimental results also show that the performance of the system is improved when lattice search is combined with the ASR based keyword spotting system. The best F1 score 0.7766 is obtained for ASR-LS-KWS algorithm using RMM. From the experimental results it is evident that the best approach that can be used for the implimentation of Malayalam keyword spotting in continues speech is ASR based keyword spotting using word lattice search with relaxed matching method.

Chapter 6

Automatic Content based Classification of Speech Audio using Multiple Instance Learning Approach

6.1 Introduction

The keyword spotting system discussed in Chapter 5 finds exactly where a term was spoken in an audio sample. Speech analytics system based on KWS provides much better details of the speech content by extracting usefull information from it. As audio forms a major portion of information disseminated in the world every day many researchers are attempting to classify it based on various criteria [204]. In today's digital world people have access to a tremendous amount of news audio and video; on radio, television and the internet. The amount of multimedia data that are available is now so immense that it is infeasible for a human to go through it all and distinguish required files among them.

Automatic content based analysis provides useful information for audio classification as well as segmentation. Information about the audio

content can be used for classifying the file. Audio content understanding is thus an active research problem in speech analytics. A key step in this direction is a classifier that can predict the category of the input audio. In this chapter, a novel audio classification technique is proposed based on Multiple Instance Learning (MIL).

Different supervised and unsupervised learning algorithms are available in machine learning approaches. Multiple Instance Learning (MIL) is proposed as a variant of supervised learning for problems with incomplete knowledge about labels of training examples. Melih Kandeir *et al.* [86] conducted a benchmark study over the performance of different MIL methods. In their study, it is reported that mi-Graph and mi-SVM gave considerably better result compared to other MIL methods.

For the purpose of multimedia classification, features are drawn mainly from the text, audio, and visual modalities. Usually, multimedia approaches are found more often in the literature than text-only approaches. The audio content based approach usually requires fewer computational resources than visual methods [205]. There are different features which provide a compact representation of the given audio signal. Among them, Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear prediction (PLP) coefficients are widely used features [206].

In this work, a novel approach for content based speech audio classification using MIL methods is preposed. The experiments are conducted over the own developed news audio database MNAC familiaized in 5.2. The results obtained using these methods are evaluated using different performance metrics. The rest of this chapter is organized as follows. Section 6.2 describes feature extraction methods used in this study. Section 6.3 describes the proposed news audio classification methodology. Section 6.4 discusses the experimental results, and section 6.5 concludes the work.

6.2 Feature Extraction from News Audios for Classification

In this section audio feature extraction for MIL classifier is discussed in detail. To classify the news audio input, as a pre-processing the audio part is extracted and is split into overlapping segments. The signals are divided into constant-time segment of 25ms blocks [207]. Speech signal analysis is generally performed over short-time frames with a fixed frame length (FFL) and a fixed frame rate (FFR), based on the assumption that these signals are non-stationary and exhibit quasi-stationary behaviour in short durations [208]. This method benefits from the simplicity of implementation and the ease of comparing blocks of the same length.

Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) coefficients are extracted as features and further used for classification purpose. The algorithms used for MFCC and PLP based feature extraction techniques are described below.

6.2.1 Mel Frequency Cepstral Coefficient (MFCC) Feature Extraction

The most popular spectral based parameter used in recognition approach is the Mel Frequency Cepstral Coefficients called MFCC. MFCC takes human perception sensitivity with respect to frequencies under consideration, and therefore are the best for audio recognition [209]. Detailed implementation of MFCC is discussed in section 5.3.The procedure to determine the feature is described in algorithm 4.

Algorithm 4 MFCC feature extraction

1: Segmentation of voiced speech signal into 25 ms length frames.
2: Compute the spectral density of the power spectrum for each frame.
3: Apply the mel filterbank to the power spectra, sum the energy in each filter.
4: Compute the logarithm of all filterbank energies.
5: Compute the DCT of the log filterbank energies.
6: MFCCs are the amplitudes of the resulting spectrum.
The mel-scale frequency mapping is formulated as: $m(f) = 1125(1 + f/700)$

6.2.2 Perceptual Linear Prediction (PLP) Feature Extraction

The Perceptual Linear Prediction (PLP) model has been developed by Hermansky [210]. PLP models the human speech based on the concept of psychophysics of hearing [206]. PLP discards irrelevant information of the speech and thus improves speech recognition rate. The procedure to determine PLP coefficients are described in the algorithm 5.

Algorithm 5 PLP feature extraction

1: The N- point DFT is applied on the segmented input signal $x(n)$.
2: The critical-band power spectrum is computed through discrete convolution of the power spectrum with the piece-wise approximation of the critical-band curve.
3: Equal loudness pre-emphasis is applied on the down-sampled $\theta(B)$ and then intensity-loudness compression is performed.
4: Inverse DFT is performed for getting the equivalent autocorrelation function.
5: PLP coefficients are computed after autoregressive modelling and conversion of the autoregressive coefficients to cepstral coefficients.

6.3 Content based Audio Classification using MIL

This work aims for the automatic classification of news audios by catagorizing it based on the content. This categorisation will reduce the search cost of analytic applications. Given a list of news audios of interest, the proposed method will produce a discriminative model to distinguish them. In the following sections, implimentation details of the MIL approach have been discussed along with a description on the mi-Graph and mi-SVM methods which have been used for classification are explained in detail.

6.3.1 MIL for News Audio Classification

Initially, the input news audio files taken from the MNAC audio corpus are split into 25 ms length overlapping segments for feature extraction. Instances are created from each audio segments by extracting features from it. All the feature sets (the group of instances) belonging to the same news files are grouped into a bag. Labels are assigned for instances and for the bags as a whole, assuming the bag label to be the maximum of the instance labels within the bag. Finally, these bags along with their labels are fed into MIL classifier. A bag with a positive label indicates that there is at least one positively labelled instance and for a negatively labelled bag, all instances are known to have negative labels. Thus, as shown in figure 6.1, interested newsgroups are represented by the positive bag and other news sets by the negative bag. The schematic diagram of the proposed MIL based audio classification methodology is shown in figure 6.2.

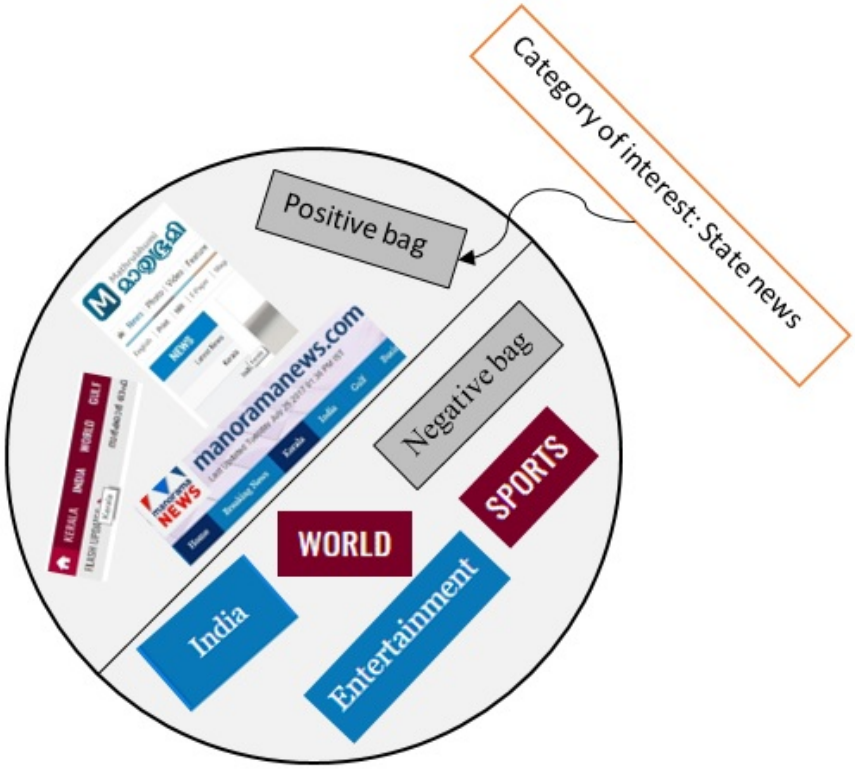


Fig. 6.1 MIL approach for news audio classification considering state news as the area of interest

Multiple Instance Learning (MIL) is a variation of supervised learning for problems with incomplete knowledge about labels of training examples. In MIL, the labels are assigned to bags of instances. The binary classifier labels a bag positive if no less than one instance in that bag is positive, otherwise bag is labelled as negative [211]. That is the MIL training set consists of bags $X_1, X_2, \ldots X_n$ and bag labels $y_1, y_2, \ldots y_n$, where $X_i = x_{i1}, x_{i2}, \ldots x_{im}$, $x_{ij} \in X$ and $y_i \in \{-1, 1\}$. The goal of MIL is to either train an instance classifier $h(X) : X \rightarrow Y$ or a bag classifier $H(X) : X_m \rightarrow Y$.

The brief description of two MIL based classification methods *viz.* mi-Graph and mi-SVM, used in this study are given in the following subsections.

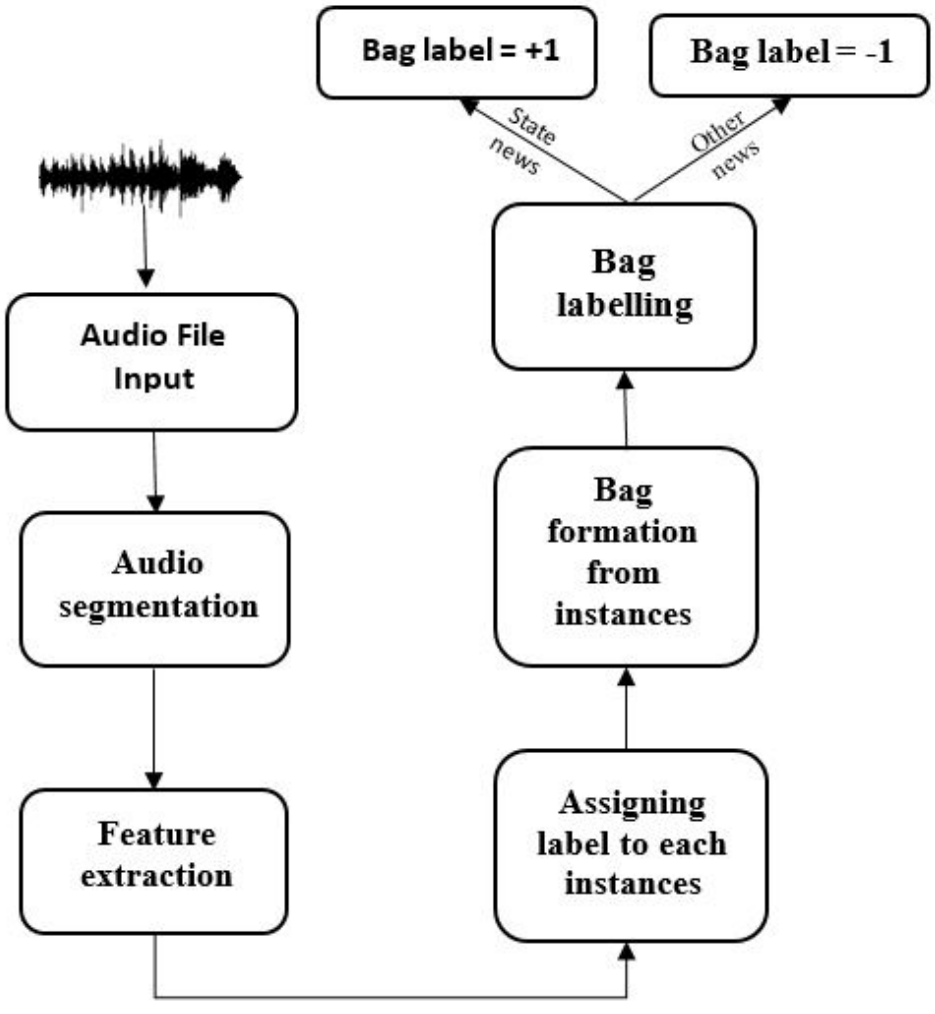


Fig. 6.2 Schematic diagram of proposed news audio classification methodology

6.3.2 mi-Graph based Classification Method

mi-Graph is a simple but effective method that represents each bag by a similarity graph [212]. initially, the cross-similarities of bag instances are computed by an instance-level kernel function $k_{inst}(x_i, x_j)$. A graph is then constructed by placing a node per each instance within a bag and each node pair is connected by an edge if the two corresponding instances are more similar to each other than a threshold δ. Let W_b

be the affinity matrix of bag b, whose entry is $w^b_{nm} = 1$, if there is an edge between the nodes of instances n and m, and $w^b_{nm} = 0$ otherwise. Consequently, similarity between bags b and c are computed by the following kernel function:

$$K_{bag}(X_b, X_c) = \frac{\sum_{n=1}^{N_b} \sum_{m=1}^{N_c} v_{bn} v_{cm} k_{inst}(X_{bn}, X_{cm})}{\sum_{n=1}^{N_b} \sum_{m=1}^{N_c} v_{cm}} \tag{6.1}$$

Where $v_{bn} = \frac{1}{\sum_{u=1}^{N_b} W^b_{nu}}$, $v_{cm} = \frac{1}{\sum_{u=1}^{N_c} W^c_{mu}}$ are the sum of the weights of the edges incident to nodes (instances) n and m of bags b and c, respectively. Based on the resultant bag-level Gram matrix, the arbitrary kernel learner is trained. The intuition behind this kernel is that for instances that are similar to a large number of other instances within the bag, W_{ia} has a smaller value, and for instances different from the rest of the bag, $W_i a$ is large. Hence, the influence of odd instances within bags are enhanced, and others are down weighted.

6.3.3 mi-SVM based Classification Method

This method approaches MIL as a semi-supervised learning problem, treating the labels of positive bag instances as latent variables [85]. These latent variables are added to the optimization problem that inferred from data.

$$\begin{aligned} &\min_{y} \min_{w,b,\xi} \frac{1}{2} \parallel W^2 \parallel + C \sum_{i=1}^{N} \xi_b, \\ &s.t. \quad y_i(W^T \phi(X_i)) \geq 1 - \xi_i, \forall i, \\ &\qquad \max^{\xi_i \geq 0, \forall i,} = Y_b, \forall b \end{aligned} \tag{6.2}$$

where w is the vector of model parameters defining the planar decision boundary, C is the regularization constant, ξ_b are slack variables, and $\phi(.)$ is a function that maps an instance from the original feature space to a Reproducing Kernel Hilbert Space (RKHS) [213]. At each iteration the approximate solution can be found as follows: train an instance-level standard SVM based on the current assignments of the latent variables, then update these variables by making predictions with the learned SVM. The following section presents the simulation experiments conducted for the news audio classification using MIL approach and summary of the result obtained.

6.4 Simulation Experiments and Results

The evaluation of the proposed MIL based audio classification is performed on MNAC news audio archive. The MFCC and PLP features and two MIL techniques *viz.* mi-Graph and mi-SVM have been used for the experiments. The experiment is conducted over the resultant audio samples obtained after the conduct of keyword spotting experiments. The evaluation of the proposed method is conducted by considering the news audio samples present in the dataset as two classes *viz.* state news audio and non-state news audio. Similarly, non-state news can be further categorized into different binary classes like national and non-national, sports and non-sports as well as news with cultural and non-cultural importance. The block diagram of the evaluation model for the preposed MIL based news audio classification is shown in figure 6.3. The keyword spotted audio files are given as an input to the MIL classifier. The MIL classifier classifies the audio file into either positive bag or negative bag. The file is considered as state news if it is labelled as positive. The details of the performance evaluations are discussed in this section.

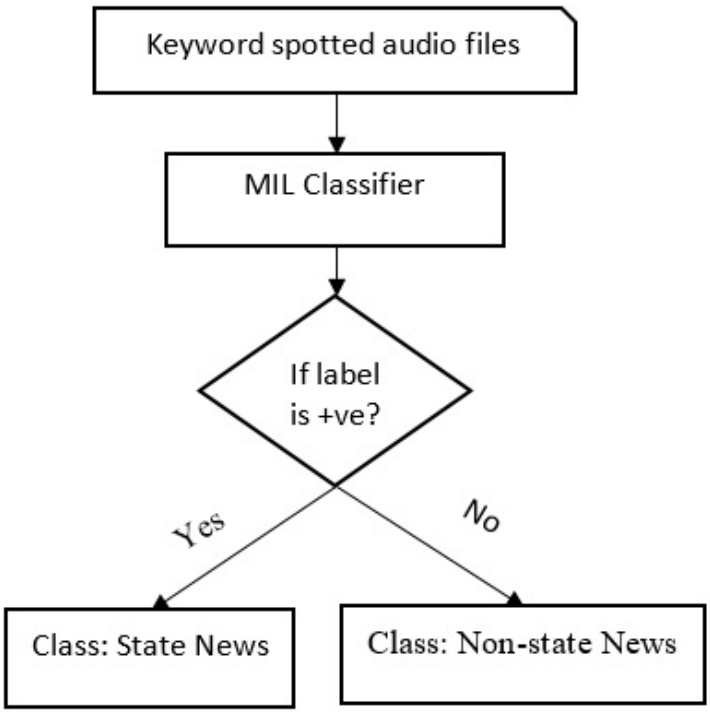


Fig. 6.3 Evaluation model for the MIL based news audio classifier

As the first stage the audio signals are segmented into 25 ms frames. Frames are considered as the instances of the audio signal. MFCC and PLP features have been extracted from each frame. Following four performance metrics are used for audio classification evaluation of the proposed MIL classifier.

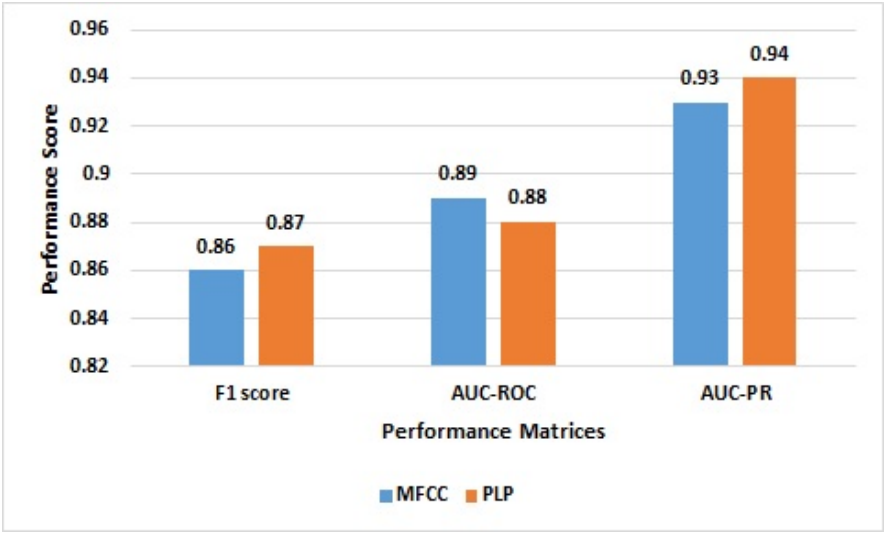
- Accuracy: Percentage of correctly classified test points.
- F1 score: Harmonic mean of precision and recall.
- AUC-ROC: Area under Receiver Operating Characteristics (ROC) curve.
- AUC-PR: Area under precision–recall curve.

The news audio classification experiments are conducted using MFCC and PLP features separately based on two different MIL techniques *viz.* mi-Graph and mi-SVM. The news audio classification results are performance matrices obtained by taking the state news as positive bags is given in table 6.1.

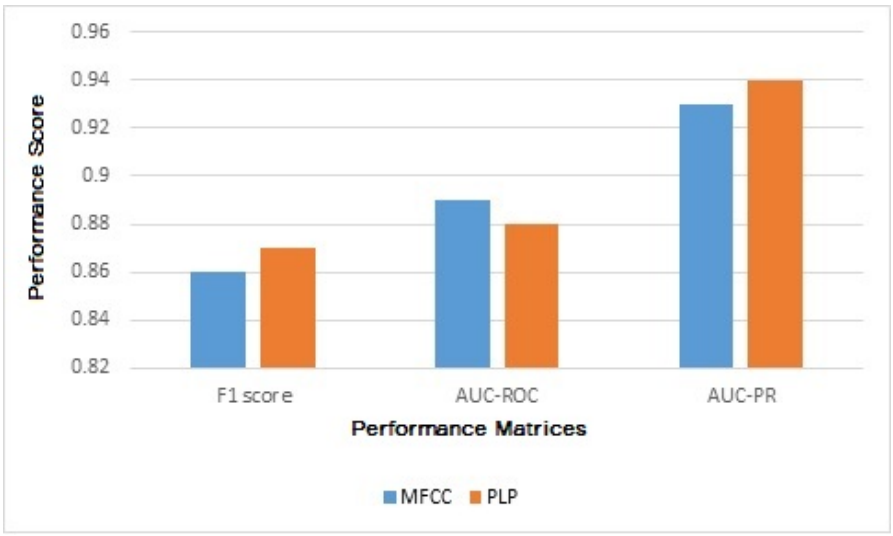
Table 6.1 MIL based news audio classification results and performance matrices

MIL method	**Feature**	**Accuracy (%)**	**F1 score**	**AUC-ROC**	**AUC-PR**
mi-Graph	MFCC	90.0	0.91	0.89	0.81
	PLP	85.0	0.86	0.94	0.96
mi-SVM	MFCC	80.3	0.86	0.89	0.93
	PLP	79.4	0.87	0.88	0.94

From the experimental result it is evident that the MIL classification method works effectively in speech audio classification. It is also evident that mi-Graph with MFCC feature give better result compared to other methods. Figure 6.4 shows the graphical representation of the performance score obtained for mi-graph and mi-SVM based audio classification.



(a) mi-Graph



(b) mi-SVM

Fig. 6.4 Performance scores for (a) mi-Graph (b) mi-SVM based news audio classification

6.5 Conclusion

In this study, a novel method for content based audio classification using MIL approach is presented. The news audio files taken from the indigenous MNAC audio dataset are classified using mi-Graph and mi-SVM

techniques. mi-Graph directly models within bag instance relationships and mi-SVM is semi-supervised in its nature. The news audio classification experiments are conducted using MFCC and PLP features. Performance evaluation of the proposed mi-Graph and mi-SVM methods using MFCC and PLP parameters are also carried out. mi-Graph using MFCC features appears as the best-performing method with 90.0% audio classification accuracy and 0.91 F1 score which is comparative with the other audio classification results reported earlier. Many audio, multimedia and speech analytics applications would certainly benefit from the ability of the proposed MIL based audio classifier to classify and retrieve audio samples into different categories based on its content.

Chapter 7

Effective Speaker Spotting based on Nonlinear Properties of Vocal Tract

7.1 Introduction

Speech analytics is the process of analysing speech data to gather relevant information from it. Automatic extraction of speaker information is one among the major speech analytic application. A novel speaker spotting method discussed in this thesis can be effectively used for speaker specific short listing of KWS result. Speaker recognition can be broadly divided into two classes: Speaker Verification (SV) and Speaker Spotting or speaker identification. The basic objective of speaker spotting is to associate an identifier to the speech of an individual speaker which is different from all other unique speakers. Linear Time Invariant (LTI) modelling of speech signal is widely used in speaker recognition works.

Nonlinear methods for speech processing are also a rapidly growing area of research. Nonlinearities are routinely included in attempts to model the physical process of vocal cord vibration, which have focused on two or more mass models [89]. Observation of glottal waveform reinforce

this evidence, where it has been shown that this wave form can change shape at different amplitudes. Such change would not be possible in a strictly linear system where the wave form shape is unaffected by the amplitude changes. Nonlinear signal processing techniques have several potential advantages over traditional linear signal processing methodologies [90–94, 214, 215]. They are capable of recovering the nonlinear dynamics of signals of interest possibly preserving natural information. In this context, the Eigen values of the reconstructed phase space, capacity dimension, correlation dimension, Kolmogorov entropy and largest positive Lyapunov exponents extracted from the vowel phonemes of thirty five different speakers are analysed.

Lyapunov exponents related with a trajectory give a measure of the average rates of convergence and divergence of nearby trajectories [216]. Fractal dimension is a measure that quantifies the number of degrees of freedom and the extent of self-similarity in the attractor's structure [217]. Kolmogorov entropy measures the rate of information loss or gain over the trajectory [217]. These measures search for a signature of chaos in the observed time series. Since these measures quantify the structure of the underlying nonlinear dynamical system, they are prime candidates for feature extraction of a signal with strong nonlinearities.

In this work the speaker identity based on the non-linear properties of the power spectral measures of the speech samples are analysed. These features are normally not considered in any of the conventional feature extraction methods. The power spectral measures show interesting similarities with the theoretical chaotic models. The source and system are separated by cepstral method and power spectral measures are carried out. It is observed that there exists an exponential decay in the power spectrum of the speech samples as predicted by Lorenz and Rossler chaotic dynamical system models [218]. These different features are

combined to model each unique speaker. The speaker identification experiments are conducted based on the proposed features using Feed Forward Multilayer Perceptron (FFMLP) classifier simulated using the error back propagation learning algorithm. Section 7.2 describes the algorithm used for vowel segmentation from continues speech samples. Section 7.3 presents the nonlinear dynamics of the vocal tract together with the proposed nonlinear feature extraction methodologies are discussed. Section 7.4 describes speaker spotting experiments conducted based on nonlinear features and ANN and section 7.5 concludes the work.

7.2 Segmentation of Vowel Units from Continues Speech

As a first step, the vowel units are segmented from continues speech for feature extraction. A segmentation algorithm proposed by Natarajan, V. *et al.* is implemented for vowel region segmentation [219]. Segmentation experiment is conducted based on first two formant frequencies using Support Vector Machine (SVM). The proposed algorithm is composed of three stages. In the first stage, the input audio is segmented into 20 ms-long frames with a 5 ms shift, where formant frequencies for each frame is computed. In order to group the frames into Vowel/Consonant in phoneme level, a silence detection algorithm using spectral centroid and signal energy is proposed. In the second stage the formant frequencies for each frame is computed using the Linear Prediction Analysis. In the third stage each frame is identified as either vowel or consonant using the support vector machine. The segmented vowel units are then given as input to the proposed speaker spotting module. For verifying the effectiveness of the proposed method, five Malayalam short vowels

(അ/a/, ഏ/e/,ഇ/i/, ഒ/o/, ഉ/u/) are segmented from different news audio files spoken by different speakers.

7.3 Nonlinear Dynamics of Vocal Tract

In the case of a purely deterministic system, once its present state is fixed, all future states can be determined as well. Hence it is important to establish a vector space called phase space or state space for the system, such that specifying a point in the space specifies the state of the system and vice versa. Then the information about the dynamics of the system can be obtained by studying the various features of the corresponding phase space distribution. The state of a particle moving in one dimension is specified by its position (x) and velocity (v). Its phase space is a plane. On the other hand a particle moving in three dimensions would have a six dimensional phase space with three position and three velocity directions. In phase space, momentum can be used instead of velocities.

In the case of speech signal what we have is not a phase space object but a time series, a sequence of scalar measurements. We therefore have to convert the observations into state vectors. This is the important problem of phase space reconstruction, which is technically solved by the method of time delay embedding. One of the profound results established in chaotic theory is the Takens' embedding theorem [91]. Takens' theorem states that under certain assumptions, phase space of a dynamical system can be reconstructed through the time-delayed versions of the original scalar measurements. This new state space is commonly referred to in the literature as Reconstructed Phase Space (RPS), and has been proven to be topologically equivalent to the original phase space of the dynamical system.

Packard *et al.* have first proved the concept of phase space reconstruction in 1980 [220]. Soon after,Takens has shown that a delay-coordinate mapping from a generic state space to a space of higher dimension preserves the topology [221]. This theorem provides important theoretical justification for the use of RPS's for system identification and pattern classification. Because the topology of the RPS is identical to the topology of the underlying system's phase space, we can expect the shape and density of the RPS attractor to provide valuable information of the system that generates a signal. According to Takens' embedding theorem a reconstructed phase space can be produced for a measured state variable x_n, where $n = 1,2,3,4,\ldots N$, via the method of delays by creating vectors given by

$$X_n = [x_n \quad x_{n+r} \quad x_{n+2r}\ldots \quad x_{n+(d-l)r}] \tag{7.1}$$

where d is the embedding dimension and τ is the chosen time delay value. The row vector X_n, defines the single point in the RPS. The row vectors then can be compiled into a matrix called a trajectory matrix to completely define the dynamics of the system and create a reconstructed phase space as

$$X_d = \begin{bmatrix} x_1 & x_{1+r} & x_{1+2r} & \cdots & x_{1+(d-1)r} \\ x_2 & x_{2+r} & x_{2+2r} & \cdots & x_{2+(d-1)r} \\ x_3 & x_{3+r} & x_{3+2r} & \cdots & x_{3+(d-1)r} \\ \cdots & & & & \\ x_N & x_{N+r} & x_{N+2r} & \cdots & x_{N+(d-1)r} \end{bmatrix} \tag{7.2}$$

The Reconstructed Phase Space (RPS) constructed for the isolated vowel phoneme അ /a/ with embedding dimension d=2 and the time delay value $\tau = 1$ is shown in figure 7.1.

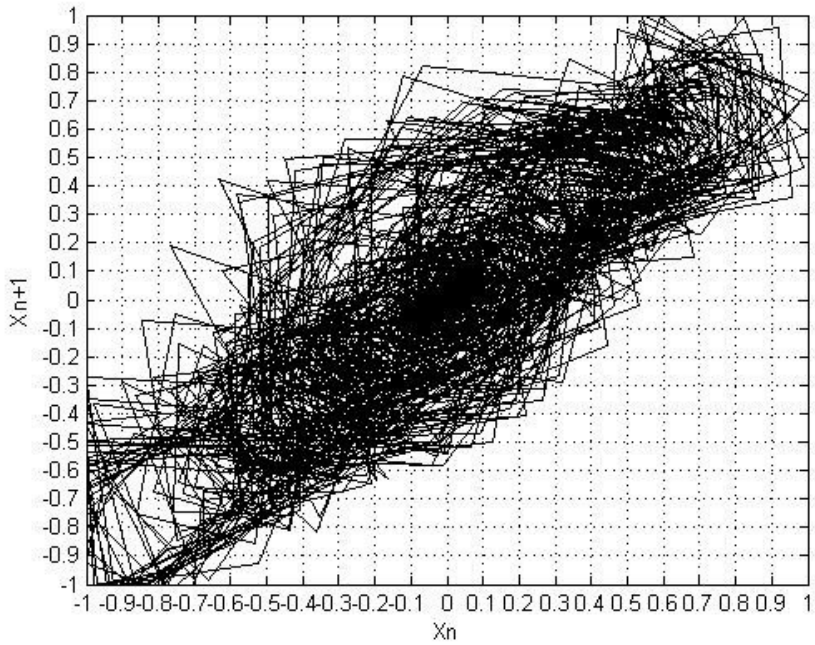


Fig. 7.1 Reconstructed Phase Space (RPS) for the vowel അ/a/ with $d = 2$ and $\tau = 1$

The following subsections describe various nonlinear features used in speaker modelling. Two novel nonlinear features, Eigen Value of Reconstructed Phase Space (RPC-EV) and Spectral Decay Coefficient (SDC) are proposed as part of this work.

7.3.1 Nonlinear Features used in Speaker Modelling

Nonlinear features used in this study for speaker modelling including Lyapunov exponent (λ_{max}), capacity dimension, correlation dimension ($D2$) and Kolmogorov entropy ($K2$) are discussed in this section.

a. **Lyapunov exponent**

The analysis of separation in time of two trajectories with infinitesimally close initial points is measured by Lyapunov exponents [216].

For a system whose evolution function is defined by a function f, we need to analyse:

$$\Delta x(t) \approx \Delta x(0) \frac{\partial}{\partial x}(f^n)\Delta x(0) \tag{7.3}$$

To quantify this separation, we assume that the rate of growth (or decay) of the separation between the trajectories is exponential in time. Hence we define the exponents, λ_i as:

$$\lambda_i \lim_{n\to\infty} \frac{1}{n} \ln((eig_i) \prod_{p=0}^{n} J(p)) \tag{7.4}$$

Where, J is the Jacobian of the system as the point p moves around the attractor. These exponents are invariant characteristics of the system and are called Lyapunov exponents, and can be computed by applying the above equation to points on the reconstructed attractor. The exponents read from a reconstructed attractor measure the rate of separation of nearby trajectories averaged over the entire attractor. Largest Lyapunov Exponent can be used as a feature value for uniquely modelling the speaker. Figure 7.2 shows the Lyapunov exponent obtained after repeated iteration for the vowel അ/a/.

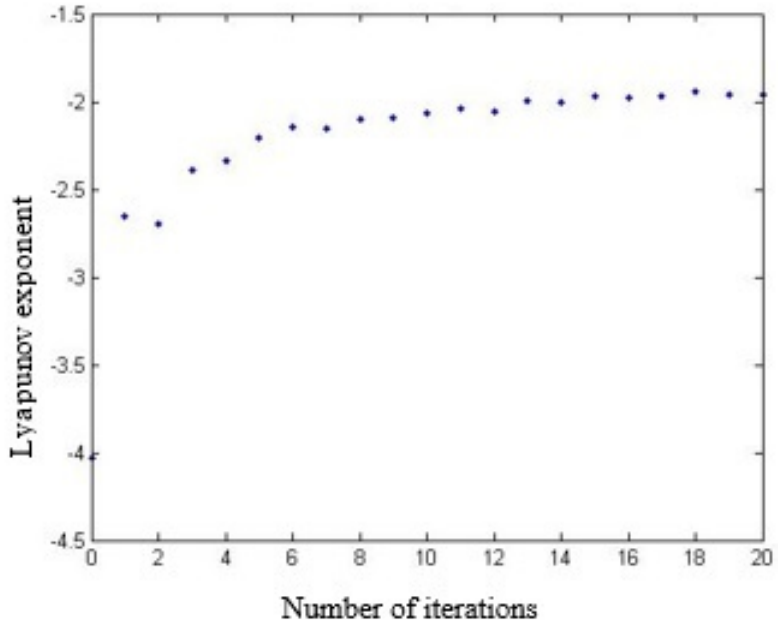


Fig. 7.2 Lyapunov exponents obtained for the Malayalam the vowel അ/a/

after repeated iteration

b. **Capacity dimension**

There are different ways to define the dimension, $d(A)$, of a set A. One approach is the capacity dimension dB. For a one dimensional figure such as straight line or curve of length L, it can be covered by $N(\varepsilon)$ one dimensional boxes of size ε. The capacity dimension is defined as,

$$dB \approx \frac{log(\varepsilon)}{log(\frac{1}{\varepsilon})} \tag{7.5}$$

To compute the capacity dimension using box counting method, break up the embedding space into a grid of boxes of size ε. Then count the number of boxes $N(\varepsilon)$ inside which at least one point of the attractor lies.

c. **Correlation dimension**

Fractals are objects which are self-similar at various resolutions.

Self-similarity in a geometrical structure is a strong signature of a fractal object. Correlation dimension [218] is a popular choice for numerically estimating the fractal dimension of the attractor. Correlation dimension d_C measures the variation of average fraction of neighbouring points with distance. The correlation integral $C_d(R)$ in d dimensional space is given by

$$C_d(R) = \lim_{N\to\infty}[\frac{1}{N^2}\sum_{i,j=1}^{N} H(R- \mid X_i - X_j \mid)] \tag{7.6}$$

Where X_i and X_j are points on attractor, $H(y)$ is the Heaviside function, N is the number of point randomly chosen from the data set. The Correlation dimension d_C is the variation of $C_d(R)$ with R.

$$d_c = \lim_{R\to 0} \log \frac{[C_d(R)]}{\log R} \tag{7.7}$$

d. **Kolmogorov entropy**

Entropy is a well-known measure used to quantify the amount of disorder in a system. It has also been associated with the amount of information stored in general probability distributions. Numerically, the Kolmogorov entropy can be estimated as the second order Kolmogorov entropy (K_2) and can be related to the correlation integral of the reconstructed attractor [218] as:

$$C_d(\varepsilon) \sim \lim_{\substack{\varepsilon\to 0 \\ d\to\infty}} \varepsilon^D exp(-\tau D K_2) \tag{7.8}$$

Where D is the fractal dimension of the system's attractor, d is the embedding dimension and τ is the time-delay used for attractor reconstruction. This leads to the relation:

$$K_2 \sim \frac{1}{\tau} \lim_{\substack{\tau \to 0 \\ d \to \infty}} \ln \frac{C_d(\varepsilon)}{C_{d+1}(\varepsilon)} \tag{7.9}$$

In practical, the values of ε and d are restricted by the resolution of the attractor and the length of the time series.

7.3.2 Eigen Value of Reconstructed Phase Space

The computation of the novel nonlinear feature, Eigen value of reconstructed phase space proposed in this study is discussed in this section. An Eigenvector of a square matrix A is a non-zero vector V that, when the matrix is multiplied by V, yields a constant multiple of V, the multiplier being commonly denoted by λ. That is:

$$AV = \lambda V \tag{7.10}$$

The number λ is called the eigenvalue of A corresponding to the Eigenvector V. The Eigenvalues can be used for dimensionality reduction [222]. The Eigenvalues of Reconstructed Phase Space (RPS-EV) are computed for different d values [223]. A detailed analysis of eigenvalues obtained in different dimensions of RPS are conducted to identify the optimum d value. RPS constructed for Malayalam vowel അ/a/ spoken by five different speakers ($d = 3$) is shown in figure 7.3. From the figure, it is evident that the RPS of the same vowel spoken by different speakers has a significant variation which can be effectively used for speaker modeling. Figure 7.4 shows the proposed RPS-EV feature vector (normalized) of different samples of Malayalam vowel അ /a/ (computed from RPS fixing d=5). Figure 7.5 shows the RPS-EV feature vectors extracted for $d = 5$ from 5 different vowels uttered by five different speakers.

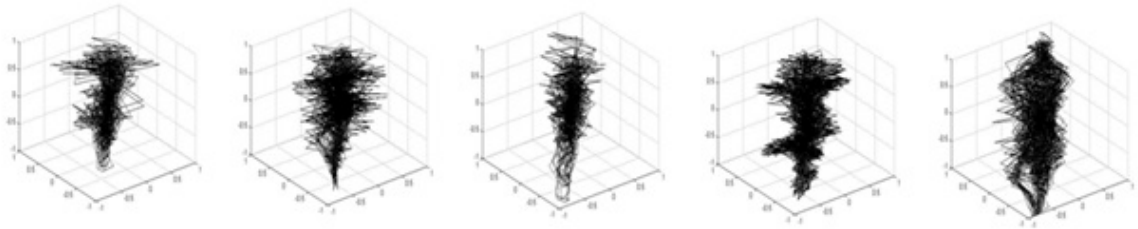


Fig. 7.3 Reconstructed Phase Space (RPS) for vowel അ/a/ spoken by five different speakers (d=3)

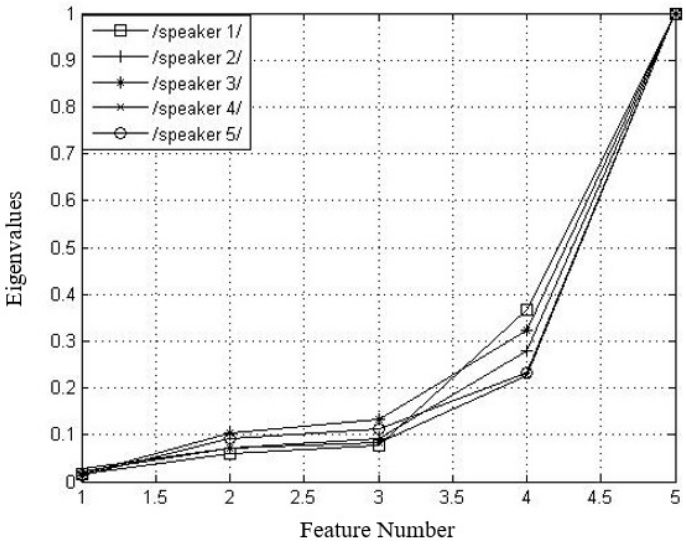


Fig. 7.4 RPS-EV feature vector $(d = 5)$ for the Malayalam vowel അ/a/ uttered by five different speakers

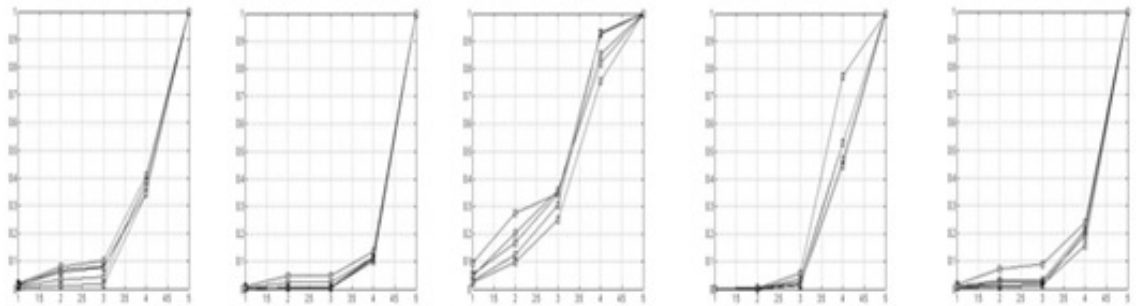


Fig. 7.5 Normalized RPS-EV feature vector $(d = 5)$ extracted from five different vowels uttered by five different speakers.

the following section describes the proposed nonlinear speaker modeling based on chaotic properties of the power spectrum in detail.

7.3.3 Speaker Modelling based on Chaotic Properties of the Power Spectrum

The power density spectrum of the speech signal are analysed for the purpose of speaker modelling [224]. The problem addressed in this work concerns the investigation of novel acoustical modelling techniques that exploits the theoretical result of nonlinear dynamics [218].

7.3.3.1 Evidence of Chaos in Power Spectrum

Power Spectral Density (PSD) is the frequency response of a random or periodic signal. It indicates where the average power is distributed as a function of frequency. M.C. Valsakumar *et al.* have focused on the analysis of the spectral measure, the tools employed in the study of stationary stochastic processes, as well as deterministic dynamical systems [218]. It is also known that a periodic system with frequency f has a pure point spectral measure with a spectral density (power spectrum) constituted by sharp (delta) peaks at f, and all its harmonics in general. On the other hand, the spectral measure of a generic stationary stochastic process has an absolutely continuous part and a corresponding broad power spectrum. They have also observed that the power spectrum of a chaotic time series differs in some way or the other from that of a stochastic time series, especially at high frequencies.

In an another study D.E.Sigeti have shown numerically that the power spectra of time series extracted from continuous time chaotic dynamical systems exhibit an exponential decay at high frequencies [225]. In view of this results, M.C. Valsakumar *et al.* have further investigated the

power spectra of a variety of chaotic dynamical systems including the Rossler [226], Raman [227], and Anantha [228] oscillators, the Lorenz model [229], the quasi-periodically kicked oscillator [215] and the Lorenz intermittency model [230], and presented their critical remarks about the theoretical arguments in the literature for the exponential decay of the power spectrum present at high frequencies for chaotic systems. The power spectra of chaotic dynamical systems are found to exhibit an exponential decay followed by a much slower decay (resembling an algebraic decay). The power spectra of two representative models exhibiting chaos, namely, the Lorenz and Rossler models, used in this study for investigating the chaotic property of the power spectrum of speech signal, are explained in the following session.

a. Lorenz and Rossler models

The Lorenz system is a system of ordinary differential equations (the Lorenz equations) first studied by Edward Lorenz in 1963 [229]. It is notable for having chaotic solutions for certain parameter values and initial conditions. The Rossler system is a system of three non-linear differential equations originally studied by Otto Rossler [226]. Otto Rossler has designed the Rossler attractor in 1976, and the equations are later found to be useful in modelling equilibrium in chemical reactions. The defining equations of the Lorenz and Rossler system are shown in Table 7.1.

Table 7.1 The defining equations of the Lorenz and Rossler systems

Model	**Equation**	**Parameters**	**Route**
Rossler	$\frac{dx}{dt} = -y - z$ $\frac{dy}{dt} = x + ay$ $\frac{dz}{dt} = b + z(x - c)$	a = 0.15 b= 0.2 c =10.00	Period doubling
Lorenz	$\frac{dx}{dt} = \rho(y - x)$ $\frac{dy}{dt} = x(r - z) - y$ $\frac{dz}{dt} = xy - cz$	ρ=16.00 r= 45.92 c=4	Quasiperiodic

Figure 7.6 shows the power spectrum obtained for the Lorenz and Rossler chaotic models [218]. From the figure it is evident that the power spectra of these models exhibit an exponential decay followed by a slower decay and the exponential decay constant derived from this could be identified as an important attribute of the chaotic system. The following section describes the speaker modelling based on the Spectral Decay Coefficient (SDC) extracted from the power spectrum of the speech signal.

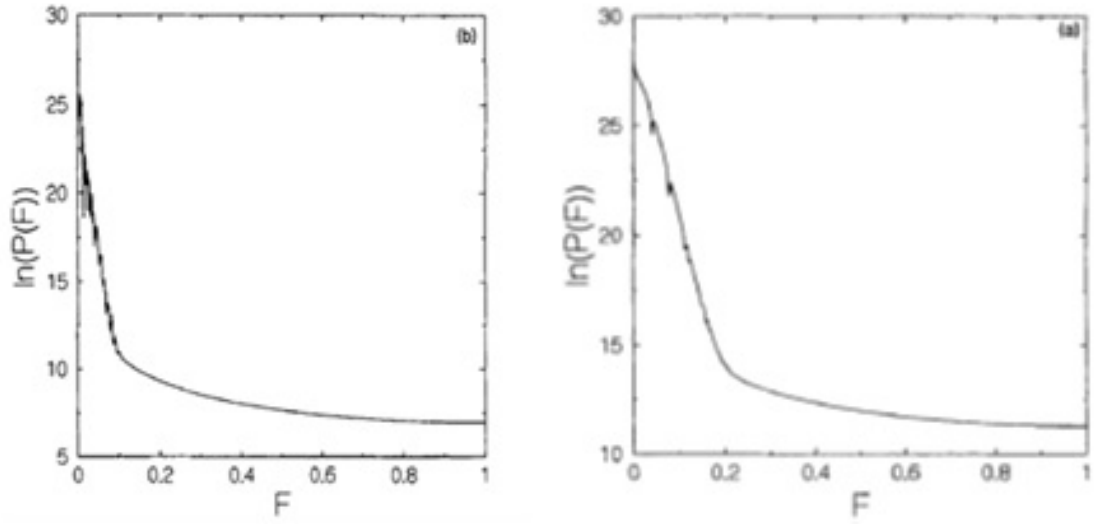


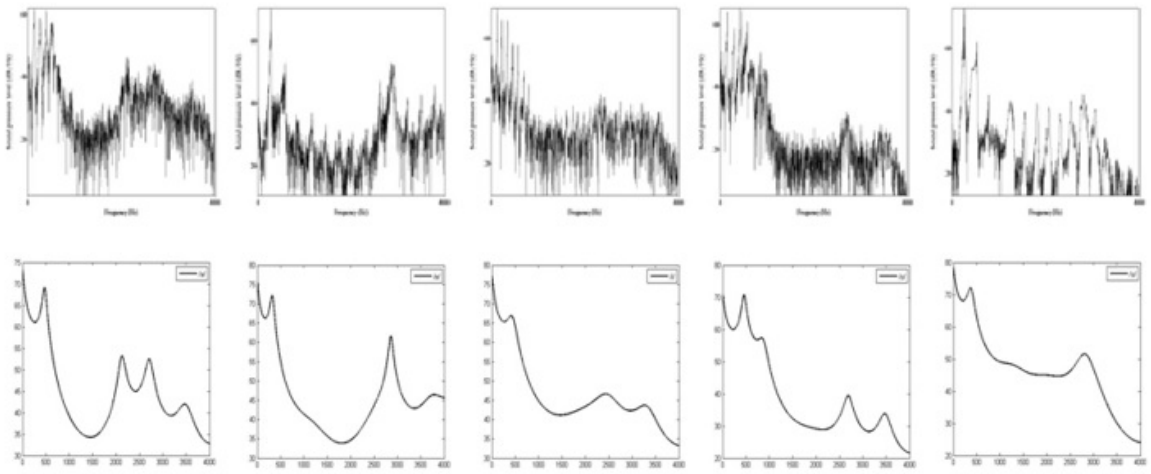
Fig. 7.6 Computed Power spectrum (a) Lorenz model and (b) Rossler model

7.3.3.2 Speaker Modelling based on Spectral Decay Coefficient (SDC) Extracted from the Power Spectrum

The power spectrum analysis (focused on chaotic properties) presented in this section pointed to the fact that the power spectrum of a chaotic system exhibits an exponential decay at high frequencies and the decay coefficient characterizes the system. As the speech production system possesses chaotic nature, it is decided to analyse the power spectrum of speech signal so as to extract certain nonlinear features which can be used for speaker modelling.

An analysis on the nonlinear properties of the power spectrum is conducted on five Malayalam short vowels (അ/a/, എ/e/, ഇ/i/, ഒ/o/, ഉ/u/) taken from the dataset. The power spectrum of the speech signals is then computed using Discrete Fourier Transform (DFT) and compared against the vowel samples spoken by different speakers. For further examination, the power spectrum is smoothed using an LPC based smoothing technique.

The power spectrum extracted for each vowel from five Malayalam short vowel spoken by a single speaker and the corresponding LPC smoothed power spectrum are shown in Figure 7.7. From the figure it is evident that the spectra is distinct for each vowel and there exist an exponential decay in each case.



(a) Vowel /a/ (b) Vowel /e/ (c) Vowel /i/ (d) Vowel /o/ (e) Vowel /u/

Fig. 7.7 Original power spectra (upper row) and the corresponding LPC smoothed power spectra (lower row) of five short vowels (/a/, /e/, /i/, /o/ and /u/) spoken by a single speaker

Figure 7.8 shows the LPC smoothed power spectrum of Malayalam vowels അ /a/ uttered by five different speakers and Figure 7.9 shows the power spectrum of the different samples of the vowel അ /a/ spoken by a single speaker. The visual analysis of the power spectrum shows that the exponential decay corresponding to a particular vowel spoken by a single speaker is similar in nature and decay part for the same vowel spoken by different speakers differs considerably.

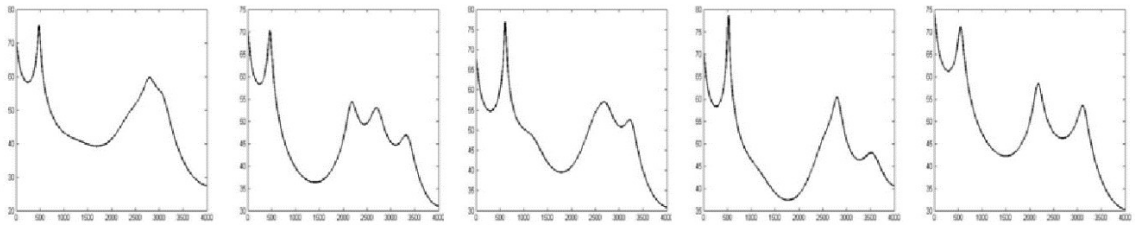


Fig. 7.8 LPC smoothed power spectra for vowel അ /a/ spoken by five different speakers

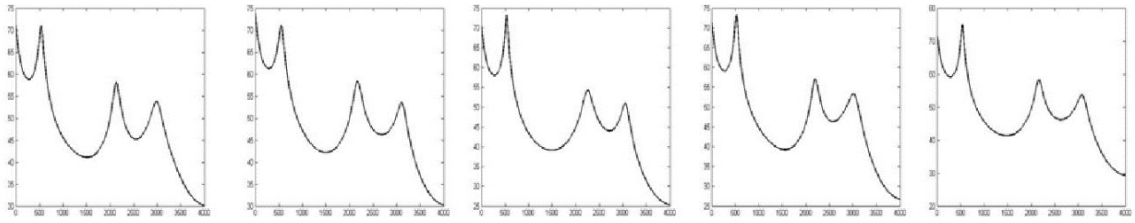


Fig. 7.9 LPC smoothed power spectra of different samples of vowel അ /a/ spoken by a single speaker

The LPC smoothed power spectrum extracted for the vowel അ/a/ and the corresponding exponential fit are shown in Figure 7.10. Further the Spectral Decay Coefficients (SDC) are extracted and used as feature for speaker recognition.

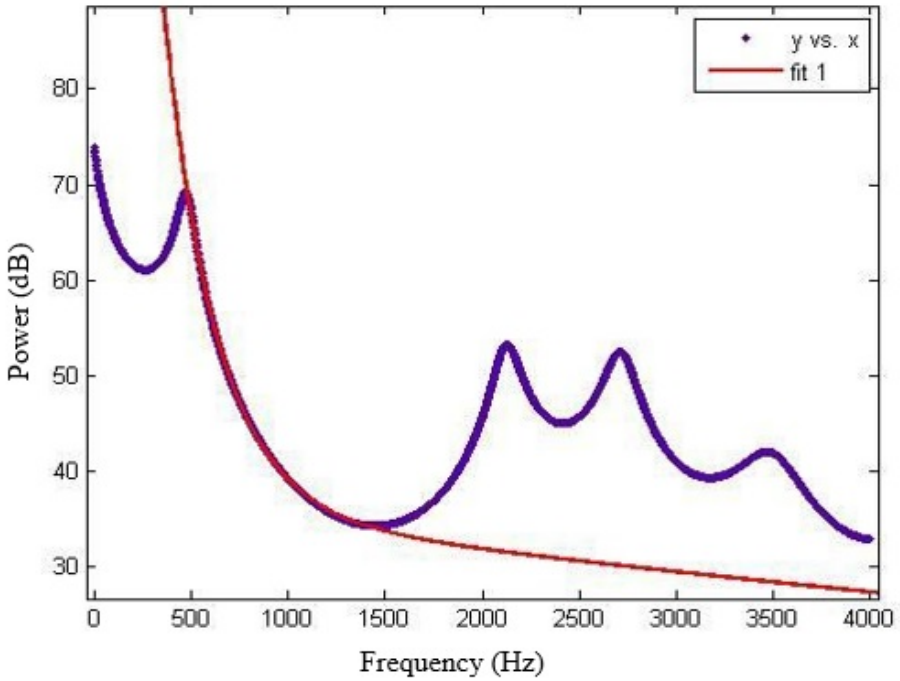


Fig. 7.10 Exponential fit over the LPC smoothed power spectrum for vowel അ/a/.

We have modelled the speaker identity using the Spectral Decay Coefficients (SDC) extracted from the speech power spectrum extracted from the vowel phonemes. An exponential curve fitting method is implemented for the extraction of these coefficients using the equation,

$$f(x) = ae^{-bx} + ce^{-dx} \tag{7.11}$$

The exponential decay always occurs at the high frequency part of the power spectrum and hence the curve fitting is also done on that particular portion of the power spectrum. The spectral decay coefficients (SDC) corresponding to the constants a, b, c and d denoted in the equation 7.11 are then extracted. The following section describes the speaker spotting experiments conducted based on the proposed nonlinear features using Artificial Neural Network (ANN.

7.4 Speaker Spotting Experiments based on Nonlinear Features and ANN

The speaker spotting experiments are conducted based on different samples of the five Malayalam vowel (അ /a/, എ /e/, ഇ /i/, ഒ /o/, ഉ /u/) segmented from the MNAC audio archive. This dataset contains vowel samples spoken by 35 different speakers (30 utterances of each vowel from every speaker). The speaker spotting experiments are simulated as an independent module. The Capacity Dimension (CD), the Correlation Dimension (CRD), Kolmogorov Entropy (KE) and Largest Lyapunov Exponent (LLE) are extracted from all the samples taken from the vowel database. Eigenvalues from the Reconstructed Phase Space (RPS) over different dimensions are computed from a total of 5,250 vowel samples and analysed thoroughly. Thus the experiments are also conducted to examine the speaker identification capabilities of the proposed Spectral Decay Coefficients (SDC). The SDCs are extracted from speech samples of five Malayalam vowels taken from the dataset.

The speaker identification experiments are conducted based on the proposed features using Feed Forward Multilayer Perceptron (FFMLP) Classifier simulated using the error back propagation learning algorithm. The number of input layers is fixed according to the feature vector size and five output nodes are fixed to represent five vowel classes. A constant learning rate 0.01 is used with the initial random weights obtained by generating random numbers ranging from 0.1 to 1. The recognition experiments are repeated by changing the number of hidden layers and number of nodes in each hidden layer for obtaining the successful architecture.

Initially, the network is trained using RPC- EV and SDC features extracted from the vowel samples. A set of 2,625 samples (15 samples

corresponding to five vowels) spoken by 35 different speakers are used for iteratively computing the final weight matrix (training) and a disjoint set of vowels of same size from the vowel dataset is used for speaker spotting purpose (testing) phase. The experiments are conducted in three stages. Initially, the SDC feature is used as parameters to train the system. In the second stage RPC- EV feature combined with SDC is used to form the parameter set to conduct the speaker spotting experiment. Finally, the experiment is also repeated by adding the nonlinear features including capacity dimension, the correlation dimension, Kolmogorov entropy and Lyapunov exponent along with SDC and RPC-EV feature set. The speaker spotting accuracy obtained for fifty different speakers based on above said features extracted from each of the five vowels using FFMLP classifier are tabulated in table 7.2. The experimental results indicate that the combined feature approach provides better speaker identification accuracy and hence substantiate the result substantiates the use of nonlinear features in speaker modeling.

Table 7.2 Speaker spotting Results

Parameters used	**Feature size**	**Speaker spotting accuracy (%) (for each vowel)**					**Accuracy (%)**
		അ/a/	**എ/e/**	**ഇ/i/**	**ഒ/o/**	**ഉ/u/**	
SDC	4	71.2	59.80	64.00	54.74	69.00	63.73
SDC + RPC - EV	9	86.6	63.45	71.00	67.00	77.74	71.17
SDC+RPC -EV+ LLE+ CD+CRD+KE	13	88.00	66.00	74.00	70.00	82.00	73.00

7.5 Conclusion

In this work, the speaker identity is modelled based on the non-linear properties of the speech samples which are normally not considered in the conventional feature extraction methodologies. The Eigen values of the reconstructed phase space, capacity dimension, correlation dimension, Kolmogorov entropy, largest positive Lyapunov exponent and spectral decay coefficient of selected phoneme for fifty different speakers are computed. The speaker recognition experiments are conducted using Multilayer Feed Forward Neural Networks using combined nonlinear features. The experiment results show that the speaker spotting capability of the proposed SDC parameter is enhanced when RPC-EV is used as an additional parameter. Further, it shows that the combined feature approach, in which the SDC features when used with RPC-EV, LLE, CD, CRD, and KE offer considerable improvement with an average speaker spotting accuracy of 73.00%. Future work can look at the effect of using linear and nonlinear feature merging technique to offer a robust practical speaker spotting system.

Chapter 8

Conclusions and Future Research Directions

8.1 Conclusion

The modern era has witnessed a drastic evolution of ASR technology. New research in speech processing is emerging in which ASR engines are being adopted for the development of different domain specific applications. Speech analytics is a similar newly emerging research area in which ASR engine is used for Keyword spotting (KWS) and hence to assist automated content based speech analysis. This thesis primarily addresses the problem of keyword spotting in Malayalam continuous speech for analytic applications.

Malayalam phonology and rule set for forming Malayalam allophones are studied in detail. In addition, this work has also analysed the allophonic variability of Malayalam vowels. An extensive statistical analysis is performed on the durational properties and the first two formant frequencies of the vowel allophones. Experiment results show that the characterisation of the allophonic variations in Malayalam vowel phones, obtained as part of the study can be efficiently used in automatic

speech processing applications. In this line, this work is the first of its kind in Malayalam language.

A rule-based Grapheme-to-Phoneme transcription for Malayalam is proposed. The transcriptor converts Malayalam text into a sequence of phonemes. The set of Malayalam graphemes is divided into six subsets, and separate processing routines are employed for the G2P transcription of each of the subsets. The proposed system transcripts text into Malayalam phonemes as well as to its corresponding International Phonetic Alphabet (IPA) representations. This tool can efficiently be used in the automatic speech processing applications including text-to-speech converter and speech recogniser. The phoneme occurrence frequency based on the word corpus (olam dataset) and sentence corpus (news text archive) are computed using the G2P conversion tool. The phoneme statistics derived out of the corpuses can be considered as a salient factor in designing language models for various speech processing applications.

The Malayalam continuous speech database has been created by recording the sentences collected from the online news portals of leading Malayalam dailies spoken by 35 male and female speakers of different age groups. For the purposes of the experiment, 5,250 sentences have been categorized into five classes, including state news, national news, international news, sports news and news related to cultural importance.

A discriminative method for detecting and spotting keywords in spoken utterances is preposed. The most widely used speech recognition algorithm, Hidden Markov Model (HMM) is used for the implementation of this method. The Keyword Spotting (KWS) system is implemented using two methods. The first one is ASR based Keyword spotting technique (ASR-KWS) and the second one is thefiller model based acoustic approach (FMA-KWS). In the KWS evaluation stage, two methods, Exact Matching Method (EMM) and Relaxed Matching Method (RMM)

are performed for evaluation. In EMM, the keyword form is recognized precisely, where as in RMM mismatch of inflected form is allowed. Precision, Recall and F1 scores are measured for verifying the effectiveness of both the systems. From the experimental results it is observed that the ASR-KWS gives better results compared to the FMA-KWS with precision as 0.79 and recall as 0.75. FMA-KWS gives higher false alarm rate and low F1 score. The experimental results also show that the efficiency of system is improved when lattice search is combined with the ASR based keyword spotting system. The best F1 score of 0.7766 is obtained for the relaxed matching mode based ASR-KWS algorithm with lattice search.

A novel approach for the content based audio classification using Multiple Instance Learning approach is implemented. There are two types of features used for audio content detection, namely, MFCC and PLP. Two MIL techniques mi-Graph and mi-SVM are used for classification purposes. The results obtained using these methods are evaluated using different performance metrics. From the experimental result, it is evident that the MIL works excellently on the audio classification. .It is also noted that mi-Graph which uses MFCC features appears as the best-performing method with 90.0% classification accuracy and 0.91 F1 score.

A novel method for speaker spotting is proposed with prominence on nonlinear speaker modelling techniques. The speech signal is non-stationary in nature as it is produced from a time-varying vocal tract system with time-varying excitation. However, most of the signal processing algorithms like LPC and MFCC model speech as a linear time-invariant system. In this work we make an attempt to model the vocal tract using different nonlinear features including Eigen values of the reconstructed phase space, capacity dimension, correlation dimension,

Kolmogorov entropy, largest positive Lyapunov exponent and spectral decay coefficients. The speaker spotting experiment is simulated using Artificial Neural Network. The experimental results indicate that the proposed keyword spotting approach provides an F1 score of 0.77. This module can be further used for speaker specific short listing of keywords.

8.2 Contributions

The major contributions to the field of digital speech processing as well as speech analytics that has been reported as part of this research work are detailed below.

An inclusive Malayalam phoneme dataset has been developed working in collaboration with the TEMU Malayalam phonetic archive project owned by Thunchath Ezhuthachan Malayalam University (TEMU), Kerala. The Malayalam phoneme segments are recorded in its standardized orthography at studio environment followed by a number of examples of its occurrence in phonologically relevant different positions. Allophones are listed together and pronunciation of each example recorded from the natural speech is demonstrated in both male and female voices. An extensive Malayalam speech dataset is also created by recording news sentences collected from the online news portals of the leading Malayalam dailies. A total of 5,250 spoken sentences are properly categorized and labelled to keep in the dataset along with its transcripted form.

An extensive study of Malayalam phonology with allophonic variation in vowel phonemes based on their durational and spectral characteristics are conducted. This work could be considered as a first step towards a paradigm shift to allophone-based Malayalam speech processing. A comprehensive rule-based Grapheme-to-Phoneme (G2P) transcriptor for Malayalam is designed and implemented. The transcription tool can be

effectively used in the automatic speech processing applications including text-to-speech converter and speech recognizer. Phoneme frequencies based on word and sentence corpuses are computed as a part of the performance evaluation of the proposed G2P transcription tool.

The development of a novel Keyword Spotting (KWS) system in Malayalam speech using continuous Hidden Markov Modelling is another major contribution of this dissertation. A KnowledgeBase Preparation Tool (KBPT-M) is designed as part of this work to generate language model and textual resourses required for accustic modeling from the given text corpus. The Automatic Speech Recognition based KWS (ASR-KWS) and Filler model based acoustic KWS (FMA-KWS) experiments are conducted. From the experimental results, it is observed that the ASR-based keyword spotting using lattice search outperforms other methods with a F1 score of 0.7766.

As part of the speech analytics work, a novel speech audio classifier based on Multiple Instance Learning is proposed. This classifier analyses the content of the news audios and classifies news audios under state news label from rest of the other categories.The outcome of the work can be used to identify the category of the output audio samples obtained from the KWS system. Another speech analytics work conducted is the speaker spotting performed on keyword spotted audios. An effective speaker spotting model is developed based on nonlinear properties of the vocaltract. Two novel nonlinear features RPS-EV and SDC have been proposed as part of this work. It is observed that the proposed method offer enormous improvement in speaker spotting with an average of 73% accuracy.

8.3 Future Direction

The research work provides a comprehensive study on KWS systems with application to speech analytics. However, there is scope for further research in this field and some of the future prospects are listed below.

The indigenous dataset used in the study has to be scaled up in such a way that all the dialectical variations of Malayalam are incorporated in the dataset, as dialect survey of Malayalam (1974) has identified 12 classes of dialect areas of Malayalam where each dialect class has its own subclasses.

The present study has analysed allophonic variation of Malayalam vowels based on its durational and spectral characteristics. Durational and spectral properties of consonant and diphthong allophones need to be explored and analysed in detail, as it is evident that the durational and spectral features of allophones can be effectively used to improve the performance of the ASR and speech synthesis systems designed for continuous speech.

One of the major contributions of this study is a rule based Grapheme to Phoneme (G2P) transcription algorithm for Malayalam. A Grapheme-to-Allophone transcriptor can also be developed to improve the performance of speech processing applications like speech synthesis as well as speech recognition.

A detailed analysis of the computational complexity of the proposed keyword spotting system is not reported in this thesis. It is an important metric that needs to be considered when implementing real-time tasks and applications. Hence a study on space and time complexity of the proposed algorithm is identified as another possible direction for future research work.

Finally, yet importantly, approaches differing from GMM-HMM-based acoustic modelling, such as Deep Neural Network (DNN) based framework can also be considered. DNN-based acoustic models are discriminative models which have a completely different model adaptation framework from generative models. Implementing KWS as well as speaker spotting process in this new framework may improve the system performance.

In this work, two speech analytics applications, *viz.* audio classifier and speaker spotting techniques are discussed in detail. The development of KWS based speech analytic applications in the field of speaker separation, emotion detection, sentiment analysis and audio summarization may also be explored.

References

[1] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[2] KH Davis, R Biddulph, and Stephen Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952.

[3] Harry F Olson and Herbert Belar. Phonetic typewriter. *The Journal of the Acoustical Society of America*, 28(6):1072–1081, 1956.

[4] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. The microsoft 2016 conversational speech recognition system. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5255–5259. IEEE, 2017.

[5] Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, and Yoshua Bengio. A network of deep neural networks for distant speech recognition. *arXiv preprint arXiv:1703.08002*, 2017.

[6] Cini Kurian and Kannan Balakriahnan. Continuous speech recognition system for malayalam language using plp cepstral coefficient. *Journal of Computing and Business Research*, 3(1), 2012.

[7] Sonia Sunny, S David Peter, and K Poulose Jacob. Development of a speech recognition system for speaker independent isolated malayalam words. *International Journal of Computer Science & Engineering Technology*, 3(4):69–75, 2012.

[8] VR Vimal Krishnan, Athulya Jayakumar, and Anto P Babu. Speech recognition of isolated malayalam words using wavelet features and

artificial neural network. In *Electronic Design, Test and Applications, 2008. DELTA 2008. 4th IEEE International Symposium on*, pages 240–243. IEEE, 2008.

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

[10] K. Saumudravijaya. Hindi speech recognition. *J. Acoustic Society India*, 29(1):385–95, 2001.

[11] S Rajendran and B Yegnanarayana. Word boundary hypothesization for continuous speech in hindi based on f0 patterns. *Speech communication*, 18(1):21–46, 1996.

[12] K Samudravijaya, R Ahuja, N Bondale, T Jose, S Krishnan, P Poddar, R Raveendran, et al. A feature-based hierarchical speech recognition system for hindi. *Sadhana*, 23(4):313–340, 1998.

[13] Mohit Kumar, Nitendra Rajput, and Ashish Verma. A large-vocabulary continuous speech recognition system for hindi. *IBM journal of research and development*, 48(5.6):703–715, 2004.

[14] Somnath Majumder and AK Dutta. Automatic recognition of isolated bengali words. In *Proc. Speech Workshop*, pages 10–13, 1990.

[15] P Vijai Bhaskar, S Rama Mohan Rao, and A Gopi. Htk based telugu speech recognition. *Int J Adv Res Comput Sci Softw Eng*, 2(12):307–314, 2012.

[16] M Chandrasekar and M Ponnavaikko. Tamil speech recognition: a complete model. *Electronic Journal «Technical Acoustics*, 20, 2008.

[17] MA Anusuya and SK Katti. Wavelet transform for noisy kannada speech recognition. *International Journal of Computational Intelligence Research*, 16(4), 2010.

[18] K Samudravijaya. Computer recognition of spoken hindi. *Training 198*, 56(2000):93.

[19] T.M. Thasleemaand N.K Narayanan V.L. Lajish. Isolated word–speech recognition using lexicon based hidden markov models.

Proc. of the IEEE – Int. Conf. on Signal and Image Processing, 2:988–991, 2006.

[20] T.M. Thasleemaand N.K Narayanan V.L. Lajish. Hidden markov models for isolated wordspeech recognition based on large vocabulary. *Proc. of the National Symposium on Acoustics,(NSA2006), NPL, New Delhi*, 1, 2006.

[21] Sreekanteswaram G PadmanabhaPillai. *Sabdhatharavali*. SahithyaPravarthakaSahakarana-Sangham, 1983.

[22] Piotr Kozierski, Talar Sadalla, Szymon Drgas, and Adam Dąbrowski. Allophones in automatic whispery speech recognition. In *Methods and Models in Automation and Robotics (MMAR), 2016 21st International Conference on*, pages 811–815. IEEE, 2016.

[23] Long Nguyen, Xuefeng Guo, and John Makhoul. Modeling frequent allophones in japanese speech recognition. In *Seventh International Conference on Spoken Language Processing*, 2002.

[24] Ji Xu, Yujing Si, Jielin Pan, and Yonghong Yan. Automatic allophone deriving for korean speech recognition. In *Computational Intelligence and Security (CIS), 2013 9th International Conference on*, pages 776–779. IEEE, 2013.

[25] FayÇal Imedjdouben and Amrane Houacine. Generation of allophones for speech synthesis dedicated to the arabic language. In *New Technologies of Information and Communication (NTIC), 2015 First International Conference on*, pages 1–4. IEEE, 2015.

[26] Pavel Skrelin. Allophone-based concatenative speech synthesis system for russian. In *Text, Speech and Dialogue*, pages 842–842. Springer, 1999.

[27] Wafa Barkhoda, Bahram ZahirAzami, Anvar Bahrampour, and Om-Kolsoom Shahryari. A comparison between allophone, syllable, and diphone based tts systems for kurdish language. In *Signal Processing and Information Technology (ISSPIT), 2009 IEEE International Symposium on*, pages 557–562. IEEE, 2009.

[28] Louis CW Pols, Xue Wang, and Louis FM ten Bosch. Modelling of phone duration (using the timit database) and its potential benefit for asr. *Speech Communication*, 19(2):161–176, 1996.

[29] Dennis H Klatt. Linguistic uses of segmental duration in english: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, 59(5):1208–1221, 1976.

[30] Yong-Ju Lee and Sook-hyang Lee. On phonetic characteristics of pause in the korean read speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 1, pages 118–120. IEEE, 1996.

[31] Omer Sayli. Duration analysis and modeling for turkish text-to-speech synthesis. *Master's thesis, Department of Electrical and Electronics Engineering, Bogaziei University*, 2002.

[32] Robert Batusek. A duration model for czech text-to-speech synthesis. In *Speech Prosody 2002, International Conference*, 2002.

[33] K Samudravijaya. Durational characteristics of hindi phonemes in continuous speech. *Technical Report, TIFR*, 2003.

[34] K Samudravijaya. Durational characteristics of hindi stop consonants. In *Eighth European Conference on Speech Communication and Technology*, 2003.

[35] K Sreenivasa Rao and B Yegnanarayana. Modeling syllable duration in indian languages using support vector machines. In *Intelligent Sensing and Information Processing, 2005. Proceedings of 2005 International Conference on*, pages 258–263. IEEE, 2005.

[36] N Sridhar Krishna and Hema A Murthy. Duration modeling of indian languages hindi and telugu. In *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[37] SR Savithri. Durational analysis of kannada vowels. *Journal of Acoustical Society of India*, 14(2):34–31, 1986.

[38] Deepa P Gopinath, J Divya Sree, Reshmi Mathew, SJ Rekhila, and Achuthsankar S Nair. Duration analysis for malayalam text-to-speech systems. In *Information Technology, 2006. ICIT'06. 9th International Conference on*, pages 129–132. IEEE, 2006.

[39] W Ainsworth. A system for converting english text into speech. *IEEE Transactions on audio and electroacoustics*, 21(3):288–290, 1973.

[40] Michel Divay and Marc Guyomard. Grapheme-to-phoneme transcription for french. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'77.*, volume 2, pages 575–578. IEEE, 1977.

[41] M Douglas McIlroy. Synthetic english speech by rule. *The Journal of the Acoustical Society of America*, 55(S1):S55–S56, 1974.

[42] H Elovitz, Rodney Johnson, Astrid McHugh, and J Shore. to-sound rules for automatic translation of english text to phonetics. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(6):446–459, 1976.

[43] Sue Hertz. Appropriateness of different rule types in speech synthesis. *The Journal of the Acoustical Society of America*, 65(S1):S130–S130, 1979.

[44] Rabia Belrhali, Véronique Aubergé, and Louis-Jean Boe. From lexicon to rules: toward a descriptive method of french text-to-phonetics transcription. In *Second International Conference on Spoken Language Processing*, 1992.

[45] Michel Divay and Anthony J Vitale. Algorithms for grapheme-phoneme translation for english and french: Applications for database searches and speech synthesis. *Computational linguistics*, 23(4):495–523, 1997.

[46] Askars Salimbajevs and Marcis Pinnis. Towards large vocabulary automatic speech recognition for latvian. In *Baltic HLT*, pages 236–243, 2014.

[47] Attila Novák and Borbála Siklósi. Grapheme-to-phoneme transcription in hungarian. *Int. J. Comput. Linguistics Appl.*, 7(1):161–173, 2016.

[48] Daniela Braga, Luís Coelho, and Fernando Gil Vianna Resende. A rule-based grapheme-to-phone converter for tts systems in european portuguese. In *Telecommunications Symposium, 2006 International*, pages 328–333. IEEE, 2006.

[49] Yu-Chun Wang and Richard Tzong-Han Tsai. Rule-based korean grapheme to phoneme conversion using sound patterns. In *PACLIC*, pages 843–850, 2009.

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

[51] Sumi S Nair, CR Rechitha, and C Santhosh Kumar. Rule-based grapheme to phoneme converter for malayalam. *International Journal of Computational Linguistics and Natural Language Processing*, 2(7):417–420, 2013.

[52] J Robin Rohlicek, William Russell, Salim Roukos, and Herbert Gish. Continuous hidden markov modeling for speaker-independent word spotting. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pages 627–630. IEEE, 1989.

[53] R Wohlford, A Smith, and M Sambur. The enhancement of wordspotting techniques. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'80.*, volume 5, pages 209–212. IEEE, 1980.

[54] Richard C Rose and Douglas B Paul. A hidden markov model based keyword recognition system. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 129–132. IEEE, 1990.

[55] Joseph Keshet, David Grangier, and Samy Bengio. Discriminative keyword spotting. *Speech Communication*, 51(4):317–329, 2009.

[56] Aravind Ganapathiraju and Ananth Nagaraja Iyer. Method and system for real-time keyword spotting for speech analytics, June 6 2017. US Patent 9,672,815.

[57] Zhehuai Chen, Yimeng Zhuang, Yanmin Qian, Kai Yu, Zhehuai Chen, Yimeng Zhuang, Yanmin Qian, Kai Yu, Kai Yu, Yimeng Zhuang, et al. Phone synchronous speech recognition with ctc lattices. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(1):90–101, 2017.

[58] Martha Larson. Sub-word-based language models for speech recognition: implications for spoken document retrieval. *Whorkshop on Language Modeling and Information Retrieval*, 2001.

[59] Wade Shen, Christopher M White, and Timothy J Hazen. A comparison of query-by-example methods for spoken term detection. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[60] Aren Jansen and Partha Niyogi. Point process models for spotting keywords in continuous speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8):1457–1470, 2009.

[61] Ami Moyal, Vered Aharonson, Ella Tetariy, and Michal Gishri. Keyword spotting methods. In *Phonetic Search Methods for Large Speech Databases*, pages 7–11. Springer, 2013.

[62] Igor Szöke, Petr Schwarz, Pavel Matejka, Lukás Burget, Martin Karafiát, Michal Fapso, and Jan Cernockỳ. Comparison of keyword spotting approaches for informal continuous speech. In *Interspeech*, pages 633–636, 2005.

[63] Rafid A Sukkar, Anand R Setlur, Mazin G Rahim, and Chin-Hui Lee. Utterance verification of keyword strings using word-based minimum verification error (wb-mve) training. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 518–521. IEEE, 1996.

[64] Eric D Sandness and I Lee Hetherington. Keyword-based discriminative training of acoustic models. In *Sixth International Conference on Spoken Language Processing*, 2000.

[65] Yassine Benayed, Dominique Fohr, Jean Paul Haton, and Gérard Chollet. Confidence measures for keyword spotting using support vector machines. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2003.

[66] Anupam Mandal, KR Prasanna Kumar, and Pabitra Mitra. Recent developments in spoken term detection: a survey. *International Journal of Speech Technology*, 17(2):183–198, 2014.

[67] John S Garofolo, Cedric GP Auzanne, and Ellen M Voorhees. The trec spoken document retrieval track: A success story. In *Content-Based Multimedia Information Access-Volume 1*, pages

1–20. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2000.

[68] Lidia Mangu, Eric Brill, and Andreas Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4):373–400, 2000.

[69] Ciprian Chelba and Alex Acero. Position specific posterior lattices for indexing speech. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 443–450. Association for Computational Linguistics, 2005.

[70] Yi-cheng Pan and Lin-shan Lee. Performance analysis for lattice-based speech indexing approaches using words and subword units. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1562–1574, 2010.

[71] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaten. Content-based classification, search, and retrieval of audio. *IEEE multimedia*, 3(3):27–36, 1996.

[72] Jonathan T Foote. Content-based retrieval of music and audio. In *Multimedia Storage and Archiving Systems II*, volume 3229, pages 138–148. International Society for Optics and Photonics, 1997.

[73] Silvia Pfeiffer, Stephan Fischer, and Wolfgang Effelsberg. Automatic audio content analysis. In *Proceedings of the fourth ACM international conference on Multimedia*, pages 21–30. ACM, 1997.

[74] Don Kimber, Lynn Wilcox, et al. Acoustic segmentation for audio browsers. *Computing Science and Statistics*, pages 295–304, 1997.

[75] Tong Zhang and C-C Jay Kuo. Video content parsing based on combined audio and visual information. In *Proc. SPIE*, volume 4, pages 78–89, 1999.

[76] Phung Quoc Dinh, Chitra Dorai, and Svetha Venkatesh. Video genre categorization using audio wavelet coefficients. *ACCV 2002*, 2002.

[77] Radu S Jasinschi and Jennifer Louie. Automatic tv program genre classification based on audio patterns. In *Euromicro Conference, 2001. Proceedings. 27th*, pages 370–375. IEEE, 2001.

[78] Thomas F Quatieri. *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2006.

[79] Zhu Liu, Jincheng Huang, and Yao Wang. Classification tv programs based on audio information using hidden markov model. In *Multimedia Signal Processing, 1998 IEEE Second Workshop on*, pages 27–32. IEEE, 1998.

[80] Erik Visser, Yinyi Guo, Lae-Hoon Kim, Raghuveer Peri, and Shuhua Zhang. Deep neural net based filter prediction for audio event classification and extraction, May 30 2017. US Patent 9,666,183.

[81] James D Keeler, David E Rumelhart, and Wee Kheng Leow. Integrated segmentation and recognition of hand-printed numerals. In *Advances in neural information processing systems*, pages 557–563, 1991.

[82] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1):31–71, 1997.

[83] Yixin Chen and James Z Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5(Aug):913–939, 2004.

[84] Jun Yang, Rong Yan, and Alexander G Hauptmann. Multiple instance learning for labeling faces in broadcasting news video. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 31–40. ACM, 2005.

[85] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*, pages 577–584, 2003.

[86] Melih Kandemir and Fred A Hamprecht. Computer-aided diagnosis from weak supervision: a benchmarking study. *Computerized Medical Imaging and Graphics*, 42:44–50, 2015.

[87] Lawrence George Kersta. Voiceprint identification. *Nature*, 196(4861):1253–1257, 1962.

[88] Jonathan Harrington and Steve Cassidy. *Techniques in speech acoustics*, volume 8. Springer Science & Business Media, 2012.

[89] H Teager. Some observations on oral air flow during phonation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(5):599–601, 1980.

[90] Petros Maragos, Thomas F Quatieri, and James F Kaiser. Speech nonlinearities, modulations, and energy operators. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 421–424. IEEE, 1991.

[91] Andrew C Lindgren, Michael T Johnson, and Richard J Povinelli. Speech recognition using reconstructed phase space features. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–60. IEEE, 2003.

[92] Michael T Johnson, Andrew C Lindgren, Richard J Povinelli, and Xiaolong Yuan. Performance of nonlinear speech enhancement using phase space reconstruction. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2003.

[93] Adriano Petry and Dante Augusto Couto Barone. Speaker identification using nonlinear dynamical features. *Chaos, Solitons & Fractals*, 13(2):221–231, 2002.

[94] Vassilis Pitsikalis and Petros Maragos. Speech analysis and feature extraction using chaotic models. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, pages I–533. IEEE, 2002.

[95] Niko Brümmer, Albert Swart, and David van Leeuwen. A comparison of linear and non-linear calibrations for speaker recognition. *arXiv preprint arXiv:1402.2447*, 2014.

[96] Shabnam Gholamdokht Firooz, Farshad Almasganj, and Yasser Shekofteh. Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals. *Computers & Electrical Engineering*, 58:215–226, 2017.

[97] Tejas Godambe and K Samudravijaya. Speech data acquisition for voice based agricultural information retrieval. In *Proc. Of 39th All India DLA Conference, Punjabi University, Patiala, June*, 2011.

[98] Ashish Verma Chalapathy Neti, Nitendra Rajput. A large vocabulary continuous speech recognitionfor hindi. *Proceedings of the National conference on Communications, Mumbai*, 56(2002):366–370.

[99] R Gupta and G Sivakumar. Speech recognition for hindi language. *IIT BOMBAY*, 2006.

[100] MT Bala Murugan, M Balaji, and B Venkataramani. Sopc-based speech-to-text conversion. *National Institute of Technology, Trichy*, 2006.

[101] R Mathur and Kansal Babita. "domain specific speaker independent continuous speech recognizer using julius". *Proceedings of ASCNT–2010, CDAC, Noida, India*, pages 55–60, 2010.

[102] Sunita Arora, Babita Saxena, Karunesh Arora, and SS Agarwal. Hindi asr for travel domain. *Proceedings of OCOCOSDA*, 2010.

[103] K Kumar and RK Aggarwal. Hindi speech recognition system using htk. *International Journal of Computing and Business Research ISSN (Online): 2229-6166*, 2(2), 2011.

[104] RK Aggarwal and M Dave. Using gaussian mixtures for hindi speech recognition system. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 4(4):157–170, 2011.

[105] A Mishra, A Biswas, M Chandra, and N Sharan. Robust hindi connected digit recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 4(2), 2011.

[106] K. Sivaraman. G.; Samudravijaya. Hindi speech recognition and online speaker adaptation. *International Conference on Technology Systems and Management: ICTSM-2011.*

[107] Kaushik S. Kumar R., Singh C. Isolated and connected word recognition for punjabi language using acoustic template matching technique. (2004).

[108] Wiqas Ghai and Navdeep Singh. Analysis of automatic speech recognition systems for indo-aryan languages: Punjabi a case study. *Int J Soft Comput Eng*, 2(1):379–385, 2012.

[109] Kumar Ravinder. Comparison of hmm and dtw for isolated word recognition system of punjabi language. In *Iberoamerican Congress on Pattern Recognition*, pages 244–252. Springer, 2010.

[110] A Nayeemulla Khan and B Yegnanarayana. Development of speech recognition system for tamil for small restricted task. In *Proceedings of national conference on communication*, number 3, 2001.

[111] S Saraswathi and TV Geetha. Building language models for tamil speech recognition system. In *Asian Applied Computing Conference*, pages 161–168. Springer, 2004.

[112] A Lakshmi and Hema A Murthy. A syllable based continuous speech recognizer for tamil. In *Ninth International Conference on Spoken Language Processing*, 2006.

[113] R Thangarajan, AM Natarajan, and M Selvam. Word and triphone based approaches in continuous speech recognition for tamil language. *WSEAS transactions on signal processing*, 4(3):76–86, 2008.

[114] R Thangarajan, AM Natarajan, and M Selvam. Syllable modeling in continuous speech recognition for tamil language. *International Journal of Speech Technology*, 12(1):47, 2009.

[115] VS Dharun and M Karnan. Voice and speech recognition for tamil words and numerals. *International Journal of Modern Engineering Research (IJMER)*, 2(5):3406–3414, 2012.

[116] V Radha et al. Speaker independent isolated speech recognition system for tamil language using hmm. *Procedia Engineering*, 30:1097–1102, 2012.

[117] S Karpagavalli, KU Rani, R Deepika, and P Kokila. Isolated tamil digits speech recognition using vector quantization. *International Journal of Engineering Research and Technology*, 1(4):1–12, 2012.

[118] A Akila and E Chandra. Isolated tamil word speech recognition system using htk. *International Journal of Computer Science Research and Application*, 3(2):30–38, 2013.

[119] Manash Pratim Sarma and Kandarpa Kumar Sarma. Speech recognition of assamese numerals using combinations of lpc-features and heterogenous anns. In *International Conference on Advances in Information and Communication Technologies*, pages 8–12. Springer, 2010.

[120] Manash Pratim Sarma and Kandarpa Kumar Sarma. Assamese numeral speech recognition using multiple features and cooperative lvq-architectures. *International Journal of Electrical and Electronics*, 5:1, 2011.

[121] Md Rafiul Hassan, Baikunth Nath, and Mohammed Alauddin Bhuiyan. Bengali phoneme recognition: a new approach. In *Proc. 6th international conference on computer and information technology (ICCIT03)*, 2003.

[122] AKMM Houque. Bengali segmented speech recognition system. *Undergraduate thesis, BRAC University, Bangladesh*, 2006.

[123] Pratyush Banerjee, Gaurav Garg, Pabitra Mitra, and Anupam Basu. Application of triphone clustering in acoustic modeling for continuous speech recognition in bengali. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[124] Sandipan Mandal, Biswajit Das, and Pabitra Mitra. Shruti-ii: A vernacular speech recognition system in bengali and an application for visually impaired community. In *Students' Technology Symposium (TechSym), 2010 IEEE*, pages 229–233. IEEE, 2010.

[125] NS Nehe and RS Holambe. New feature extraction techniques for marathi digit recognition. *vectors*, 1:2, 2009.

[126] Kayte Charansing Nathoosing. A multi-hmm marathi iolated wod recognizer. *Science Research Report*, 2(2):175–177, April 2012.

[127] Vaishali Patil and Preeti Rao. Acoustic features for detection of aspirated stops. In *Communications (NCC), 2011 National Conference on*, pages 1–5. IEEE, 2011.

[128] Vaishali Patil and Preeti Rao. Acoustic features for detection of phonemic aspiration in voiced plosives. In *INTERSPEECH*, pages 1761–1765, 2013.

[129] Tejas Godambe, Namrata Karkera, and K Samudravijaya. Adaptation of acoustic models for improved marathi speech recognition. Acoustics, 2013.

[130] Saroj Bajirao Jadhav, Jayshree Ghorphade, and Rishikesh Yeolekar. Speech recognition in marathi language on android os. *IJRCCT*, 3(8):815–818, 2014.

[131] Bharti W Gawali, Santosh Gaikwad, Pravin Yannawar, and Suresh C Mehrotra. Marathi isolated word recognition system using mfcc and dtw features. *ACEEE International Journal on Information Technology*, 1(01):21–24, 2011.

[132] Santosh Gaikwad, Bharti Gawali, and SC Mehrotra. Polly clinic inquiry system using ivr in marathi language. *International Journal of Machine Intelligence*, 3(3), 2011.

[133] S Mohanty and BK Swain. Continuous oriya digit recognition using bakis model of hmm. *International Journal of Computer Information Systems*, 2(1):2011, 2011.

[134] Sanghamitra Mohanty and Basanta Kumar Swain. Markov model based oriya isolated speech recognizer-an emerging solution for visually impaired students in school and public examination. *Special Issue of IJCCT*, 2(2):3, 2010.

[135] Agha Ali Raza, Sarmad Hussain, Huda Sarfraz, Inam Ullah, and Zahid Sarfraz. An asr system for spontaneous urdu speech. *the Proc. of Oriental COCOSDA*, pages 24–25, 2010.

[136] Huda Sarfraz, Sarmad Hussain, Riffat Bokhari, Agha Ali Raza, Inam Ullah, Zahid Sarfraz, Sophia Pervez, Asad Mustafa, Iqra Javed, and Rahila Parveen. Large vocabulary continuous speech recognition for urdu. In *Proceedings of the 8th International Conference on Frontiers of Information Technology*, page 1. ACM, 2010.

[137] S K Katti M A Anusuya. Kannada speech recognisation using discrete wavelet tansform pca. *International Conference on Computer Applications*, pages 24–27, Dec 2010.

[138] Sarika Hegde, KK Achary, and Surendra Shetty. Isolated word recognition for kannada language using support vector machine. In

Wireless networks and computational intelligence, pages 262–269. Springer, 2012.

[139] MA Anusuya and SK Katti. Speaker independent kannada speech recognition using vector quantization. In *IJCA Proceedings on National Conference on Advancement in Electronics and Telecommunication Engineering NCAETE (1)*, pages 32–35. Citeseer, 2012.

[140] Sivakumar C A. A simple approach for speech recognition: Kannada digits speech recognition for permanent deaf patiants. *Lap Lambet Academic Publishing*, Sep 2012.

[141] Sarika Hegdea, KK Acharyb, and Surendra Shettyc. Analysis of isolated word recognition for kannada language using pattern recognition approach.

[142] V S Girija Sivapasad P S. Speech recognition of isolated telugu vowels using neural networks. *Proceedings of the first Indina International Conference on Artificial Intelligence, Hyderabad*, Dec 2003.

[143] Rajesh Mahanand Hegde, Hema A Murthy, and Venkata Ramana Rao Gadde. Continuous speech recognition using joint features derived from the modified group delay function and mfcc. In *Eighth International Conference on Spoken Language Processing*, 2004.

[144] Rajesh M Hegde, Hema A Murthy, and GV Ramana Rao. Speech processing using joint features derived from the modified group delay function. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 1, pages I–541. IEEE, 2005.

[145] N Kalyani and Dr KVN Sunitha. Syllable analysis to build a dictation system in telugu language. *arXiv preprint arXiv:1001.2263*, 2010.

[146] KVN Sunitha, N Kalyani, et al. Isolated word recognition using morph knowledge for telugu language. *International Journal of Computer Applications*, 38(12):47–54, 2012.

[147] N Usha Rani and PN Girija. Error analysis to improve the speech recognition accuracy on telugu language. *Sadhana*, 37(6):747–761, 2012.

[148] Himanshu Nitinbhai Patel and PV Virparia. A small vocabulary speech recognition for gujarati. *International Journal of Advanced Research in Computer Science*, 2(1), 2011.

[149] Purnima Pandit and Shardav Bhatt. Automatic speech recognition of gujarati digits using dynamic time warping. *International Journal of Engineering and Innovative Technology*, 3(12), 2014.

[150] Patel Pravin and Harikrishna Jethva. Neural network based gujarati language speech recognition. *Int J Comput Sci Manage Res*, 2(5):2623–2627, 2013.

[151] MK Deka, CK Nath, SK Sarma, and PH Talukdar. An approach to noise robust speech recognition using lpc-cepstral coefficient and mlp based artificial neural network with respect to assamese and bodo language. In *International Symposium on Devices MEMS, Intelligent Systems & Communication (ISDMISC)*, pages 23–26, 2011.

[152] Utpal Bhattacharjee. Recognition of the tonal words of bodo language. In *International Journal of Recent Technology & Engineering. ISSN: 2277-3878, Volume-1, Issue-6, January 2013 IJCA TM: www. ijcaonline. org.* Citeseer, 2013.

[153] N K Narayanan V L Lajish, Thasleema T M. Isolated word speech recognition using lexicon based hidden markov models. *Proceedings of the IEEE International Conference of Signal and Image Processing*, pages 988–991, 2006.

[154] N K Narayanan V L Lajish, Thasleema T M. Hidden markov models for isolated word speech recognition based on large vocabulary. *Proceedings of National Symposium on Accoustics*, 2006.

[155] R Syama and Suma Mary Idikkula. Hmm based speech recognition system for malayalam. In *The International Conference on Artificial Intelligence*, 2008.

[156] A Raji Sukumar, A Firoz Shah, and P Babu Anto. Isolated question words recognition from speech queries by using artificial

neural networks. In *Computing Communication and Networking Technologies (ICCCNT), 2010 International Conference on*, pages 1–4. IEEE, 2010.

[157] Anuj Mohamed and KN Nair. Continuous malayalam speech recognition using hidden markov models. In *Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India*, page 59. ACM, 2010.

[158] Anuj Mohamed and KN Ramachandran Nair. Hmm/ann hybrid model for continuous malayalam speech recognition. *Procedia Engineering*, 30:616–622, 2012.

[159] VL Lajish, P Vivek, and RK Sunil Kumar. Hmm word modeling and short query based directory access for automatic phone dialing in malayalam. In *Proc. of the Indo-French Conference on Acoustics, Acoustics 2013*, pages 10–15, 2013.

[160] Lajish V L Vivek P, R K Sunil kumar. Automatic keyword spotting from malayalam conversational speech using hmm. *Proceedings of National Symposium on Accoustics*, 2005.

[161] Sweety Sarma and Anupam Barman. Multilingual speech identification using artificial neural network.

[162] E Veera Raghavendra, Sachin Joshi, Vamshi Ambati, and Kishore S Prahallad. Rapid development of speech to speech systems for tourism and emergency services in indian languages.

[163] Seeram Tejaswi and S Umesh. Addressing data sparsity in dnn acoustic modeling. In *Communications (NCC), 2017 Twenty-third National Conference on*, pages 1–5. IEEE, 2017.

[164] *Report of the Commissioner for linguistic minorities: 50th report (July 2012 to June 2013)*. Commissioner for Linguistic Minorities, Ministry of Minority Affairs, Government of India, 17 September 2016.

[165] Ronald E Asher and TC Kumari. *Malayalam*. Psychology Press, 1997.

[166] Arun Soman, Sachin Kumar, VK Hemanth, M Sabarimalai Manikandan, and KP Soman. Corpus driven malayalam text-to-speech synthesis for interactive voice response system. *International journal of computer application*, 29(4), 2011.

[167] T.M. Thasleema V.L. Lajish and N.K Narayanan. *Hidden Markov Models for isolated wordspeech recognition based on large vocabulary.* Proc. of the National Symposium on Acoustics, (NSA2006), NPL, New Delhi., 2006.

[168] Wesley Mattheyses and Werner Verhelst. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66:182–217, 2015.

[169] Lawrence R Rabiner and Biing-Hwang Juang. Fundamentals of speech recognition. 1993.

[170] VR Prabodhachandran Nayar. Malayalam verbal forms. *Trivandrum: DLA*, 1972.

[171] http://www.cmltemu.in/.

[172] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[173] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2016.

[174] Gosse Bouma. A finite state and data-oriented method for grapheme to phoneme conversion. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 303–310. Association for Computational Linguistics, 2000.

[175] Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451, 2008.

[176] Thomas Burrow. *The Sanskrit Language.* Motilal Banarsidass Publ., 2001.

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

[178] Haowen Jiang. Malayalam: a grammatical sketch and a text. *Department of Linguistics, Rice University*, 2010.

[179] *Malayalam Script—Adoption of New Script for Use—Orders Issued.* Government of Kerala, 1971.

[180] Bartosz Ziółko, Jakub Gałka, Suresh Manandhar, Richard C Wilson, and Mariusz Ziółko. The use of statistics of polish phonemes in speech recognition.

[181] Vishal Chourasia[1], K Samudravijaya, Maya Ingle[1], and Manohar Chandwani. Statistical analysis of phonetic richness of hindi text corpora. 2007.

[182] Ms Sunder and Ms Pratima Sharma. Research on phoneme sequences for language identification and concurrent voice transmission. 2014.

[183] Paul Dalsgaard, Ove Andersen, Hanne Hesselager, and Bojan Petek. Language-identification using language-dependent phonemes and language-independent speech units. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1808–1811. IEEE, 1996.

[184] Pavel Matejka, Igor Szöke, Petr Schwarz, and Jan Cernockỳ. Automatic language identification using phoneme and automatically derived unit strings. *Lecture notes in computer science*, pages 147–154, 2004.

[185] Suzan Verberne. Context-sensitive spell checking based on word trigram probabilities. *Unpublished master's thesis, University of Nijmegen*, 2002.

[186] Bhupesh Bansal, Monojit Choudhury, Pradipta Ranjan Ray, Sudeshna Sarkar, and Anupam Basu. Isolated-word error correction for partially phonemic languages using phonetic cues. In *International Conference on Knowledge based Computer Systems (KBCS 2004)*, pages 509–519, 2004.

[187] Stéphane Dupont and Juergen Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE transactions on multimedia*, 2(3):141–151, 2000.

[188] Lawrence Rabiner and B Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[189] Bernd Weinberg and Jan Westerhouse. A study of buccal speech. *Journal of Speech, Language, and Hearing Research*, 14(3):652–658, 1971.

[190] Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. Irstlm: an open source toolkit for handling large scale language models. In *Interspeech*, pages 1618–1621, 2008.

[191] Mark Gales and Steve Young. The application of hidden markov models in speech recognition. *Foundations and trends in signal processing*, 1(3):195–304, 2008.

[192] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel. Sphinx-4: A flexible open source framework for speech recognition. 2004.

[193] Alex Rudnicky. Sphinx knowledge base tool (2010). *URL http://www. speech. cs. cmu. edu/tools/lmtool. html.[Online].*

[194] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. The htk book. *Cambridge university engineering department*, 3:175, 2002.

[195] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

[196] Steven B Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *Readings in speech recognition*, pages 65–74. Elsevier, 1990.

[197] Nick G Kingsbury and Peter JW Rayner. Digital filtering using logarithmic arithmetic. *Electronics Letters*, 7(2):56–58, 1971.

[198] Harald Singer and Mari Ostendorf. Maximum likelihood successive state splitting. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 2, pages 601–604. IEEE, 1996.

[199] Slava Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401, 1987.

[200] Andrew J Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. In *The Foundations Of The Digital Wireless World: Selected Works of AJ Viterbi*, pages 41–50. World Scientific, 2010.

[201] Andreas Stolcke, Yochai Konig, and Mitchel Weintraub. Explicit word error minimization in n-best list rescoring. In *Fifth European Conference on Speech Communication and Technology*, 1997.

[202] Lalit R Bahl, Peter F Brown, Peter V de Souza, and Robert L Mercer. A new algorithm for the estimation of hidden markov model parameters. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 493–497. IEEE, 1988.

[203] Allan Knight and Kevin Almeroth. Fast caption alignment for automatic indexing of audio. In *Methods and Innovations for Multimedia Database Content Management*, pages 204–220. IGI Global, 2012.

[204] Michael Christel, Scott Stevens, and Howard Wactlar. Informedia digital video library. In *Proceedings of the second ACM international conference on Multimedia*, pages 480–481. ACM, 1994.

[205] Darin Brezeale and Diane J Cook. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(3):416–430, 2008.

[206] Iosif Mporas, Todor Ganchev, Mihalis Siafarikas, and Nikos Fakotakis. Comparison of speech features on the speech recognition task. *Journal of Computer Science*, 3(8):608–616, 2007.

[207] Steve Young. A review of large-vocabulary continuous-speech. *IEEE signal processing magazine*, 13(5):45, 1996.

[208] Zheng-Hua Tan and Ivan Kraljevski. Joint variable frame rate and length analysis for speech recognition under adverse conditions. *Computers & Electrical Engineering*, 40(7):2139–2149, 2014.

[209] Rivarol Vergin, Douglas O'shaughnessy, and Azarshid Farhat. Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. *IEEE Transactions on Speech and Audio Processing*, 7(5):525–532, 1999.

[210] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[211] Jun Yang. Review of multi-instance learning and its applications. *Technical report, School of Computer Science Carnegie Mellon University*, 2005.

[212] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th annual international conference on machine learning*, pages 1249–1256. ACM, 2009.

[213] Bernhard Schölkopf and Alexander J Smola. Support vector machines, regularization, optimization, and beyond. *MIT Press*, 656:657, 2002.

[214] Michael Banbrook, Steve McLaughlin, and Iain Mann. Speech characterization and synthesis by nonlinear methods. *IEEE Transactions on Speech and Audio Processing*, 7(1):1–17, 1999.

[215] Arun Kumar and SK Mullick. Nonlinear dynamical analysis of speech. *The Journal of the Acoustical Society of America*, 100(1):615–629, 1996.

[216] J-P Eckmann and David Ruelle. Ergodic theory of chaos and strange attractors. *Reviews of modern physics*, 57(3):617, 1985.

[217] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*, volume 7. Cambridge university press, 2004.

[218] MC Valsakumar, SVM Satyanarayana, and V Sridhar. Signature of chaos in power spectrum. *Pramana*, 48(1):69–85, 1997.

[219] V Anantha Natarajan and S Jothilakshmi. Segmentation of continuous speech into consonant and vowel units using formant frequencies. *International Journal of Computer Applications*, 56(15), 2012.

[220] Norman H Packard, James P Crutchfield, J Doyne Farmer, and Robert S Shaw. Geometry from a time series. *Physical review letters*, 45(9):712, 1980.

[221] Dr S Broomhead and Gregory P King. Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena*, 20(2-3):217–236, 1986.

[222] DS Guru, BH Shekar, and P Nagabhushan. A simple and robust line detection algorithm based on small eigenvalue analysis. *Pattern Recognition Letters*, 25(1):1–13, 2004.

[223] RK Sunil Kumar, VL Lajish, and P Vivek. Power spectrum analysis of speech signals (focused on chaotic properties) for speaker identification, 2013.

[224] P Vivek, VL Lajish, and Sunil Kumar RK. Improved vowel phoneme classification over the eigenvalues of the reconstructed phase space using ann. 2015.

[225] David E Sigeti. Exponential decay of power spectra at high frequency and positive lyapunov exponents. *Physica D: Nonlinear Phenomena*, 82(1-2):136–153, 1995.

[226] Otto E Rössler. An equation for continuous chaos. *Physics Letters A*, 57(5):397–398, 1976.

[227] D Shanka Ray A Nath. Physical review. *A35*, 1959.

[228] G Ananthakrishna and MC Valsakumar. Repeated yield drop phenomenon: a temporal dissipative structure. *Journal of Physics D: Applied Physics*, 15(12):L171, 1982.

[229] Edward N Lorenz. Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, 20(2):130–141, 1963.

[230] T Kapitaniak. Distribution of transient lyapunov exponents of quasiperiodically forced systems. *Progress of Theoretical Physics*, 93(4):831–833, 1995.

Appendix A

Windows OS

TeXLive package - full version

1. Download the TeXLive ISO (2.2GB) from
   https://www.tug.org/texlive/
2. Download WinCDEmu (if you don't have a virtual drive) from
   http://wincdemu.sysprogs.org/download/
3. To install Windows CD Emulator follow the instructions at
   http://wincdemu.sysprogs.org/tutorials/install/
4. Right click the iso and mount it using the WinCDEmu as shown in
   http://wincdemu.sysprogs.org/tutorials/mount/
5. Open your virtual drive and run setup.pl

or

Basic MikTeX - TEX distribution

1. Download Basic-MiKTEX(32bit or 64bit) from
   http://miktex.org/download

2. Run the installer
3. To add a new package go to Start » All Programs » MikTex » Maintenance (Admin) and choose Package Manager
4. Select or search for packages to install

TexStudio - TEX editor

1. Download TexStudio from

Appendix B

LaTeX.cls files can be accessed system-wide when they are placed in the <texmf>/tex/latex directory, where <texmf> is the root directory of the user's TeXinstallation. On systems that have a local texmf tree (<texmflocal>), which may be named "texmf-local" or "localtexmf", it may be advisable to install packages in <texmflocal>, rather than <texmf> as the contents of the former, unlike that of the latter, are preserved after the LaTeXsystem is reinstalled and/or upgraded.

It is recommended that the user create a subdirectory <texmf>/tex/latex/CUED for all CUED related LaTeXclass and package files. On some LaTeXsystems, the directory look-up tables will need to be refreshed after making additions or deletions to the system files. For TeXLive systems this is accomplished via executing "texhash" as root. MIKTeXusers can run "initexmf -u" to accomplish the same thing.

Users not willing or able to install the files system-wide can install them in their personal directories, but will then have to provide the path (full or relative) in addition to the filename when referring to them in LaTeX.

LaTeX.cls files can be accessed system-wide when they are placed in the <texmf>/tex/latex directory, where <texmf> is the root directory of the user's TeXinstallation. On systems that have a local texmf tree (<texmflocal>), which may be named "texmf-local" or "localtexmf",

it may be advisable to install packages in <texmflocal>, rather than <texmf> as the contents of the former, unlike that of the latter, are preserved after the LaTeXsystem is reinstalled and/or upgraded.

It is recommended that the user create a subdirectory <texmf>/tex/latex/CUED for all CUED related LaTeXclass and package files. On some LaTeXsystems, the directory look-up tables will need to be refreshed after making additions or deletions to the system files. For TeXLive systems this is accomplished via executing "texhash" as root. MIKTeXusers can run "initexmf -u" to accomplish the same thing.

Users not willing or able to install the files system-wide can install them in their personal directories, but will then have to provide the path (full or relative) in addition to the filename when referring to them in LaTeX.

LaTeX.cls files can be accessed system-wide when they are placed in the <texmf>/tex/latex directory, where <texmf> is the root directory of the user's TeXinstallation. On systems that have a local texmf tree (<texmflocal>), which may be named "texmf-local" or "localtexmf", it may be advisable to install packages in <texmflocal>, rather than <texmf> as the contents of the former, unlike that of the latter, are preserved after the LaTeXsystem is reinstalled and/or upgraded.

It is recommended that the user create a subdirectory <texmf>/tex/latex/CUED for all CUED related LaTeXclass and package files. On some LaTeXsystems, the directory look-up tables will need to be refreshed after making additions or deletions to the system files. For TeXLive systems this is accomplished via executing "texhash" as root. MIKTeXusers can run "initexmf -u" to accomplish the same thing.

Users not willing or able to install the files system-wide can install them in their personal directories, but will then have to provide the path

(full or relative) in addition to the filename when referring to them in LaTeX.

LaTeX.cls files can be accessed system-wide when they are placed in the <texmf>/tex/latex directory, where <texmf> is the root directory of the user's TeXinstallation. On systems that have a local texmf tree (<texmflocal>), which may be named "texmf-local" or "localtexmf", it may be advisable to install packages in <texmflocal>, rather than <texmf> as the contents of the former, unlike that of the latter, are preserved after the LaTeXsystem is reinstalled and/or upgraded.

It is recommended that the user create a subdirectory <texmf>/tex/latex/CUED for all CUED related LaTeXclass and package files. On some LaTeXsystems, the directory look-up tables will need to be refreshed after making additions or deletions to the system files. For TeXLive systems this is accomplished via executing "texhash" as root. MIKTeXusers can run "initexmf -u" to accomplish the same thing.

Users not willing or able to install the files system-wide can install them in their personal directories, but will then have to provide the path (full or relative) in addition to the filename when referring to them in LaTeX. LaTeX.cls files can be accessed system-wide when they are placed in the <texmf>/tex/latex directory, where <texmf> is the root directory of the user's TeXinstallation. On systems that have a local texmf tree (<texmflocal>), which may be named "texmf-local" or "localtexmf", it may be advisable to install packages in <texmflocal>, rather than <texmf> as the contents of the former, unlike that of the latter, are preserved after the LaTeXsystem is reinstalled and/or upgraded.

It is recommended that the user create a subdirectory <texmf>/tex/latex/CUED for all CUED related LaTeXclass and package files. On some LaTeXsystems, the directory look-up tables will need to be refreshed after making additions or deletions to the system files. For TeXLive systems

this is accomplished via executing “texhash” as root. MIKTEXusers can run “initexmf -u” to accomplish the same thing.

Users not willing or able to install the files system-wide can install them in their personal directories, but will then have to provide the path (full or relative) in addition to the filename when referring to them in LATEX.

Appendix A

Table: List of Malayalam Graphemes in ascending order

SI. No.	Diphone frequency distribution- Sentence Corpus	IPA	SI. No.	Diphone frequency distribution- Sentence Corpus	IPA
1	അ	/a/	2	ം	/am/
3	ആ	/a:/	4	ഇ	/i/
5	ഉ	/u/	6	ഋ	/r̥/
7	എ	/e/	8	ഏ	/e:/
9	ഒ	/o/	10	ക	/ka/
11	ക്ക	/kka/	12	ക്ല	/kl̥a/
13	ഖ	/kha/	14	ഗ	/ga/
15	ഗ്ല	/gl̥a/	16	ഘ	/gha/
17	ങ	/ŋa/	18	ങ്ക	/ŋka/
19	ങ്ങ	/ŋŋa/	20	ച	/ca/
21	ച്ച	/cca/	22	ഛ	/cha/
23	ജ	/ɟa/	24	ഝ	/ɟha/
25	ഞ	/ɲa/	26	ഞ്ച	/ɲca/
27	ഞ്ഞ	/ɲɲa/	28	ട	/ʈa/
29	ട്ട	/ʈʈa/	30	ഠ	/ʈha/
31	ഡ	/ɖa/	32	ഢ	/ɖha/
33	ണ	/ɳa/	34	ൺ	/ɳ/
35	ണ്ട	/ɳʈa/	36	ണ്ണ	/ɳɳa/
37	ത	/ta/	38	ത്ത	/tta/
39	ഥ	/tha/	40	ദ	/da/
41	ധ	/dha/	42	ന	/n̪a/
43	ൻ	/n/	44	ന്ത	/n̪ta/
45	ന്ന	/n̪na/	46	പ	/pa/
47	പ്പ	/ppa/	48	പ്ല	/pl̥a/
49	ഫ	/pha/	50	ബ	/ba/

51	ബ്ല	/bl̪a/	52	ഭ	/bha/
53	മ	/ma/	54	മ്പ	/mpa/
55	മ്മ	/mma/	56	മ്ല	/ml̪a/
57	യ	/ya/	58	യ്യ	/yya/
59	ര	/ra/	60	ർ	/r̲/
61	ല	/la/	62	ൽ	/l/
63	ല്ല	/lla/	64	വ	/ va/
65	വ്വ	/vva/	66	ശ	/ʃa/
67	ശ്ല	/ʃl̪a/	68	ഷ	/ʂa/
69	സ	/ sa/	70	സ്ല	/sl̪a/
71	ഹ	/ ha/	72	ഹ്ല	/ hl̪a/
73	ള	/ l̪a/	74	ൾ	/ l̪/
75	ഴ	/ ɣa/	76	റ	/ r̲a/
77	ാ	/ a:/	78	ി	/ i/
79	ീ	/ i:/	80	ു	/ u/
81	ൂ	/ u:/	82	ൃ	/r̥/
83	െ	/ e/	84	േ	/ e:/
85	ൗ	/ au/	86	്	/ ə/

Appendix B

Diphone frequency distribution-Word Corpus

	ഇ	ഉ	ഈ	ഊ	എ	ഒ	ഏ	ഓ	അ	ആ	ഉ്	ഐ	ഔ
ഇ	0	3	0	0	5	2	0	0	0	1	0	0	0
ഉ	0	0	0	0	3	0	0	0	1	0	0	0	0
ഈ	0	0	0	0	0	0	0	0	0	0	0	0	0
ഊ	0	0	0	0	0	0	0	0	0	0	0	0	0
എ	0	0	0	0	0	0	0	0	0	0	0	0	0
ഒ	0	0	0	0	2	0	0	0	1	0	0	0	0
ഏ	0	0	0	0	4	1	0	0	1	0	0	0	0
ഓ	0	0	0	0	0	0	1	0	1	0	0	0	0
അ	17	6	0	0	7	3	0	1	4	10	0	0	0
ആ	5	0	0	0	4	2	0	0	9	0	0	0	0
ഉ്	0	0	0	0	0	0	0	0	1	0	0	0	0
ഐ	0	0	0	0	0	0	0	0	0	0	0	0	0
ഔ	0	0	0	0	0	0	0	0	0	0	0	0	0
പ്	1500	1734	241	588	350	394	263	313	8948	2420	0	52	61
ത്	4623	1728	359	176	168	176	254	340	10661	1882	0	76	20
റ്	181	260	3	1	11	7	7	12	340	44	0	0	0
്	2486	1478	80	12	66	0	41	71	4590	547	0	12	1
ച്	1626	686	262	278	623	155	313	223	3765	1080	0	51	39
ക്	1323	5177	429	621	246	730	475	997	19370	3807	0	337	173
ഫ്	19	56	7	7	2	0	20	24	346	48	0	1	0
ട്	380	48	5	33	0	0	17	21	1146	358	0	6	6
്	95	4	16	4	0	0	9	10	542	78	0	0	0
ഛ്	119	27	4	1	3	0	74	8	340	147	0	0	0
ഖ്	110	42	5	0	1	0	43	14	1151	136	0	2	0
ബ്	237	284	96	30	4	1	22	100	1009	322	0	8	3
ദ്	1027	435	370	133	3	0	483	148	3343	753	0	71	30
ഡ്	292	95	63	34	7	0	9	34	1055	157	0	1	0
ജ്	480	98	350	30	2	0	28	46	2849	537	0	28	2
ട്	384	599	101	74	2	1	42	352	3458	467	0	8	78
ഝ്	720	184	130	364	0	0	109	192	2005	678	0	15	27
ഡ്	937	258	113	141	0	0	62	47	2745	674	0	9	21
ഢ്	15	0	6	4	0	0	2	4	197	19	0	0	2
ഞ്	11	1	0	0	0	0	0	0	62	9	0	0	1
ഘ്	23	64	4	23	0	0	3	108	773	190	0	0	1
മ്	1229	1674	238	422	110	74	423	310	7115	2666	0	56	58
ന്	3636	976	727	113	226	42	315	188	7876	1705	0	106	29
ങ്	1484	250	166	3	10	1	64	42	3870	372	0	0	2
ഞ്	87	12	6	8	72	35	24	7	371	294	0	0	3
ങ്	36	154	1	1	1	0	19	10	383	69	0	0	0
സ്	715	730	75	308	7	6	181	100	3811	797	0	44	94
ഷ്	839	245	105	9	2	0	207	61	2662	308	0	3	13
ശ്	823	403	231	91	2	0	155	163	2581	764	0	57	47
ഹ്	660	163	107	45	2	0	115	84	2134	590	0	22	6
ര്	3526	2068	314	240	30	6	224	443	9752	1760	0	8	24
ർ	1313	898	359	43	0	0	25	25	6986	1402	0	2	30
ൽ	1630	443	166	56	41	14	224	499	5927	1057	0	12	27
ൾ	1292	290	61	17	24	3	85	52	1883	230	0	4	5
ഴ്	560	523	2	3	8	2	8	3	657	45	0	0	0
വ്	3893	219	448	10	356	2	893	67	8204	2902	0	357	4
യ്	742	963	23	166	40	15	101	666	9228	1904	0	7	37

Contd...

Diphone frequency distribution-Word Corpus

	പ്	ത്	റ്	ട്	ച്	ക്	ഫ്	മ്	ം	ഷ്	ഖ്	ബ്
ഇ	1409	3476	43	883	618	6000	31	60	4	0	110	136
ഉ	1887	2068	140	1047	425	6597	15	12	21	1	380	97
ഈ	295	475	58	186	115	604	14	39	24	0	3	27
ഊ	289	519	48	386	101	204	1	5	1	0	9	18
എ	60	112	56	629	38	121	0	1	0	0	2	1
ഒ	55	142	67	422	30	105	1	0	0	0	0	0
ഏ	155	525	89	200	95	520	6	3	7	0	141	10
ഓ	360	374	39	419	218	439	9	5	3	0	1	17
അ	4502	6519	170	2903	1507	9471	118	532	91	7	153	602
ആ	1110	2243	125	1324	479	1617	35	170	48	0	130	107
ഇ്	0	0	0	0	0	0	0	0	0	0	0	0
ഐ	55	183	1	9	21	137	1	15	2	0	8	7
ഔ	52	99	0	27	23	75	0	2	0	0	8	4
പ്	3115	365	2	0	5	13	6	31	0	0	0	0
ത്	196	3903	1	3	2	232	10	134	0	0	4	2
റ്	0	0	0	0	0	0	0	0	0	0	0	0
ട്	2	0	0	2532	4	8	2	0	2	0	0	1
ച്	0	0	0	0	1502	0	0	0	0	439	0	0
ക്	36	1155	4	13	24	6830	0	19	0	1	4	2
ഫ്	0	0	0	0	0	0	0	0	0	0	0	0
മ്	0	0	0	0	0	0	0	0	0	0	0	0
ം	0	0	0	0	0	0	0	0	0	0	0	0
ഷ്	0	0	0	0	0	0	0	0	0	0	0	0
ഖ്	0	0	0	0	0	0	0	0	0	0	0	0
ബ്	0	0	0	0	0	2	0	0	0	0	0	11
ഭ്	5	8	0	0	4	39	0	0	0	0	0	25
ഡ്	0	2	0	0	0	2	0	0	0	0	0	1
ജ്	0	1	0	0	0	0	0	0	0	0	0	0
ഗ്	0	2	0	0	0	0	0	0	0	0	0	2
ഋ്	0	97	0	0	0	0	0	2	0	0	0	0
ഡ്	0	35	0	0	0	1	0	0	0	0	0	0
ഢ്	0	0	0	0	0	0	0	0	0	0	0	0
ഝ്	0	0	0	0	0	0	0	0	0	0	0	0
ഛ്	0	53	0	0	0	5	0	0	0	0	0	0
ല്	1384	191	0	1	5	46	8	0	0	0	63	561
ന്	87	2154	0	0	9	48	1	118	0	1	0	0
ണ്	24	6	0	944	5	31	1	0	142	0	0	0
ഞ്	1	0	0	0	870	0	0	0	0	33	0	0
ങ്	0	0	0	0	0	1161	0	0	0	0	34	0
സ്	131	899	19	4	2	219	91	593	0	0	11	4
ഷ്	254	0	0	1037	0	201	4	0	397	0	0	0
ശ്	0	3	0	0	85	1	0	0	0	2	0	0
ഹ്	0	47	0	0	3	1	0	0	0	0	0	0
ര്	352	747	1	19	177	416	6	261	0	0	8	72
ർ	2	0	0	0	0	0	0	0	0	0	0	0
ൽ	365	31	2	1	28	210	2	0	0	0	0	14
ൾ	37	20	0	1	25	50	3	0	0	0	0	1
ഴ്	16	19	0	0	64	39	0	0	0	0	0	0
വ്	1	265	0	1	0	66	0	9	0	0	0	0
യ്	47	29	0	0	1	266	0	0	0	0	0	0

Contd...

Diphone frequency distribution-Word Corpus

	ദ്	ഡ്	ജ്	ഗ്	ഭ്	ധ്	ഝ്	ഢ്	ഘ്	മ്	ന്	ണ്	ഞ്
ഇ	647	60	344	362	332	331	0	1	100	1319	1692	596	208
ഉ	1058	126	235	263	216	173	1	0	21	1436	963	876	164
ഈ	36	117	110	21	95	50	14	0	14	136	419	117	8
ഊ	38	47	82	12	5	4	107	0	1	160	136	101	22
എ	2	8	3	2	0	0	0	0	1	175	177	218	48
ഒ	1	3	3	0	1	0	0	0	0	66	139	86	16
ഏ	399	28	79	36	9	104	2	0	41	118	293	157	8
ഓ	340	44	167	385	88	267	11	0	24	331	274	178	7
അ	2099	173	1429	1541	1704	1112	2	1	242	31436	18545	4415	872
ആ	1113	100	475	736	278	611	50	0	108	2197	3190	1145	160
ഉ്	0	0	0	0	0	0	0	0	0	0	0	0	0
ഐ	75	11	19	6	7	19	0	0	2	93	93	12	1
ഔ	63	25	11	15	7	8	6	0	8	54	40	56	16
ട്	1	0	0	0	0	0	0	0	0	1	11	1	0
ത്	4	1	1	10	22	1	0	0	6	369	112	37	0
റ്	0	0	0	0	0	0	0	0	0	0	0	0	0
ട്	0	1	0	0	0	0	0	0	0	0	0	1	0
ച്	0	0	0	0	0	0	0	0	0	0	0	0	1
ക്	4	0	0	0	2	0	0	0	0	42	15	2	0
ഫ്	0	0	0	0	0	0	0	0	0	0	0	0	0
ഥ്	0	0	0	0	0	0	0	0	0	0	0	0	0
ഠ്	0	0	0	0	0	0	0	0	0	0	0	0	0
ഷ്	0	0	0	0	0	0	0	0	0	0	0	0	0
ഖ്	0	0	0	0	0	0	0	0	0	0	0	0	0
ബ്	74	0	35	0	4	86	1	0	0	4	1	0	0
ദ്	71	0	0	95	82	541	23	0	28	29	3	0	0
ഡ്	0	19	6	29	3	0	7	0	0	7	0	0	0
ജ്	0	0	104	1	1	0	0	9	0	18	2	0	432
ഗ്	19	0	5	15	10	78	0	0	1	27	357	5	0
ഭ്	0	0	0	6	0	0	0	0	0	17	0	0	0
ധ്	0	0	0	0	0	0	0	0	0	0	0	0	0
ഝ്	0	0	0	0	0	0	0	0	0	0	0	0	0
ഢ്	0	0	0	0	0	0	0	0	0	0	0	0	0
ഘ്	0	0	0	0	0	0	0	0	0	0	0	17	0
മ്	34	7	19	1252	383	2	0	0	31	430	27	9	0
ന്	1295	7	4	3	2	715	0	0	0	423	1134	3	0
ണ്	0	1024	1	2	0	0	4	0	0	62	1	510	0
ഞ്	0	0	360	0	0	0	0	4	0	0	0	0	241
ങ്	0	0	0	0	0	0	0	0	51	21	3	0	0
സ്	2	0	3	8	0	0	0	0	0	158	107	7	1
ഷ്	0	0	0	0	0	0	0	0	0	149	0	311	0
ശ്	1	0	0	2	0	1	0	0	0	148	23	3	0
ഹ്	46	0	0	0	0	0	0	0	0	187	20	4	0
ര്	252	7	233	261	210	265	1	12	102	892	67	582	0
ർ	0	0	0	0	0	0	0	0	0	0	1	0	0
ൽ	11	0	3	37	8	0	0	0	0	83	11	0	0
ൾ	0	0	0	2	1	0	0	0	0	19	4	0	0
ഴ്	1	0	3	1	0	0	0	0	0	25	5	0	0
വ്	46	0	4	0	0	0	2	0	0	0	16	0	0
യ്	1	0	0	0	0	0	0	0	0	64	5	0	0

Contd...

Diphone frequency distribution-Word Corpus

	ങ്	സ്	ഷ്	ശ്	ഹ്	ര്	ർ	ൽ	ൾ	ഴ്	വ്	യ്
ഇ	173	567	1116	745	282	2521	322	1653	479	278	1139	1481
ഉ	194	297	862	177	144	2049	511	892	611	355	883	99
ഊ	18	66	95	150	36	1047	51	274	44	170	446	538
ഋ	22	32	180	32	89	1013	110	369	101	63	137	48
എ	90	14	3	8	0	269	164	222	217	165	55	118
ഒ	58	12	2	5	0	169	75	190	160	181	19	30
ഏ	27	36	375	536	118	410	87	289	117	45	399	338
ഓ	20	44	306	89	229	372	63	470	163	62	145	83
അ	1227	3983	383	1779	1567	12327	782	5809	1715	503	5956	4115
ആ	222	1490	276	903	827	5065	184	2179	846	238	2140	1678
ഐ	0	0	0	0	0	0	0	0	0	0	0	0
ഐ	8	18	49	66	10	173	4	110	1	0	137	55
ഔ	4	18	43	35	33	161	4	60	12	0	40	2
ു	0	31	64	8	9	0	2303	0	91	0	1	101
ൂ	0	292	19	1	4	0	2449	0	1	0	274	758
ൃ	0	0	0	0	0	0	0	0	0	0	0	0
ൗ	0	1	0	0	0	0	85	0	4	0	14	39
ച്	0	0	0	0	0	0	2	0	0	0	1	90
ക്	0	26	2602	35	0	0	916	0	119	1	64	121
ഫ്	0	0	0	0	0	0	0	0	0	0	0	0
ട്	0	0	0	0	0	0	0	0	0	0	0	0
ഠ്	0	0	0	0	0	0	0	0	0	0	0	0	0
ഛ്	2	0	0	0	0	0	0	0	0	0	0	0
ഖ്	0	0	0	0	0	0	0	0	0	0	0	0
ബ്	0	0	0	1	35	0	196	0	8	0	1	9
ദ്	0	0	70	31	0	0	1071	0	0	0	514	473
ഡ്	0	0	0	0	0	0	19	0	0	0	14	34
ജ്	0	0	0	0	1	0	52	0	1	0	79	222
ഗ്	0	0	1	0	153	0	741	0	27	0	30	90
ഋ്	0	0	3	3	0	0	0	0	0	0	0	0
ഢ്	0	0	17	0	0	0	0	0	0	0	0	0
ഝ്	0	0	0	0	0	0	0	0	0	0	0	0
ഭ്	0	0	0	0	0	0	0	0	0	0	0	0
ഫ്	0	0	5	0	0	0	0	0	0	0	0	0
മ്	1	305	42	273	91	5	98	10	79	0	116	188
ന്	0	29	4	5	2	0	34	1	0	0	134	445
ൺ	0	7	0	0	0	0	1	0	0	0	24	206
ഞ്	0	0	0	0	0	0	0	0	0	0	0	0
ങ്	774	0	6	0	0	0	0	0	0	0	0	1
സ്	0	498	33	0	0	1	197	0	12	0	446	231
ഷ്	0	0	2	0	0	0	0	0	0	0	32	212
ശ്	7	0	0	102	0	0	497	0	85	0	435	159
ഹ്	0	0	6	0	0	0	35	3	16	0	113	96
ര്	2	12	366	210	118	0	0	37	3	0	678	625
ർ	0	0	0	0	0	0	0	0	0	0	0	4
ൽ	0	1	0	2	7	1	0	1032	0	0	85	216
ൾ	0	0	0	1	0	0	0	0	617	0	20	6
ഴ്	0	5	0	2	0	0	0	0	1	0	23	0
വ്	0	0	63	4	4	0	146	0	1	0	54	883
യ്	0	0	0	0	1	0	57	0	0	0	17	226

List of Publications of the Author

[1] **Vivek P**, Kumar Rajamani, Lajish V L. "Effective News Video Classification Based On Audio Content: A Multiple Instance Learning Approach." *International Journal of Computer Science and Information Technologies (IJCSIT)*, Vol. 7 (6), pp. 2556-2560, ISSN: 0975-9646, 2016.

[2] R K Sunil Kumar, K M Muraleedharan, **Vivek P**, Lajish V L. "Study of Nonlinear Properties of Vocal Tract and its effectiveness in Speaker Modeling." *Journal of Acoustical Society of India*, Vol. 43, No. 2, pp. 116-124, 2016.

[3] Lajish V L, Sunil Kumar R K, **Vivek P**. "Speaker Identification using a Nonlinear Speech model and ANN." *International Journal of Advanced Information Technology (IJAIT)*, Vol. 2, No.5, October 2012, DOI: 10.5121/ijait.2012.2502.

[4] **Vivek P**, R K Sunil Kumar, V L Lajish. "Automatic Keyword Spotting from Malayalam Conversational Speech Using HMM." *National Symposium on Acoustics, NSA-2015 Goa, Acoustics for Ocean Environment,* 2015.

[5] V L Lajish, Habeebath Kakkattuputhiyottil, **Vivek P**. "A Morpheme based Language Model for Malayalam Spoken Short Query Processing." *National Symposium on Acoustics, NSA-2015 Goa, Acoustics for Ocean Environment*, 2015.

[6] R K Sunil Kumar, K M Muraleedharan, **Vivek P**, V L Lajish, "Study of Nonlinear Properties of Vocal Tract and its

effectiveness in Speaker Modeling." *National Symposium on Acoustics, NSA-2015 Goa, Acoustics for Ocean Environment*, 2015.

[7] Anoop K Anoop, **Vivek P**, Lajish V L. "A Keyword Spotting approach for Content Based Indexing and Retrieval of Malayalam News Videos." *National Symposium on Acoustics, NSA-2015 Goa, Acoustics for Ocean Environment*, 2015.

[8] **Vivek P**, Lajish V L, Sunil Kumar R.K. "Improved Vowel Phoneme Classification over Eigen values of the Reconstructed Phase Space Using ANN." *5th National Conference on Indian Language Computing (NCILC – 2015)*, 2015.

[9] Lajish V L, **Vivek P**, R K Sunil Kumar. "Malayalam speech controlled multipurpose Robotic arm." *26th Kerala Science Congress* Pookode, Wayanad, 2014.

[10] R K Sunil Kumar, V L Lajish, **Vivek P**. "Power Spectrum Analysis of Speech Signals (Focused On Chaotic Properties) for Speaker Identification." *Acoustics 2013* New Delhi, 2013.

[11] V L Lajish, **Vivek P**, R K Sunil Kumar. "HMM Word Modeling and Short Query based Directory Access for Automatic Phone Dialing in Malayalam." *Acoustics 2013*, New Delhi, 2013.