Analysis of Emotional, Noisy and Pathological Speech Signals from the Perspective of Nonlinearity and Multifractality of Reconstructed System Hyperspace

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

A Thesis Submitted by

MURALEEDHARAN K M

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

DEPARTMENT OF PHYSICS
GOVERNMENT COLLEGE MADAPPALLY
VADAKARA, CALICUT, KERALA,
INDIA – 673102
(Affiliated to University of Calicut)
DECEMBER 2021

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

CERTIFICATE

This is to certify that the thesis entitled "Analysis of Emotional, Noisy and Pathological Speech Signals from the Perspective of Nonlinearity and Multifractality of Reconstructed System Hyperspace" is a report of original work carried out by **Mr. Muraleedharan K. M.** under my supervision and guidance in the Department of Physics, Govt. College, Madappally, Vadakara, Calicut, Kerala and that no part thereof has been presented for the award of any other degree.

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

DECLARATION

I hereby declare that the work presented in this thesis entitled "Analysis of Emotional, Noisy and Pathological Speech Signals from the Perspective of Nonlinearity and Multifractality of Reconstructed System Hyperspace", is the original work done by me under the guidance of Dr. R. K. Sunil Kumar, Department of Information Technology, Kannur University, Kerala and no part thereof has been presented for the award of any other degree.

Muraleedharan K. M.
Research Scholar
Govt. College, Madappally, Vadakara
Calicut, Kerala

ACKNOWLEDGEMENTS

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

I extend my sincere gratitude to Dr Udaya Kumar O. K., Principal, Govt. College, Madappally, Calicut, Kerala for providing the right resources and facilities in the department to accomplish my research work. I thank Dr. K. P Harikrishnan, Professor (Rtd), Cochin college, Kerala for his valuable suggestions and advice regarding the work. I am extremely thankful to Prof. K. Suresh Babu, Former Head of my department and Dr P. Ramakrishnan, Former Principal of the same institution, Prof. K. C. Abraham, former head of the department, St.Mary's College, Sulthan Batheri for extending their valuable suggestions and support during the work tenure. I am deeply indebted to my department teachers, especially Dr Suneera T. P. (Head of the Department), Dr Harikrishnan G. and Dr Nithyaja B, for their inestimable support

My special thanks to the M.Sc. project students for their indebted support during the recording and preparation of the database mentioned in this work. I extend my thanks to all my friends, especially Mr. Rajeev Arakkal, Software Senior Principal Engineer, Dell EMC, Bangalore, for the technical support and Mr. Aneesh Thomas, HSST, GVHSS Karthikapuram, Mr. Saneesh T Peter, Senior Documentation Manager at MetricStream, Bangalore

and Mr. Vimal Sebastian for their support in English language editing of the thesis.

Words are beyond to express my gratitude towards my colleagues, Mr Bibish Kumar K.T., Mr Sunil John and Ms Aljinu Khadar K.V. for their consistent whole-hearted co-operation and encouragement during this work. I owe a huge debt of gratitude to my parents, wife and children, whose unwavering encouragement and support served as a constant source of motivation for this work

Muraleedharan K. M.

To my family & friends….

ABSTRACT

During the course of a conversation, a great deal of information is communicated to the listener by way of the propagation of the speech signal. A person may detect emotions in a voice, notice changes in the voices of known individuals, and can distinguish one voice from a collection of voices by listening to it. The characteristics of the speech signal that may be detected are inherent in the signal and are formed by the underlying dynamics of the speech production system that generates the signal. Various studies have been conducted by many researchers over the past few decades in the areas of speaker identification, speech identification, pathological voice analysis, noise analysis, and emotional speech analysis. The majority of these investigations are based on the linear approximations and make use of spectral and prosodic features to conduct their investigations. Even though these parameters improved the accuracy of speech recognition, they were ineffective when dealing with pathological, noisy, and emotional speech signals.

This work aims to investigate the performance of pathological, noisy and emotional speech recognition systems which utilises nonlinear and multifractal features extracted from the reconstructed hyperspace. Experimental data are used to validate the proposed methodology and system components, which are then thoroughly explained. As a result, better systems, based on nonlinear and multiracial features, are proposed for the detection of pathology, noise, and emotional expression.

The availability of a phonetically balanced audio speech database in the language in which the application is to be used is the foundation of any speech-based application. A new Malayalam audio speech database has been developed and presented. This collection contains 50 isolated Malayalam phonemes and 207 related words that comprise all allophonic variations. The

database is created in open and closed modes. The database was segmented and labelled using the spectral subtraction method.

The developed database is utilised to optimise time delay and embedding dimension of the chaotic attractor of the Reconstructed Phase Space (RPS). The hypothetical abstract space that represents the system under investigation will help analyse its dynamics. The subject of optimizing the embedding dimension is studied using Lorenz and Rossler systems as models. The time delay of embedding was determined using the Mutual Information method, and it varied between samples. The embedding dimension for Lorenz and Rossler systems is three, confirming the applicability of False Nearest Neighbour (FNN) and Principal Component Analysis (PCA) for dimensionality reduction. It was found that the embedding dimension of the speech production system is unaffected by age, gender, or sample frequency in Malayalam phoneme time series. The tested samples' mode is six, with a mean close to it. In this case, the standard deviation is so small that the mode value can be used.

For ensuring the underlying nonlinear structure in the signal a surrogate analysis was performed at the optimised delay and embedding dimension. While comparing the statistical significance level of Malayalam phoneme time series with standard Lorenz and Rossler systems, the significance level of different phonemes was found to be different but comparable. The significance level for Correlation dimension at minimum embedding dimension (D_{2m}) and Correlation entropy at minimum embedding dimension (K_{2m}) analysis shows that the values are closer to those of standard systems for vowels അ /a/, ഇ/i/, ഏ/e/ and all the analysed syllables. Thus, D_{2m} and K_{2m} can be used as better tools for the study of nonlinear dynamical structures in the speech production system, emotion

recognition, and pathological analysis in place of saturated values of D_2 and K_2.

The nonlinear features are utilized in distinguishing pathological voice signals from healthy voice signals. The features used are D_{2m}, K_{2m} , and four fitting coefficients of the f(α) spectrum of strange attractor. The study relied on the VOICE database. FNN and MI have optimized the embedding dimension and time delay of RPS. The data was subjected to a statistical surrogate analysis to ensure that the characteristics used in the analysis were discriminated, and a reasonable significance level indicated the presence of nonlinearity. Based on the measures examined, a classification system is proposed. SVM was used to assess the performance of the proposed classification system in distinguishing between pathological and normal voices. The precision is 99%, and the accuracy is 97%. When compared to recognition algorithms based on linear feature vectors and other nonlinear parameters, this accuracy is promising.

The multifractal features derived from the multifractal detrended fluctuation analysis (MFDFA) is used for noise identification. The singularity spectrum width and extremal Holder exponents are seemed to be reduced in the MFDFA of the voice samples due to additive noise. The Malayalam speech database developed together with its noise simulated signals are utilised to distinguish pink, red, and white noises. The SNR has an effect on the reduction, and the SNR rate can be computed by multiplying the percentage reduction in the parameters. The noise categories are recognised using feature vectors and an SVM classifier, and the accuracy attained indicates that multifractal features are an efficient tool for recognising different types of noise.

Finally, nonlinear features and multifractal features are combined with spectral and prosodic features to recognise speaker emotion. The male sounds from the RAVDESS speech emotion database are used for study. Fundamental frequency (F0), formant frequencies (F1 and F2) and Mel frequency cepstral coefficients (MFCCs) are the most popular prosodic and spectral features. Instead of using these features directly, its noise-tolerant version was proposed by using the autocorrelation function. D_{2m}, K_{2m}, and largest Lyapunov exponent (LLE), all at minimum embedding dimension, are used as nonlinear features, and singularity spectrum parameters (height and width of singularity spectrum, and q-order Hurst exponents) are taken as multifractal features. Surrogate analysis is performed with D_{2m} ,K_{2m} and LLE as nonlinear measures in both normal and emotional signals, and the high level of significance indicates the nonlinear structure in the signal. The suggested system's classification accuracy is assessed using a Support Vector Machine (SVM) Classifier. As per the result, the addition of nonlinear features with spectral and prosodic features improves recognition accuracy and minimises classification ambiguity. The integration of multifractal features further improves the accuracy and precision. The nonlinearity and multifractality of the signal reflects in the speaker's emotional content, which makes these features a supporting tool for recognising the speaker's emotional state.

By studying and analyzing the nonlinear and multifractal features the pathological, noisy and emotional content in speech signal can be captured and it will throw light on the inherent dynamics of the speech production system. These features not only provide high accuracies for pathology detection and excellent noise identification performance but also help in identifying emotional cues in speech.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ACR	-	Autocorrelation Function
ANN	-	Artificial Neural Network
ASR	-	Automatic Speech Recognition
CV	-	Consonant-Vowel
dB	-	Decibel
DDR	-	Double Dynamic Range
DWT	-	Discrete Wavelet Transform
ED	-	Embedding Dimension
ERM	-	Empirical Risk Minimisation
F0	-	Fundamental Frequency
FFT	-	Fast Fourier Transform
FNN	-	False Nearest Neighbour
fs	-	sampling frequency
GMM	-	Gaussian Mixture Model
HMM	-	Hidden Markov Model
IDFT	-	Inverse Discrete Fourier Transform
IFFT	-	Inverse Fast Fourier Transform
IAAFT	-	Iterative Amplitude Adjusted Fourier Transform
LLE	-	Largest Lyapunov Exponent
LPC	-	Linear Predictive Coding
MFCC	-	Mel Frequency Cepstral Coefficients
MFDFA	-	Multifractal Detrended Fluctuation Analysis
MI	-	Mutual Information
PCA	-	Principal Component Analysis
PSR	-	Phase Space Reconstruction
SER	-	Speech Emotion Recognition
SD	-	Standard Deviation

SHRP - Sub harmonics to harmonics ratio

SNR - Signal-to-Noise Ratio

SRM - Structural Risk Minimisation

SVM - Support Vector Machine

CHAPTER 1
INTRODUCTION

1.1 Background

"Speech is the most sophisticated behaviour of the most complex organism in the known universe" – Moore, R. K. (2007) [1]. *Speech processing* is a distinct discipline that encompasses a broad range of topics by incorporating a variety of technologies and applications that allow humans to communicate smoothly with intelligent computers. One of the most challenging characteristics, particularly for researchers in the field of speech processing, is the fact that it is multidisciplinary in nature, necessitating knowledge and skills from a variety of different fields. Bell Laboratories began developing automatic voice recognition systems in the 1950s, starting with basic digit recognition systems [2] and progressing to more complex systems over time. Since then, the recognition challenges have become increasingly sophisticated, ranging from speaker-dependent isolated word identification to speaker-independent continuous speech recognition with a wide vocabulary to spontaneous speech recognition in a noisy environment and everything in between. In today's world, automatic speech recognition is used in a variety of applications, including voice-enabled electronic devices, navigation systems, and inquiry systems, among other things.

During the course of a conversation, a great deal of information is communicated to the listener by way of propagation of the speech signal. A person may detect emotions in a voice, notice changes in the voices of known individuals, and can distinguish one voice from a collection of voices by listening to it. The characteristics of the speech signal that may be detected are inherent in the signal and are formed by the underlying dynamics of the speech production system that generates the signal. Various studies have been

conducted by many researchers over the past few decades in the areas of speaker identification, speech identification, pathological voice analysis, noise analysis, and emotional speech analysis. The majority of these investigations are based on the linear approximations and make use of spectral and prosodic features to conduct their investigations. Even though these parameters improved the accuracy of speech recognition, they were ineffective when dealing with pathological, noisy, and emotional speech signals.

Speech production systems are extremely complex systems and analysing the dynamics of such systems from a conventional linear perspective is a time-consuming effort. Despite the fact that a number of models have been published in the literature, the majority of them fall short of explaining properly the multi-behavioural aspects of speech. In this circumstance, a reverse mechanism, i.e., one that explains the system proprieties in terms of signal characteristics is extremely important for understanding. The study of nonlinearity in signals acquired popularity through the 1980s [3], as a result of the growth of nonlinear time series analysis. There have been a number of researches on the properties of speech that have been conducted using nonlinear dynamics. Describing the properties of the system from the reconstructed phase space of the attractor is the most prominent method that has been used widely.

For speech applications such as pathological voice analysis, noise identification, and speech emotion analysis, nonlinear time series analysis based on phase space reconstruction has been widely employed, and a significant number of classification systems can be found in the literature [4]. The use of nonlinear studies has been employed for a variety of pathological research, and the proposed recognition methods have not been shown to be sufficient in recognising the diseases, particularly in such circumstances as hyperkinetic dysphonia. A number of works have been published on noise

identification with nonlinearity; however, the process of identifying the noise remains a time-consuming endeavour. While comparing the performance of Automatic Speech Recognition (ASR) systems with human speech recognition (HSR), the former is less effective. One of the reasons for this is the inability to accurately capture the emotions in voice signals. There have been a lot of works that use spectral, periodic, and nonlinear features for successful emotion recognition; nonetheless, there is a need to improve the accuracy of emotion recognition in order to be more effective.

1.2 Motivation

To characterise a physiological system's temporal evolution, the nature of dynamical variables (minimum dimension of the dynamical system) involved in developing the system with time is the most vital element. As per 'Taken's theorem' the phase space should be reconstructed with proper delay and dimension in order to extract useful information about the system. As a result, in order to apply nonlinear time series to speech signals, it is necessary to optimise the embedding dimensions and time delays before doing the analysis. Large number of works in this direction is reported in signals like EEG and ECG and the optimisation provides immense results in these signals. The methods of False Nearest Neighbours (FNN) and Principal Component Analysis (PCA) are the most extensively utilised methodologies for optimising embedding dimension.

The lack of a freely available standard speech database poses the most significant challenge in optimising the embedding dimension of a system. The number of freely available databases does not allow for a generalisation of a result due to a lack of volume. Having a huge volume database with varied age groups and sampling frequencies is necessary for this task to be successful. This work aims to provide a Malayalam audio speech database that has been acquired in a variety of scenarios for a variety of research

purposes, with a particular emphasis on the study of nonlinearity in the speech production system. An extensive database with a high number of samples is required in order to generalise the nonlinear features of the system phase space and to optimise the embedding parameters in the system phase space.

The correlation dimension, the largest Lyapunov exponents, and the correlation entropy are the three most essential nonlinear characteristics employed in pathological analysis. It should be noted, however, that the features were retrieved from phase space in either two dimensions or three dimensions, which reduces the overall efficiency of the study. Pathological analysis is carried out in this thesis using the correlation dimension and correlation entropy (both extracted from the reconstructed hyperspace), as well as multifractal features, which is the first attempt in this direction. Despite the fact that the nonlinear structures are destructed by the presence of additive noise, the multifractal structure provides a clue as to the source of the noise. In this work, the multifractality of the speech time series is utilised for the identification of speech noise.

Emotion identification from speech is one of the most challenging problems to solve in the current world. Recently, a considerable number of researches have been published on emotion recognition utilising prosodic, spectral, and nonlinear features, among other techniques. These studies make use of nonlinear features obtained from two- or three-dimensional space. Nonlinear parameters extracted from the optimised hyper dimensional phase space and multifractal features are used as supporting components for already established spectral and prosodic features in this thesis, with the goal of improving their performance. The motivation of this research is that nonlinear and multifractal properties in the reconstructed hyperspace have not yet been successfully utilised for the analysis of diseased, noisy, and emotional speech signals.

1.3 Thesis Outline

This research aims to investigate the performance of pathological, noisy and emotional speech recognition systems which utilises nonlinear and multifractal features extracted from the reconstructed hyperspace. Experimental data are used to validate the proposed methodology and system components, which are then thoroughly explained. The rest of the thesis is organised as follows:

The objective of Chapter 2 is to lay the groundwork for the subsequent chapters. The chapter begins with an in-depth examination of the existing audio speech database. The various studies that have been conducted for the purpose of pathological voice analysis are covered in the next section. The following part examines the various research articles that have been published in the literature for the purpose of noise removal and noise identification. The last section reviews the various approaches to speech emotion recognition.

The third chapter discusses the creation and presentation of a new Malayalam audio speech database. This collection contains 50 isolated Malayalam phonemes and 207 related words that include all of the allophonic variations of the language. There are two different ways to establish a database: closed and open. A clean audio speech database is created with the help of 200 speakers (100male and 100 female) in a closed environment and noisy speech is developed with the help of 20 speakers (10 male and 10 female) in an open environment. The speakers belong to the age groups of five to ten, twenty to twenty-five and sixty to sixty-five. Each recording was to be repeated by the speakers ten times. The background noise in the database is removed using the spectral subtraction method. The database is segmented and labelled. The aforementioned database can be utilised to enhance research in a variety of speech-based signal studies.

The fourth chapter discusses the optimisation of time delay and embedding dimension for the purpose of phase space reconstruction. The Mutual Information (MI) method is utilised to find out the time delay for embedding. The False Nearest Neighbour (FNN) method and the Principal Component Analysis (PCA) method are used for optimising the embedding dimension of time series. The time series obtained from the typical non-linear systems, the Lorenz system and the Rossler system, is used to standardise the methods, and the Malayalam speech vowel time series developed in the third chapter is used for analysis. It was observed that the time delay varies from sample to sample, and, it ought to be better to figure out the time delay with the analysis. The embedding dimension is shown to be independent of gender, age, and sampling frequency and can be projected as six. Hence, a six-dimensional hyperspace will probably be adequate for reconstructing the attractor of speech time series.

In chapter 5, detailed surrogate data analysis is conducted for ensuring nonlinearity in the signal. Correlation Dimension and Correlation Entropy at minimum embedding dimension (D_{2m} and K_{2m}) are used as nonlinear discriminating measures. By taking Lorenz system and Rossler system with 30000 data points as model systems, the statistical significance level of five Malayalam vowel and consonant time series(chapter 3) for both D_{2m} and K_{2m} are analysed. The D_{2m} and K_{2m} for whole time series were determined, and the significance level is compared. It was found that the significance level for speech samples is comparable to that of Lorenz and Rossler systems. Both D_{2m} and K_{2m} analysis show better significance level for vowels and consonants except the vowels ഓ/o/ and ഊ/u/.

The viability of six nonlinear discriminating measures derived from the phase space realm, involving healthy and pathological voice signals, is studied in Chapter 6. The analysed parameters are Singularity spectrum

coefficients (α_{min}, α_{max}, γ_1 and γ_2), Correlation entropy and Correlation dimension at optimum embedding dimension (K_{2m} & D_{2m}). From the VOice ICar fEDerico (VOICED) database, comprising 208 healthy and pathological voices, 50 samples of each are used.The optimum time delay is determined by Mutual Information method, and the embedding dimension of these data sets is optimized by False Nearest Neighbor method. A statistical surrogate analysis has been performed on the data to check the discrimination of the nonlinear characteristics used in the analysis. A classification system is proposed based on the analysed features, and a classifier based on Support Vector Machines (SVM) was implemented to evaluate the proposed classification system.

In chapter 7, Multifractal Detrended Fluctuation Analysis (MFDFA) is introduced to identify the type of noise present in a speech signal. The effects of different types of noise on human speech data and their identification by MFDFA are studied in the chapter. The Malayalam vowel database and the corresponding simulated noisy signal (developed in Chapter 3) are utilised for the study. The pink noise, red noise, and white Gaussian noise are added to the database and the variation in the singularity spectrum width and the extremum values of the Holder exponent are estimated. The noise type was identified by the shift in singularity spectrum width and values of Holder exponents. A SVM Classifier was implemented to measure the performance of the proposed classification system based on multifractal features.

In Chapter 8 nonlinear features and multifractal features are combined with spectral and prosodic features to recognise speaker emotion. The male sounds from the RAVDESS speech emotion database are utilised. Correlation dimension (D_{2m}), correlation entropy (K_{2m}), and largest Lyapunov exponent (LLE), all at minimum embedding dimension, are used as nonlinear features, and singularity spectrum parameters (height and width of singularity

spectrum, and q-order Hurst exponents) are taken as multifractal features. Surrogate analysis is performed with D_{2m} , K_{2m} and LLE as nonlinear measures in both normal and emotional signals, and the high level of significance indicates the nonlinear structure in the signal. The suggested system's classification accuracy is assessed using a Support Vector Machine (SVM) Classifier.

The ninth chapter concludes the thesis by summarising the most significant findings of this work and drawing conclusions as well as making recommendations for future research. Following this chapter, references and the author's publications are listed.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

Human communication is most effective when it is done through speech. Language contains not only interpretable text but also a huge amount of paralinguistic data that might indicate a speaker's emotional and pathological shifts. Speech recognition technologies have been used to interpret human spoken language in a variety of disciplines, including automobile navigation, surveillance cameras, networking video and other human interface fields. The ability of machines to translate spoken language into written text is referred to as speech recognition. To do so, a speech recognition system must often take into account both the nonspecific and specific environment in order to effectively recognise speech content. As a result, for accurate speech recognition, feature extraction and characterization of speech signal are two crucial phases. The most commonly utilised feature extraction strategies in speech recognition are (1) phonetic features [5] (2) prosodic features [6], (3) features based on spectrum correlation [7], [8], and (4) feature fusion [9]. The piecewise linearity of voice signals characterises the above features. However, investigations have indicated that speech signal creation is a nonlinear process rather than a linear or stochastic process. As a result, extracting speech features only based on piecewise linearity of speech signals in the frequency and time domains will result in the loss of some nonlinear properties of speech signals, rendering incomplete extracted information.

Nonlinear analysis methods have been effectively implemented in a variety of fields as a result of recent advancements in the field. Although some academics have looked into the chaotic properties of speech signals,

there have been few studies into the geometric features and nonlinear features of chaotic attractors in speech signals. Nonlinear features are recovered from conventional two-dimensional and three-dimensional space in the majority of cases. As per Taken's theorem the phase space should be reconstructed with the optimized embedding delay and dimension in order to retrieve meaningful information from the system. As a result, feature extraction from the reconstructed hyperspace will help emotional, noisy, and pathological recognition systems perform better. A reliable database is essential to optimise the delay and dimension. Hence, a Malayalam audio speech database was also created as a part of this thesis.

The goal of this thesis is to develop a better classification system based on features taken from the reconstructed system hyperspace for diseased, noisy, and emotional speech analysis. To implement this system, the background and current scenario of its associated techniques or approaches must be thoroughly explored. A complete review of relevant works is discussed in this chapter. The development stages of the available audio speech database are summarised in Section 2.2. The research in pathological voice analysis using various approaches is reviewed in Section 2.3. The studies on noise identification approaches are discussed in detail in Section 2.4. Section 2.5 discusses the different studies on emotional speech analysis .Section 2.6 concludes the review.

2.2 Review on known Audio Speech Database

Years of advancement and continued study in speech-based applications have revealed the lack of standard audio speech databases. The scientific community has pushed for the construction of a large, well-organized audio speech database. A high-quality audio speech stream can be captured for a comparatively low cost. Rather than offering a detailed explanation of the qualities of the existing database, this study delivers a

tabular representation of the available audio speech databases, which enables for comparison and provides a clear picture of the features. The audio speech database's historical background is summarised in Table 2.1 in terms of gender allocation, speech corpus, hardware setup, and specific identifying characteristics. This section covers database building in both well-resourced and under-resourced languages. A complete awareness of the quantity and number of fundamental requirements and resources should be required to develop an audio speech database in an under resourced language.

Table 2.1 Available Audio Speech Database

Database – Year	Corpus – Repetition	Sampling frequency	Speaker (M,F)
TULIPS1 1995 [10]	• First four English digit – twice.	11.1 kHz.	12 (3,9)
M2VTS 1997 [11]	• French language. • Numbers (0 to 9) – 5 times.	48 kHz	37
XM2VTSDB 1999 [12] Extended M2VTS Database	• Three sentences (numbers and word) – twice.	32 kHz	295
AMP/CMU 2001 [13] Advanced Multimedia Processing Lab	• 78 Isolated words – 10 times each.	16 kHz	10 (7,3)
AV Letters 2002 [14]	• English language. • 780 utterances of letters (A to Z) .	22.05 kHz	10 (5, 5)
CUAVE 2002 [15] Clemson University Audio-Visual Experiments	• English language. • Isolated digits. • Connected digits. • Total 7000 utterances.	16 kHz	36 (17, 19) Speaker pairs -20
VidTIMIT 2002 [16]	• English language. • 10 TIMIT sentences by	32 kHz	43 (24, 19)

	each speaker.		
DUTAVSC 2002 [17]	• Dutch language. • POLYPHONE corpus.	44 kHz	8 (7,1)
BANCA 2003 [18]	• 4 Languages-English, French, Italian and Spanish. • Date of birth. • Names. • Addresses. • Numbers.	32kHz. 2 Microphone used.	52 (26, 26) for each language class.
AV-TIMIT 2004 [19]	• 450 TIMIT-SX sentences. • Each speaker utter 20 sentences. • First sentences are common and other 19 sentences are different.	16kHz.	223 (117, 106)
AVOZES 2004 [20] Audio Video OZtralian English Speech	• Australian English language. • Total of 56 sequences per speaker without repetition Digits. • Phrases • Continuous words. • Total of 56 sequences by each speaker (no repetition.)	48kHz.	20 (10, 10)
MANDARIN CHINESE 2004 [21]	• Chinese language. • Continuous speech. • Total 17,000 utterances.	48 kHz. 12 Microphones used.	225
VALID 2005 [22]	• XM2VTS speech corpus.	32 kHz.	106 (77, 29)
UWB-04-HSCAVC 2006 [23] University of West Bohemia-2004-Hundred Speakers Czech Audio-Visual Corpus	• Slavonic language (Czech and Russian). • 200 Sentences (150 unique and 50 shared).	44 kHz. 2 Microphones used.	100 (39, 61)

GRID 2006 [24]	• English language. • Each sentence with six-word sequence. • Command sentences. • Total 34,000 corpus.	25 kHz.	34 (18, 16)
UWB-07-ICAVR 2008 [25] University of West Bohemia- 2007-	• Czech language. • Total 10,000 utterances • 200 Sentences (150 unique and 50 shared).	44 kHz. 2 Microphones used.	50 (25, 25)
WAPUSK20 2010 [26]	• 100 GRID database sentences. • Total sentences 2000.	∗ 16 kHz. ∗ 4 audio channels.	20 (11,9)
AVA II 2010 [27]	• Persian language. • Phonemes, Phonemic combinations (cv, vc, vcv), 20 sentences and digits.	∗ 48 kHz. ∗ 2 Microphones used.	14 (7,7)
BL-Database 2011 [28] Blue Lips-Database	• French language. • 238 sentences. • Diphone rich utterances.	∗ 44.1 kHz. ∗ 2 Microphones used.	17 (9,8)
UNMC-VIER 2011 [29]	• 11 XM2VTS sentences. • Sequence of numerals.	∗ 48 kHz(From high quality camera). ∗ 22 kHz (Audio device).	123 (74, 49)
MoBio 2012 [30]	• English language. • 32 questions (short response questions, short response free speech, set speech, and free speech).	48 kHz.	152 (100, 52)
AVAS 2013 [31] Audio-Visual Arabic Speech	• Arabic language. • 36 daily words. • 13 casual phrases.	48 kHz.	50
†Oriya Digit Database 2013 [32]	• Oriya language. • Digits-4 times.	16 kHz	15 (5,10)
AGH 2015 [33]	• Polish language. • Isolated words and	44.1 kHz.	166 (one third

	Numbers. • Total 1,17,450 words.		female)
OuluVS2 2015 [34]	• English language. • Continuous digits. • Phrases. • TIMIT sentences.	High quality audio.	53 (40, 13)
TCD-TIMIT 2015 [35]	• 6913 phonetically rich TIMIT sentences.	16 kHz.	62 (32, 30)
†AMAUV 2015 [36] Aligarh Muslim University Audio Visual	• Hindi language. • 10 sentences out of which 2 sentences are common to all speaker.	44.1 kHz	100
MODALITY 2017 [37]	• English language. • 168 commands.	44.1 kHz. Array of 8 microphones used.	35 (26, 9)
AVID 2017 [38]	• Indonesian language. • 1040 sentences.	44.1 kHz	10 (5,5)
NTCD-TIMIT 2017 [39]	• Irish accent. • 5488 different TIMIT sentences.	16 kHz.	56
Audio-Visual Lombard Speech 2018 [40]	• 2700 Lombard and 2700 plain reference utterances. • Extension of GRID corpus.	48 kHz	54
3D Audio Visual Speech Corpus 2020[41]	• American English. • 224 sentences from CRM corpus -2 repetition. • 50 sentences from IEEE corpus.	48 kHz. 2 Microphones used	5 (2,3)
RUSAVIC 2021[42]	• Russian language. • 50 phrases related to driving condition. • 10 recording sessions.	48 kHz.	20

A wide range of standard databases have been reported, with the majority claiming to be useful for specific activities. Continuous speech, as opposed to isolated speech for the speaker verification test, is the most natural

voice material for the speech recognition problem. When compared to the speech recognition task, the speaker recognition task necessitates a large speaker population with high variability. The key requirements for building a voice database are a large phonetically balanced speech corpus uttered by many different speakers in an uncontrolled setting. It is necessary to determine the idiosyncrasies of the database's language and its linguistic history, as well as to compare it to other groups of languages, in order to overcome issues that developed during the database's formation in under-resourced languages. As a result, building an audio speech database in Malayalam that fulfils the majority of these requirements will have a huge impact on the research community.

2.3 Review on Acoustic Detection of Voice Disorders

Pathological speech classification and detection research began in the early 1980s, when machine learning methods and pattern recognition were still in their infancy. A few studies have been conducted utilising fundamental methods such as distance measurement, statistical analysis, and vector quantization, among others. However, in recent years, machine learning techniques have become popular for detecting disordered speech using the input signal's calculated acoustic properties. In this section, the research papers that employ pathological voice detection are reviewed.

By utilising digital inverse filtering, Deller and Anderson have classified and assessed laryngeal dysfunction. Automatic clustering is used to analyse the z-plane roots and a pattern feature vector. They could recognise the simulated anomalous laryngeal activity in the voice stream using an inverse filter approach. Childers and Bae [43] established time interval and amplitude difference measurements for analysing the EGG (Electroglottograph) signal. In both cases, the abnormal detection probability was 75.9%. Using the Teager Energy Operator, Cairns et al. [44] suggested a

constrained way to identify hyper nasality in spoken words. They used a probability distribution function to classify normal and hyper-nasal voices. The best category accuracy was 94.7%. These authors developed an algorithm based on fractal dimension, energy ratio, and zero-crossing properties. The normal and diseased voices are compared using the feature distance matrix. The fractal dimension accuracy was 96.1 percent, the energy ratio accuracy was 92.1 percent, and the zero-crossing feature categorization accuracy was 94.1 percent [45].

On the other hand, Parsa and Jamieson [46] looked at glottal noise as a way to distinguish between healthy and sick voices. They compared the measurements' probability distribution, ranking, and receiver operating characteristics (ROC) to classify them into two categories. The top classification rate was 96.5%. On the other hand, Hadjitodorov et al [47] suggested a method based on prototype distribution maps (PDM) to describe the probability density functions of normal and pathological speakers' input vectors. These include HNR, pitch period, low-to-high energy ratio, and pitch pulse form. Rosa et al. [48] employed a statistical approach to distinguish between healthy and diseased voices. Using PDM (prototype distribution neural map) and jitter, the greatest discriminating ability was 54.79 percent (X). Watts et al. [49] studied a professional singer's voice before and after medication. Medication reduces shimmer and jitter while increasing fundamental frequency (F0). Following this study, Guido et al [50] examined the performance of several DWTs in distinguishing between normal and diseased voices. It was 90% with Spikelet. Adaptive time-frequency transform decomposition of speech signals was proposed by Umapathy et al. [51] for classifying pathological voices. It had 93.4 percent category precision. For the clinical diagnosis of laryngeal paralysis, Zhang et al. [52] found that using both nonlinear dynamic analysis and classical perturbation analysis might help describe abnormal sounds.

Non-invasive LPC and MFCC-based features were used by Neto et al. [53]. These attributes helped them obtain classification accuracy of up to 85%, 80%, and 52%. Gomez-Vilda et al. [54], [55] used biomechanical parameters and statistical approaches such as PCA, LDA, and hierarchical clustering for pathological voice analysis. They believe that using biomechanical and acoustic factors together may effectively diagnose vocal disease. Zhang and Jiang [56] studied the acoustic properties of sustained and flowing vowels in healthy people and laryngitis sufferers. Sustained and flowing vowels were assessed using SNR, second-order entropy and correlation dimension. Normal and diseased voices differed significantly in Mann-Whitney rank sum tests. To diagnose speech pathology in real-time, Fontes et al. [57] showed that current feature extraction methods are too complicated. In this study, they introduced a new characteristic, correntropy spectral density, for the identification of laryngeal diseases. This method successfully classifies 97 percent of the MEEI voice samples.

The work by Gavidia-Ceballos and Hansen [58] advocated utilising HMM to identify vocal fold cancer. They explored computing increased spectral-pathology components, which do not require the calculation of glottal waveform, as previous publications do. The HMM is assessed on voice samples from healthy and pathological sustained vowel samples. The best accuracy in both healthy and sick voices is 92.8 percent, with a heightened spectral-pathology element. Arias-Londono et al. [59] employ HMM to convert short-term noise parameter characteristics into MFCC. The fundamental flaws of standard feature space transformations (such as PCA and MDA) are that they ignore temporal connections among data. To tackle this, a novel approach is suggested that obtains both transformation and classification stages concurrently and adjusts model parameters to minimise classification error. The method works with 96.61 percent accuracy on the MEEI database.

GMM was used in an article by Muhammad et al. [60] to classify audio recordings (Arabic Digit numbers) of patients with six different voice disorders: polyps, spasmodic dysphonia, laryngopharyngeal reflux illness, vocal fold cysts, and sulcus vocalis nodules. They showed that when sustained vowels are input, critical aspects including voice offset and voice onset qualities are neglected, which are crucial in detecting voice abnormalities. For continuous speech, they introduced a unique feature extraction approach called multidirectional regression that takes into consideration consonant and vowel positions, formant transitions, and voice start and offset distributions. The GMM classifies the characteristics with 95% accuracy. To discover and classify abnormal voices in the MEEI database, Ali et al. suggested a GMM classifier approach. These characteristics were derived using the auditory spectrum and all-pole model-based cepstral coefficients. Features of normal and diseased voices are observed and evaluated using a GMM classifier. Pathological voices are categorised into adductor, keratosis, vocal nodules, polyps, and paralysis. The auditory spectrum has a maximum disease identification and classification accuracy of 93.33 percent (adductor). The accuracy of the all-pole model's cepstral coefficients was 99.56 percent. Their key findings are as follows: While the research shows high accuracy for databases with prolonged vowels, flowing speech requires less labour and is more demanding. They emphasised the necessity of studying using continuous speech databases since they are more realistic. Processing continuous speech involves vocal activity detection (VAD), a difficult problem that leads to poor performance. The authors offer characteristics that do not need VAD. According to the literature review, MFCC is a useful characteristic for pathology identification but not for pathology categorization. Overall, they say their technique outperforms previous ongoing speech database tests without VAD. They also claim that the characteristics are aesthetically

pleasing and that the GMM classifier is optional. Ali et al. [61] presented a technique for detecting voice disorders using GMM by finding the source signal from speech using linear prediction analysis. The spectrum calculated using LP analysis characteristics shows the energy distribution in normal and diseased voices, allowing them to be distinguished. They found that lower frequencies, from 1 to 1562 Hz, help identify vocal problems. With sustained vowels, the algorithm achieved 99.94% accuracy and 99.75% accuracy with flowing speech.

A study by Behroozmand and Almasganj [62] evaluated the relevance of energy and entropy parameters taken from speech signals with unilateral paralysis of the vocal folds. In SVM with a linear kernel, the extracted characteristics are optimised genetically. Entropy characteristics have 100% accuracy, whereas energy features have 93.62%. Entropy characteristics have 32 active sub-bands, whereas energy features have 11. Using a set of 13 active sub-band entropy characteristics maximises the recognition rate. The researchers determined that entropy properties were useful in diagnosing laryngeal paralysis.

Markaki and Stylianou [63] detected pathology using an SVM classifier on MEEI database sustained vowel voice recordings and achieved 94.1 percent accuracy. They discussed the obstacles associated with some of the procedures, such as precise calculation of fundamental frequency and glottal waveform. They employed a modulation spectrum, a combination of acoustic and modulation frequency representations, to identify and classify voice abnormalities. These amplitude modulation patterns should be altered in vocal pathology, providing hints for diagnosis and categorization. Although the findings are intriguing, additional research is required on hospital datasets before such algorithms may be used in routine clinical practise. For example, Saeedi et al. [64] suggested extracting eight energy characteristics from

wavelet filter banks created using lattice factorization to distinguish between normal and diseased voices. Both MEEI and a private database had 100% classification accuracy using SVM classifiers. Arjmandi and Pooyan [65] compared short time Fourier transform, continuous wavelet transform, and wavelet packet transform with feature reduction approaches like PCA and LDA, as well as the SVM classifier. The researchers discovered that using entropy characteristics in the sixth level of WPT decomposition with LDA and SVM is the best technique for 100% recognition. Because diseased voices are nonlinear and unpredictable, entropy metrics help distinguish between normal and pathological voices. Uloza et al. [66] evaluated the efficiency of several feature sets in SVM voice classification. They classified nodular, normal, and diffuse lesion voices with over 90% accuracy using a sequential committee of SVM. The researchers evaluated the suggested method's findings with three human experts and found that it outperforms those using solely sustained vowels as a source of information.

In their study, Muhammad and Melhem [67] evaluate MPEG-7 audio low-level characteristics for disease detection and classification. They utilised an FDR to choose features. Nodules, polyps, keratoses, and adductors are all diagnosed using SVM. Saidi and Almasganj [68], [69]suggested using a linear kernel SVM classifier with M-band wavelet feature extraction to classify normal and diseased voices. They found the best four and five-band wavelets with 100% accuracy. Because the disorders' traits and symptoms vary, they have suggested several modelling methodologies. The speech signals are evaluated using four methods: noise content measurements, spectral cepstral modelling, nonlinear characteristics, and fundamental frequency stability assessments. Their study is to improve the understanding of voice pathologies. The SVM classification tests use six datasets, and the maximum classification accuracy obtained is 99%. To properly characterise voice recordings, it is critical to understand how pathology affects the tissues

involved. For example, periodicity characteristics may help with vocal fold vibration stability. The report also mentions that the presented approaches have limits and that deeper neural networks may be used for future tests.

Three sustained vowels were used by Benba et al. [70] to distinguish PD from healthy voices. To extract voice prints, they averaged the MFCC frames. The prolonged vowel /u/ includes more distinguishable analysis than other forms of voice samples. The suggested approach uses the SVM MLP kernel to attain 100% accuracy. Similarly, Benba et al. used voice samples to distinguish between neurological and PD illnesses using PLP, MFCC, and Spectral PLP. They got 90% classification accuracy using PLP and linear SVM kernels' first 11 coefficients. Ali et al. [71] use SVM classifiers to classify speech signals based on voice intensity. The peaks in the speech signals are used to construct a voice contour. The region beneath the vocal contour is used to distinguish normal from abnormal. Disordered voices have a smaller region beneath the vocal contour than normal voices. The suggested functionality avoids the need to estimate fundamental frequency. They utilised the King Abdul Aziz University database, which contains voice recordings of individuals with vocal fold cysts, polyps, unilateral vocal fold paralysis, laryngopharyngeal reflux illness, and sulcus vocalists, as well as normal people. The SVM multiband detection of vocal abnormalities was suggested by Ali et al. The fractal dimension (FD) of the power spectrum is evaluated in each band using a three-level DWT. FD is used to experiment with MDVP settings. There are 173 pathological and 53 normal voices in the MEEI database. They used 168 problematic voices since 5 of 173 had no MDVP characteristics. The findings are represented as SEN, SPE, ACC, and AUC. Combining FD of all levels with 22 MDVP parameters improved accuracy by 2.26 percent and the area under the curve by 1.45 percent.

By collecting five measures of irregularity from the vocal tract region, Muhammad et al. created an automated voice pathology detection method that uses voice production theory to identify voice pathologies. They are related to the glottis. For sustained vowel characteristics, the features recovered from pathological voice samples show inconsistent patterns across frames, aiding in voice classification. To summarise, supraglottic contributions outnumber vocal tract contributions, while vocal tract tube variation across utterances outnumbers mean vocal tract tube variance. On MEEI and SVD datasets, SVM classifies the derived features. MEEI database: 99.22 0.01, SVD: 94.7 0.021. For automated identification of voice disorders from diverse datasets, Al-Nasheri et al. [68] used SVM to analyse MDVP parameters. On the basis of these databases, they looked at three common vocal problems (cyst, paralysis, and polyp): The FDR approach is used to rank the MDVP parameters collected from a computerised speech lab. A t test is used to compare the means of normal and diseased samples to identify significant differences. Using three datasets, they found significant differences in MDVP parameter performance. The greatest accuracies found for the SVD, MEEI, and AVPD are 99.68%, 88.21%, and 72.533%, respectively. They used SVM classifiers and correlation functions to identify and classify speech pathologies in distinct frequency ranges. Correlation functions estimate the peak and its lag. The frequency bands are varied to see how they affect the automated detection and categorization. Vocal problems are more easily detected and classified in the 100–8000 Hz frequency range. From one database to another, detection and classification accuracy differed. The MEEI, SVD, and AVPD databases have the greatest detection rate at 99.81%.

The quality of speech production is affected by vocal fold pathology. The acoustic analysis of healthy and pathological voice signals as an alternative to classical diagnosis methods has been introduced recently [72].

Nonlinear dynamical analysis methods have been widely used in the study of normal and pathological vocal tract systems. The effect of non-stationary noise on the calculation of D_2 and K_2 of normal and pathological signals was first investigated by [73]. [74] pointed out that there was a noticeable difference between the fractal dimension of the electroglottographic signals affected by Parkinson's disease and those of healthy individuals. [75] proved that the largest Lyapunov exponents could be used as a tool for discriminating healthy voices and pathologic voices from patients with laryngeal paralysis. [76] revealed that both the D_2 and K_2 of pathologic human voice signals showed a statistical reduction after surgical excision of vocal polyps. [52] used D_2 and K_2 at minimum embedding dimension for the analysis and detection of voice disorder.

Amami and Smiti [77] introduced incremental DBSCAN-SVM to identify sounds, evaluate, and categorise pathological audio from normal voice. It can find clusters of any form, identify noise, employ spatial access techniques, and is efficient even for big geographical datasets. They employed SVM with a Gaussian function kernel to classify data from MEEI with 98% accuracy. The suggested technique can manage incremental and dynamic speech databases that change with time. Al-nasheri et al. [68] investigated multiple frequency bands utilising autocorrelation and entropy to build an efficient feature extraction for identifying and categorising voice disorders. Using autocorrelation, peak and lag values were retrieved from each frame of a spoken signal. The SVM classifier is used on the MEEI, SVD, and AVPD datasets. They used U tests to compare the means of normal and diseased samples. The detection and classification accuracy vary per frequency range and database. The most useful bands for detection and categorization were between 1000 and 8000 Hz. Detection accuracy was 99.69%, classification accuracy was 99.54%, detection accuracy was 92.79%.

2.4 Review on Noise Identification

At various stages in the acquisition process, audio gets contaminated by various types of noise. In audio, the goal of noise reduction is to reduce the amount of noise present without affecting the quality of the underlying signal. For the efficient reduction of noise, it is important to know the type of noise included in the speech signal. Various methodologies for noise identification and noise removal are reviewed in this chapter.

In Bayram et al. [78], a technique for audio denoising based on wavelet transforms is explored. The authors focused on audio signals that had been damaged by white noise. White noise is particularly difficult to eliminate as it may be found at all frequencies. The Discrete Wavelet Convert (DWT) was utilised by the authors to transform a noisy audio input into a wavelet domain. Using coefficient thresholding and transforming them back to the time domain, it is possible to get an audio signal that has less noise than the original. The most important criteria for evaluating experimental outcomes was the objective degree grade (ODG). The authors of the article [79] examine block attenuation approaches, which were first employed in orthogonal wavelet signal representations, and then in orthogonal wavelet signal representations. They discovered that block attenuation effectively displaces the remaining noise artefacts in restored signals. Several writers have investigated the relationship between the decision-directed a priori SNR estimator and the block attenuation developed by Ephraim and Malah.

A comparison of the performance of adaptive block attenuation with that of standard thresholding operators reveals that it performs very well. Despite the fact that short-time Fourier denoising outperforms its wavelet counterpart for stationary sections when high pitch is included, when low pitch is involved, it outperforms its wavelet counterpart by a factor of two. This is because short-time Fourier has a higher frequency resolution than

wavelet representation in high frequency bands. The denoising issue is examined in the article [80] from the perspective of sparse atomic representation. To this end, the authors suggested a comprehensive framework of time-frequency soft thresholding that includes and links well-known shrinkage operators that are used in many applications as specific examples. According to signal-to-noise ratio, the innovative technique is competitive with existing approaches. From the perspective of denoising, the neighbourhood weighting might be seen as a non-diagonal estimate method. These techniques are effective in reducing the musical noise that naturally occurs in diagonal estimation. Using the persistence features of signals as described in Article [81], significant gains in audio denoising are achieved. We present a new denoising operator based on neighbourhood smoothed Wiener filter-like shrinkage that is developed from the neighbourhood smoothed Wiener filter. Plain linear models, which vary in performance based on the quantity of noise, produce only slight improvements in performance over the optimum thresholds. In the event that the noise level is unknown, a straightforward approach for calculating it is suggested. Although they perform well when compared to current operators, the suggested operators are much more effective and resistant to modest perturbations in the noise level.

In order to reduce noise, spectral audio denoising algorithms typically use time-frequency representation magnitudes of the signal, as presented in the article [82]. Matching Pursuit (MP) is a potential method that repeatedly constructs a sparse signal representation from a signal. In the work [83], a study of the game Matching Pursuit is offered in the context of audio denoising techniques. The technique is understood as a straightforward shrinkage strategy, and the authors have highlighted aspects that are crucial to its success. They have also provided many ways to increase the algorithm's performance and resilience. The authors have given experimental findings on a diverse variety of speech signals and shown that the technique is capable of

producing results that are comparable with those obtained by existing audio denoising algorithms.

In recent years, nonlinear techniques have emerged as a viable option in the field of voice and image processing [84], [85], [86]. Speaker identification, emotion detection, voice synthesis, and speech processing are just a few of the applications of multifractality in speech time series data that have been explored in the field of speech. In combination with multifractal variables, the LPC and MFCC give high accuracy in speaker identification [87] . In noisy environments, a combination of Gammatone Frequency Cepstral Coefficients (GFCC) and Multifrequency Cepstral Coefficients (MFCC) is employed to validate the identity of the speaker [88]. According to Zhang et al. [76], both the correlation dimension and the correlation entropy of human voice signals showed a statistical decrease after surgical ablation of vocal polyps, and both metrics are reasonable approximations of the pathological voice analyses. According to Huang et al. [52], correlation dimension and correlation entropy were used at the smallest embedding dimension for the research and identification of voice disorder. Machine learning approaches combined with logistic regression have been used to predict a variety of different diseases [89], [90]. It has been proved that combining MFCC with vector quantisation increases the estimate of back ground noise [91]. Using multi-fractal de-trended fluctuation analysis, Sarkar and colleagues [92] were able to overcome the language dependence in speaker recognition while working with Bengali. A recent study [93] investigated the use of multifractal spectrum analysis and matching pursuit for the detection of audio magnetotellaric signal noise. In recent years, multi scale chaotic speaker identification systems, as well as the accuracy challenges associated with them, have been a popular study topic [94]. Despite the fact that a significant number of parameters are employed to

estimate noise in speech, the multifractality of noisy time series has yet to be exploited.

2.5 Review on Emotion Recognition

Over the past several years, an extensive inquiry into the recognition of emotions from speech data has been carried out. Narayanan [95] presented domain-specific emotion identification using voice inputs from a contact centre application, which he claims is feasible. The primary emphasis of this study is on the detection of negative and non-negative emotions (for example, rage and happiness). For emotion recognition, many sorts of information are employed, including discourse, lexical, and auditory information. In order to deal with various kinds of features, both linear and k-NN discriminant classifiers are used. The experiments, by using both auditory and verbal data in conjunction, give better results. According to the findings, when three information sources are used instead of one, categorization accuracy rises by 36.4 % for females and 40.7 % for males. When compared to previous literature, females' recognition rate increases from 0.75% to 3.96 % and males' accuracy increases from 1.4 %to 6.75 %.

Nwe et al. [96] presented a new approach for the categorization of voice samples based on their emotional content. To describe the speech signals and train the classifier, the system used a discrete high-order moving average using LFPC (short time log frequency power coefficients). This technique categorised six emotions, and then a private system was utilised to test and train the proposed system, which was subsequently implemented in production. The LFPC is evaluated in comparison to the MFCC and LPC in order to determine the effectiveness of the suggested technique (LPCC). The results show that the best and average recognition accuracy for the whole sample was 96 percent and 78 percent respectively.

A Gaussian Mixture Vector Auto-Regressive (GMVAR) method was introduced by El Ayadi et al. [97], which is a mixture of Gaussian mixture models and vector autoregressive models. The fundamental idea of GMVAR is its ability to distribute information in several modes and to define the relationship between speech feature sets. When compared to HMM, this approach has the benefit of greater discrimination between low and high arousal with neutral emotions.

It has been proposed by Rong et al. [98] that a group of random forest to trees (ERFTrees) technique may be used without the need to resort to linguistic information for emotion identification; nevertheless, this is still an open subject. A small amount of data with a large number of characteristics may be processed using this approach. It was determined via an experiment that this strategy improved the rate of emotion detection in a Chinese emotional speech dataset, which was used to test its efficacy. Aside from that, ERFTrees outperform standard dimension reduction approaches such as MDS (multi-dimensional scaling) and PCA (principal component analysis), as well as the freshly created ISOMap. For the female dataset, the highest recognition rate was reached with 16 features, which resulted in a maximum right rate of 82.54 percent, while the poorest accuracy was achieved with 84 features, which resulted in only 16 percent of correct answers.

Lee and colleagues [99] developed a hierarchical framework for binary decision trees in the realm of emotion identification. This strategy is based on locating the easiest-to-recognize hindrance at the upper class of the tree in order to reduce the accumulation of errors in the classification process. This structural technique, in addition to mapping incoming voice data into one of the classes of emotion, does so via a further binary classification layer. The outcomes using the AIBO database demonstrates an absolute improvement of 3.3 percent over the baseline model in [100], which archives 65.1 percent and

70.1 percent for the five-class and two-class problems, respectively, in the AIBO database. Measurement of probability as a soft label may be used as an alternate option to out-putting hard labels at each stage, which can result in a strong modelling solution. Yang and Lugger [101] provided a unique set of harmonic characteristics for the identification of emotional expressions in speech. In the beginning, starting with the expected pitch of a voice signal, the spherical autocorrelation of the pitch histogram is computed. The accuracy increased by 2% on average

Wu et al. [102] suggested a method for human voice emotion identification using modulation spectral features (MSFs). With the use of this technology, we were able to collect temporal and acoustic modulation frequency components that might be used to transmit critical information that was previously unavailable via standard short-term spectral characteristics. An SVM with a Gaussian kernel function is used in the process to improve accuracy. MSFs are evaluated at Vera Am Mittag (VAM) and Berlin, which are both located in Germany. As shown by the results of the experiments, MSFs outperform PLPC and MFCC in the experiments. When MSFs are used to compliment prosodic features, there is a significant increase in the overall performance of the recognition system. Furthermore, an accuracy of 91.6% has been obtained for classification.

In order to recognise emotions, Lee et al. [103] constructed a hierarchical computational framework. Through the use of successive layers of binary classifications, this approach transforms the input speech sample into the emotion class that corresponds to it. The basic notion behind the various levels in a tree is to complete the classification job in the simplest possible way in order to reduce error propagation. The categorization algorithm is evaluated using the AIBO and USC IEMOCAP datasets. The absolute increase in accuracy over the baseline SVM is 72.44 percent to 89.58

percent, with an absolute improvement of 72.44 percent to 89.58 percent. As a result, the published hierarchical strategy for categorising emotional speech in diverse datasets has been shown to be effective.

Albornoz et al. [104], [16], investigated towards a novel spectral signature that may be used to assess emotions and describe groups. It is proposed in this study that emotions may be categorised using auditory parameters and a unique hierarchical classifier. The experimental findings on the Berlin dataset demonstrate that the hierarchical approach outperforms traditional classifiers when compared to the same dataset with no hierarchy. For example, the regular HMM technique had a performance of 68.57 percent, whereas the hierarchical model achieved a performance of 71.75 percent.

According to Wu et al. [105] , a fusion-based approach that uses acoustic-prosodic (AP) properties in conjunction with semantic labels to recognise various kinds of emotions in speech has been developed (SLs). First, the AP features are extracted, and then three distinct types of base-level classifiers are used to categorise the AP features that have been extracted in this manner. The semantic labelling approach makes use of the maximum entropy model, which increases the number of potential outcomes to the greatest extent feasible. A private dataset was used for the experiments, and it was discovered that the performance based on MDT acquired is 80 percent, the performance of the same dataset based on SL recognition acquired is 80.92 percent, and the performance of the same dataset based on a blend of AP and SL acquired is 83.55 percent.

By using a three-level speech emotion recognition approach, Chen et al. [106] enhanced speaker-independent environment speech emotion recognition. This approach classifies distinct emotions into coarse and fine categories, and then selects the most relevant feature based on the Fisher rate.

ANN and PCA are used to decrease the dimensionality of four comparison trials and to classify them, respectively, using the results. A consequence of this is that Fisher outperforms PCA in dimension reduction and SVM outperforms ANN in classification for emotion identification in speaker independent experiments. The recognition rates for three levels of the Beihang University Database of Emotional Speech (BHUDES) are 86.5 percent, 68.5 percent, and 50.2 percent, respectively.

A computational technique for emotion identification and analysis of emotion specifications in vocal social media was suggested by Dai et al. [107]. The experimental findings reveal that the recognition rates for various emotions are diverse. It has been discovered that the average rate of identification reaches 82.43 percent, which is the highest rate of recognition ever discovered in a similar inquiry. A ranking SVM was suggested by Cao et al. [10], who discuss an approach for synthesising emotion information detection in order to tackle the challenge of binary classification by combining information from several sources. This ranking approach high accuracy in the training and testing steps. Ranking-based SVM algorithms performed well over standard SVM algorithms in detecting emotional speech samples in both spontaneous and acted data, which included neutral emotional voices. A 44.4% accuracy in unweight average (UA), also known as balance accuracy [104], was attained.

Because the properties of speech change throughout the creation of emotional speech, attributes that capture these fluctuations may be utilised to identify emotional states. The fundamental frequency, energy shape, quiet duration, formant, Mel-band energies, cepstral coefficients of linear prediction, Mel Frequency cepstral coefficients, and voice quality are the most often used features for the categorization of emotions [96], [108]–[112]. It was established in an experiment done by Ramamohan and Dandapat [113]

that the sinusoidal features of sounds may be utilised to distinguish emotions. It is possible to categorise the acoustic characteristics employed in SER into two categories: prosodic features and spectral features. Prosodic features, which are often utilised in SER [114] are used to convey the speaker's emotional state. Pitch and energy tracking contour statistics [97], [115] are often used to estimate the properties of pitch and energy tracking contour statistics. Spectrum features, which are often derived from the speech spectrum, have received an increasing amount of attention in the last several years. It is possible that these characteristics will aid in identification by giving extra information for prosodic features [102] . In most cases, linear source-filter models [116] are used to estimate both the prosodic and spectral characteristics of the human speech production system.

2.6 Conclusion

The research on recent developments in pathological, noisy, and emotional studies in speech is examined from several angles. A thorough examination of the audio speech database was conducted. The nonlinear approaches in pathological voice classification are studied in detail. Both noise removal and noise identification from different perspectives are addressed. The linear and nonlinear methods available in the literature for emotion recognition are presented in the chapter. The need for improvement in classification systems is pointed out.

CHAPTER 3

MECHANISM OF SPEECH PRODUCTION AND CREATION OF MALAYALAM SPEECH DATABASE

3.1 Introduction

In the last few decades, humans have interacted with machines in almost every area of their lives. To improve human-computer interaction, the machine must function similar to how a human interacts with his environment. The basic source of speech recognition is the audio signal from the mouth. To study the mechanism behind speech production, it is necessary to know about the anatomy and physiology of the speech production system. Since the system study should be carried out from the signal output, the need for a standard database is inevitable.

A wide range of standard databases have been reported, with the majority claiming to be useful for the task at hand. The database that needs to be created should have a variety of features that will be useful for various research projects. The basic need for creating a voice database is to have a vast phonetically balanced speech corpus uttered by a large number of different speakers in an uncontrolled environment. For both training and testing methods, a considerable duration of annotated recording of the speech utterance is required to construct an efficient speech-based application system. It is necessary to determine the peculiarities of the database's language and its linguistic history, as well as to compare it to other language groups in order to resolve issues that arise during the database's formation.

This chapter seeks to provide a Malayalam audio speech database that was captured in diverse situations for a range of research purposes, with a focus on the study of nonlinearity of the speech production system using the database. A database with a large number of samples is necessary to generalise the nonlinear properties of the system phase space and to optimise

the embedding parameters. It can be used for other speech-based applications as well, despite the fact that it was built for this specific purpose. The database contains audio speech recorded in a controlled environment from 100 females and 100 males in three age groups, speaking 50 Malayalam single phonemes and 207 connected sentences containing all allophonic variations. Each isolated phoneme and word in the audio is segmented and labelled accordingly. This work is a modest attempt to develop an open-access audio Malayalam speech database for researchers to use. Any assistance with the updating of this database is much appreciated.

The content of this chapter is organised as follows. The anatomy and morphology of speech production is discussed in Section 3.2. The linguistic aspects of the Malayalam language are explained in session 3.3. The recording setup and time domain representation of the recorded signals are described in Session 3.4. In section 3.5, the database's segmentation and labelling process is described. The work is concluded in section 3.6.

3.2 Speech Production

Lungs, larynx, and vocal tract are the three primary groupings of speech organs. The lungs provide air flow to the larynx stage of the speech production system and act as a power source. The larynx, a complex system of cartilages, muscles, and ligaments that controls the vocal cords during speech production, is a complicated system of cartilages, muscles, and ligaments. The oral cavity, which runs from the larynx to the lips, and the nasal route, which is connected to the oral tract by the velum, makes up the vocal tract.

When humans speak, air is driven out of their lungs through the trachea, larynx, and vocal tract, which has two openings (mouth and nose) and varying constrictions for different sounds. The glottis is the start of the vocal tract, measuring roughly 17 cm in length and extending to the lips. The vocal folds, also known as vocal cords, are two small muscular folds that

open and close in the larynx. The glottis is a slit-like opening that connects two folds. The vocal cords are tensed muscle tissues that vibrate when the folds are partially closed but not when they are open. Between the velum and the nostrils is a chamber called the nasal cavity. Voiced sounds are produced when the folds are apart (no vibration), and unvoiced noises are produced when the folds are apart (no vibration). Airflow is chopped into quasi-periodic pulses during vibration, which are then modulated in frequency by passing through the pharynx, mouth cavity, and nasal cavity—different sounds are produced depending on the position and manner of different articulators (lips, teeth, tongue, alveolar, and palates). Phonemes are divided into vowel phonemes, diphthong phonemes, and consonant phonemes in linguistic terms. The human speech production system is depicted schematically in Fig. 3.1.

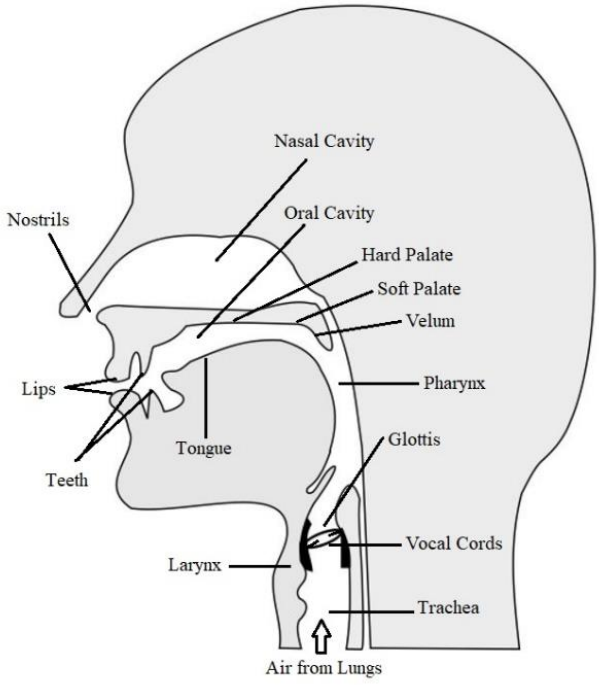


Fig. 3.1 Cross sectional view of Speech Production System

3.3 Language Material

India has 23 official languages that are recognised under the constitution. The Central Government recognises Hindi and English as

official languages. Malayalam is a Dravidian language spoken in Kerala, Lakshadweep, and Mahe, and is spoken by 38 million people globally. In 2013, it was declared as a classical language. The Malayalam alphabet has several letters among the Indian Language orthographies due to its genealogy coming from both Tamil and Sanskrit.

The number of phonemes varies between dialects and languages; for instance, British English has 44, Indian English has 38 [117], [118], and Malayalam has 50. The participation of articulators in the speech production system is used to classify the phonemes. In Malayalam language phonemes and allophones are linguistically categorized according to articulation points and manners as in [http://www.cmltemu.in/phonetic/#/], an inclusive Malayalam phonetic archive owned by Thunchath Ezhuthachan Malayalam University (TEMU), Kerala, India [119]. The audio file in this archive is utilized to understand the pronunciation of phonemes and words. In the Malayalam language, there are ten vowel phonemes, two diphthongs, and 38 consonant phonemes.

The fundamental building blocks of any language are phonemes and allophones. Phonemes are the relatively distinct and fundamental utterances of a language [1].Phonemes are classified as vowel phonemes, consonant phonemes, and diphthong phonemes. An allophone is a version of a phoneme that is phonetically distinct from another[2]. The place or phonetic surroundings in the word generally characterises the allophones of the same phoneme. Because these differences do not assist in identifying one word from another, speakers of a language sometimes have trouble recognising the phonetic differences between allophones of the same phoneme.

Vowel phonemes are produced by the vocal tract, which is powered by quasi-periodic air pulses produced by the vocal cord's vibrations. Vowels are voiced sounds that have less constriction in the vocal tract, have a longer

duration, and are generally louder than other phonemes. The height and position of the tongue, as well as the shape of the lips, are used to classify vowels. Table 3.1 shows the classification of vowel phonemes based on the position of the tongue and lips.

Table 3.1 Linguistic Classification of Vowel Phonemes

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

Consonant phonemes are created by articulators regulating the flow of air, and they can be voiced or unvoiced. The mode of articulation and the place of articulation distinguish consonants. The 38 consonant phonemes are divided into eight groups based on the position of articulation. The point of articulation is the location in the mouth cavity where the constriction is produced. Consonant phonemes appear in Malayalam as a Consonant-Vowel (CV) unit called a syllable (ക< ka > = ക്/k/ + അ /a/). Consonant phonemes must be segmented from the CV unit before the analysis of speech signal.

Bilabial, labiodental, dental, alveolar, retroflex, palatal, velar, and glottal are the eight classes of consonant phonemes based on the position of articulation. The two lips come together to make bilabial consonant phonemes, which are formed by restricting air movement. Labiodental consonant phonemes are created by restricting airflow and contacting the

upper teeth with the lower lips, and by releasing the air by expanding the mouth. Dental consonant phonemes are produced by placing the tip of the tongue behind the teeth. Limiting airflow by placing the tongue tip against the alveolar ridge produces the alveolar consonant phoneme. When the tongue articulates between the alveolar ridge and the hard palate, it makes a retroflex sound. It is Malayalam's largest consonant phoneme group. A constriction between the tongue's tip and the roof of the mouth produces palatal consonant phonemes (palate). Velar consonant phonemes are created when the back of the tongue contacts the velum (soft palate). The glottal consonant phoneme is created by sealing the rear of the glottis.

The extent of constriction made for a consonantal gesture is represented by the manner of articulation. Consonants in Malayalam were divided into four groups based on how they were articulated: Plosives (Stops), Nasal, Fricatives, and Semivowels. Plosives (Stops) are formed by totally restricting the airflow for a short period of time, referred to as closure, and then releasing the air, referred to as release. Nasal sounds are the product of the oral and nasal tracts working together. Fricative sounds are produced when the airflow through the vocal tract is partially constrained. It is caused by a constant turbulent airflow at the constriction point, which produces a hissing tone and stimulates the vocal tract. Semivowels are vowel-like sounds created by gliding vocal tract area function between adjacent phonemes, such as vowels and diphthongs. There are 23 voiced consonants (Voiced Plosives, Nasals, and Semivowels) and 15 unvoiced consonants (Unvoiced Plosives and Fricatives). Table 3.2 shows a detailed classification of Malayalam consonant phonemes.

Table 3.2 Linguistic Classification of Malayalam Consonant Phonemes

Place of Articulation	Manner of Articulation									
	Plosive				Nasal	Fricative	Semivowel			
	Unvoiced		Voiced				Trill/ Flapped	Lateral	Approximant	Glide
	Unaspirated	Aspirated	Unaspirated	Aspirated						
Bilabial	പ്/P/	ഫ്/pʰ/	ബ്/b/	ഭ്/bʰ/	മ്/m/					
Labiodental										വ്/v/
Dental	ത്/t/	ഥ്/tʰ/	ദ്/d/	ധ്/dʰ/	ന്/n̪/					
Alveolar	റ്/r̺/				ന്/n/	സ്/s/	ര്/r/ റ്/ɾ/	ല്/ൽ /l/		
Retroflex	ട്/ʈ/	ഠ്/ʈʰ/	ഡ്/ɖ/	ഢ്/ɖʰ/	ണ്/ɳ/	ഷ്/ʂ/		ള്/ൾ /ɭ/	ഴ്/ʐ/	
Palatal	ച്/c/	ഛ്/cʰ/	ജ്/ɟ/	ഝ്/ɟʰ/	ഞ്/ɲ/	ശ്/ʃ/				യ്/y/
Velar	ക്/k/	ഖ്/kʰ/	ഗ്/g/	ഘ്/gʰ/	ങ്/ŋ/					
Glottal						ഹ്/h/				

Diphthong phonemes are speech sounds produced by the vocal tract smoothly switching between two vowel configurations. There are two diphthongs in Malayalam: /ai/ and -/au/. There are 106 Malayalam allophones in 207 words, including 75 consonant allophones, 28 vowel allophones, and three diphthong-related allophones. The list of Malayalam vowel and consonant phonemes, as well as their allophones, is shown in Tables 3.3 and 3.4.

Table 3.3 Malayalam Vowel and Diphthong Phonemes with its Allophonic variations

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

Table 3.4 Malayalam Consonant Phonemes with its Allophonic variations

SI. No.	Consonant Phoneme IPA	Consonant Allophone IPA	SI. No.	Consonant Phoneme IPA	Consonant Allophone IPA	SI. No.	Consonant Phoneme IPA	Consonant Allophone IPA
1	പ് /P/	[p] [β] [b] [P]	14	സ് /s/	[s]	27	ഛ് /c^h/	[c^h] [C^h]
2	ഫ് /p^h/	[p^h]	15	ര് /r/	[r]	28	ജ് /ɟ/	[J] [j]
3	ബ് /b/	[B] [b]	16	ഺ /ɾ/	[ɾ]	29	ഝ് /$ɟ^h$/	[$ɟ^h$]
4	ഭ് /b^h/	[b^h] [$m̥^h$]	17	ല് /ൽ /l/	[l] [ɖ]	30	ഞ് /ɲ/	[ɲ]
5	മ് /m/	[M] [m] [m̥]	18	ട് /ʈ/	[ɽ] [ʈ] [T]	31	ശ് /ʃ/	[ʃ]
6	വ് /v/	[w] [v]	19	ഠ് /$ʈ^h$/	[$ʈ^h$] [T^h]	32	യ് /y/	[y]
7	ത് /t/	[t] [t"] [ð] [d̪]	20	ഡ് /ɖ/	[ɖ]	33	ക് /k/	[k] [kj] [ɣ] [ɠ] [ʈ] [K]

8	ഥ/t^h/	[t^h]	21	ഝ /$ɖ^h$/	[$ɖ^h$]	34	ഖ /k^h/	[k^h] [K^h] [K^h]
9	ദ /d/	[ḍ] [d]	22	ണ /ɳ/	[ɳ]	35	ഗ /g/	[G] [g]
10	ധ /d^h/	[d^h]	23	ഷ /ʂ/	[ʂ]	36	ഘ/g^h/	[g^h]
11	ന /n̪/	[n̪] [n]	24	ള/ൾ/ɭ/	[ɭ]			[ŋ] [ŋj]
12	ഠ /ṛ/	[d] [t]	25	ഴ /ʐ/	[ʐ]	37	ങ /ŋ/	[ŋ<] [ŋ>] [ŋ']
13	ന /n/	[n^h] [n]	26	ച /c/	[c] [ç] [ɟ] [C]	38	ഹ /h/	[H] [h]

3.4 Database Acquisition

Native speakers from northern Kerala contributed to the database. The initial job is to construct a clean audio speech database by recording 50 Malayalam isolated phonemes and 207 linked sentences by speakers. The participants in the recording include 20 males and 20 females aged 5-10, 50 males and 50 females aged 20-25, and 30 males and 30 females aged 60-65. Recordings were scheduled in a confined environment. Every phoneme and word are repeated a total of ten times. For recording, two main types of audio capture devices are employed. First, sound is captured in a controlled environment using a standard headset with a microphone. Second, in the acoustically realistic environment, a regular mobile headset with a microphone was used. The audio recordings are saved as wav files at three different sampling rates: 16 kHz, 32 kHz, and 44.1 kHz.

Each stage of recording begins with a description of the objectives to be met. The speakers should close their mouths at the beginning and ending of each utterance. Speakers commence speaking a few seconds after the recording equipment starts recording in order to catch background noise and channel distortion at the start of the speech. The majority of the data has been captured in a studio-like setting, with options to record the signal in both controlled and uncontrolled environments.

An audio-only speech database is captured in two conditions to study the effects of real-time noise in speech signal processing and the effects of ageing on human speech organs. For the first task, the audio signal is recorded in an acoustically isolated laboratory environment using a standard headphone with a microphone near the mouth region, enabling good audio acquisition. Before recording, the channel's gain is limited, thereby attaining fewer background noisy signals and reducing the channel clipping issues. The speakers are requested to repeat each utterance ten times and the recordings

are saved as a single wav file for each phoneme and allophone. In addition, speakers are requested to repeat the utterance if necessary, either because of a mistake during articulation or if the recorded sound contains noise due to respiration or channel problem. For the second task, the audio signal is recorded in an acoustically realistic environment (office, school, and house) using an ordinary mobile headset with a microphone. The speakers uttered five short vowels ten times each.

The normalized signal values of the recorded signals are plotted against time for each phoneme. The time domain representation of short vowels is shown in Fig. 3.2 and the same for diphthongs are shown in Fig. 3.3.The time domain representation of consonant-vowel unit corresponding to eight different types of consonant phonemes are shown in Fig. 3.4 to Fig 3.11.The consonant boundary is indicated by the red line.

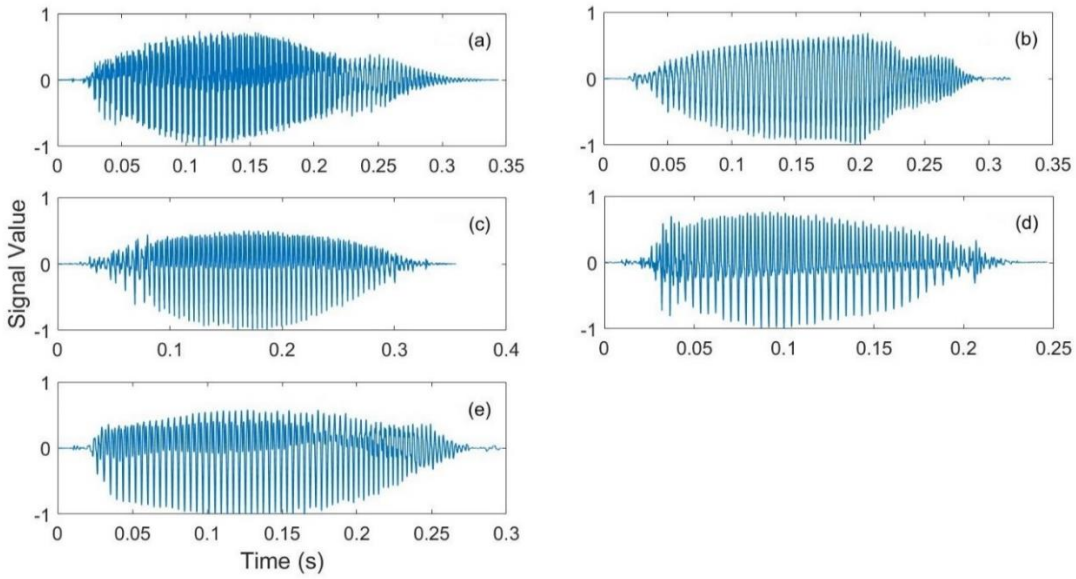


Fig. 3.2 Recorded Short Vowel Phonemes: (a) അ /a/, (b) ഇ /i/, (c) എ /e/, (d) ഒ /o/ and (e) ഉ /u/.

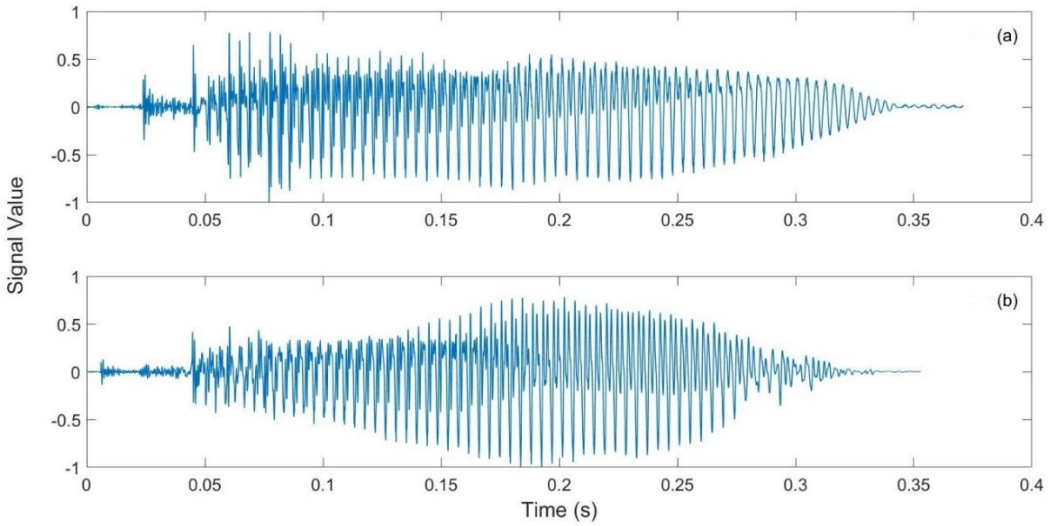


Fig. 3.3 Recorded Diphthong Phonemes: (a) ഐ /ai/ and (b) ഔ-/au/.

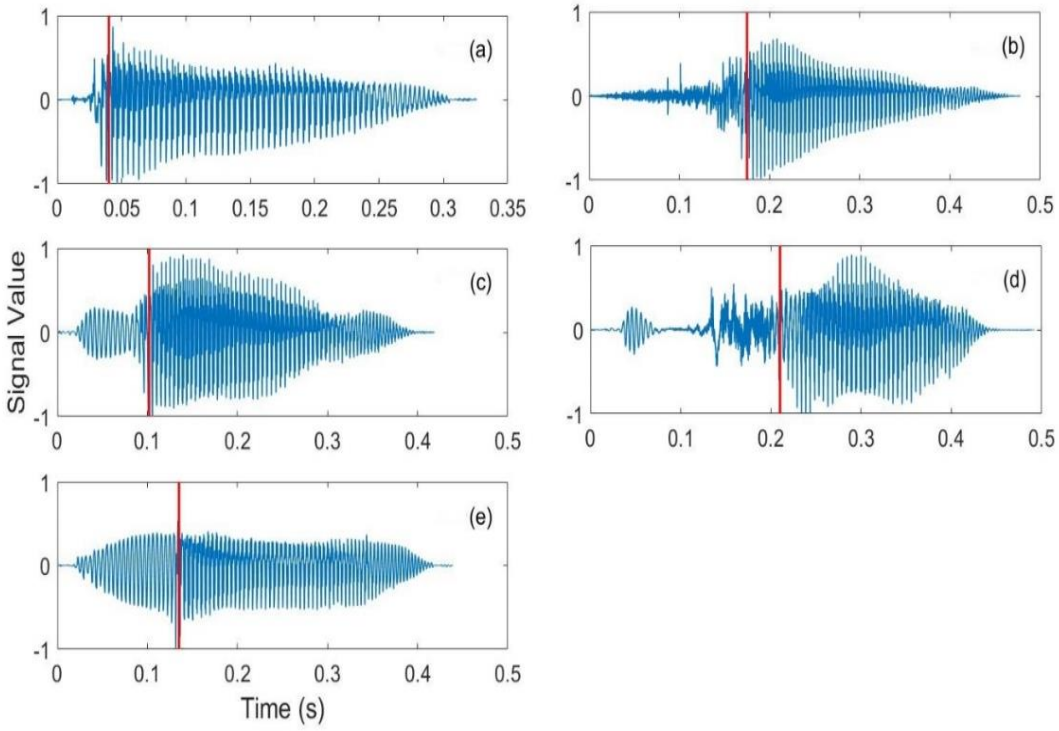


Fig. 3.4 Recorded Bilabial Consonant Phonemes: (a) പ്/P/, (b) ഫ്/p^h/, (c) ബ്/b/, (d) ഭ്/b^h/ and (e) മ്/m/.

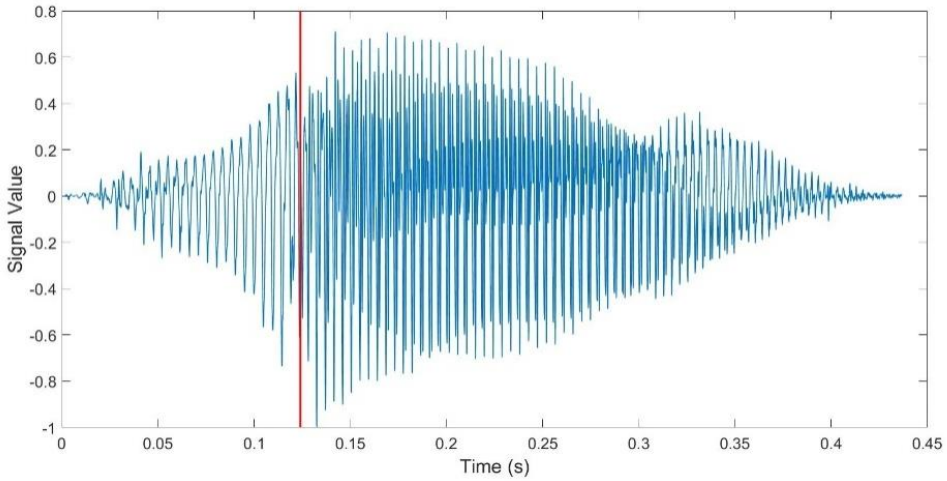


Fig. 3.5 Recorded Labiodental Consonant Phoneme: വ്/v/.

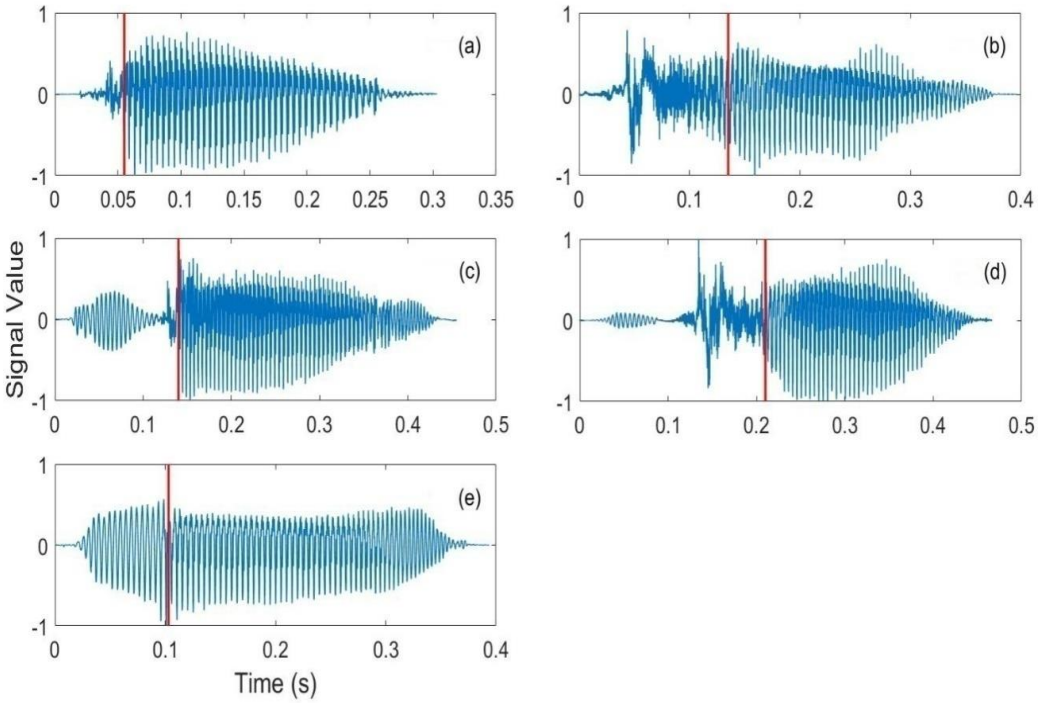


Fig. 3.6 Recorded Dental Consonant Phoneme: (a) ത്/t/, (b) ഥ്/t^h/, (c) ദ്/d/, (d) ധ്/d^h/ and (e) ന്/n̪/.

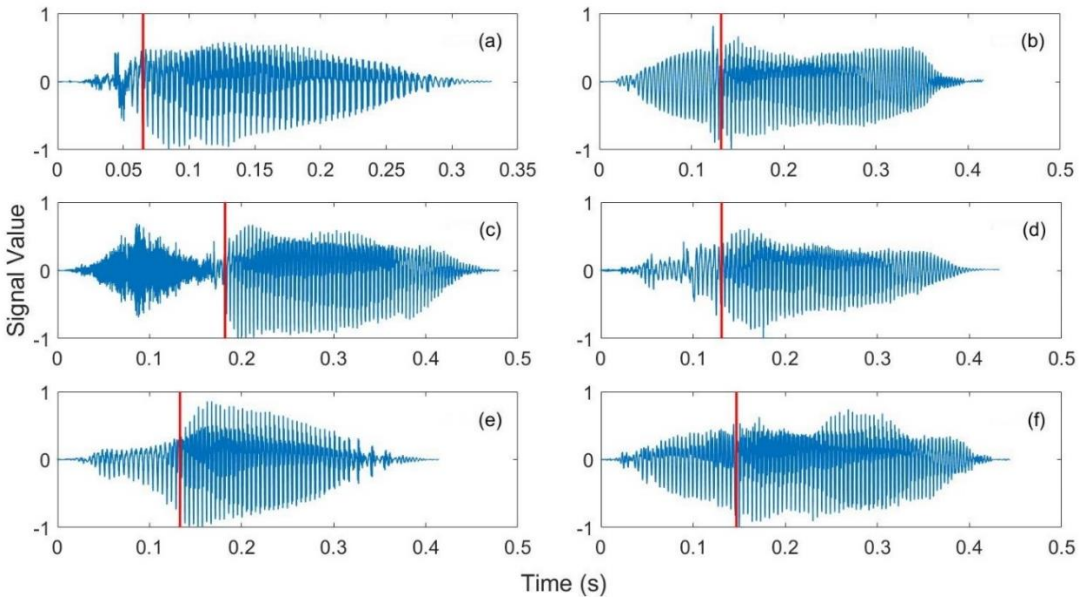


Fig. 3.7 Recorded Alveolar Consonant Phonemes: (a) ർ /r̪/, (b) ൻ /n/, (c) സ് /s/, (d) ര് /r/, (e) ർ /ṛ/and (f) ല്/ൽ /l/.

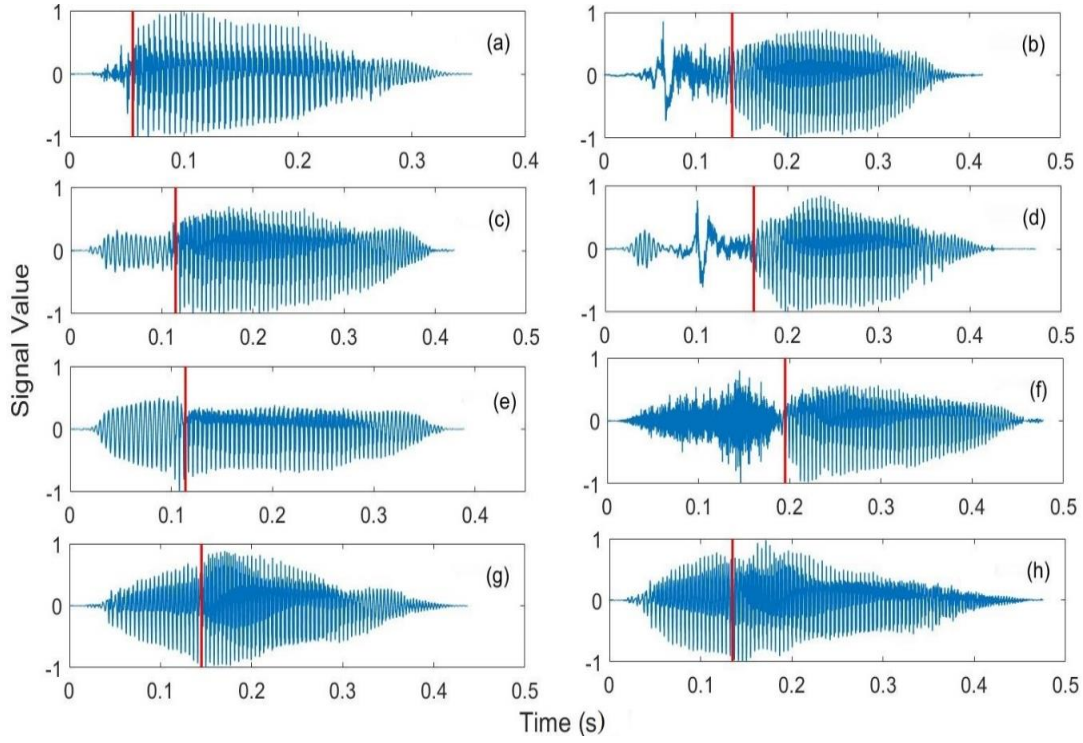


Fig. 3.8 Recorded Retroflex Consonant Phonemes: (a) ട്/ʈ/, (b) ഠ്/ʈʰ/, (c) ഡ് /ɖ/, (d) ഢ് /ɖʰ/, (e) ണ് /ɳ/, (f) ഷ് /ʂ/, (g) ള്/ൾ/ɭ/ and (h)ഴ് /ʐ/.

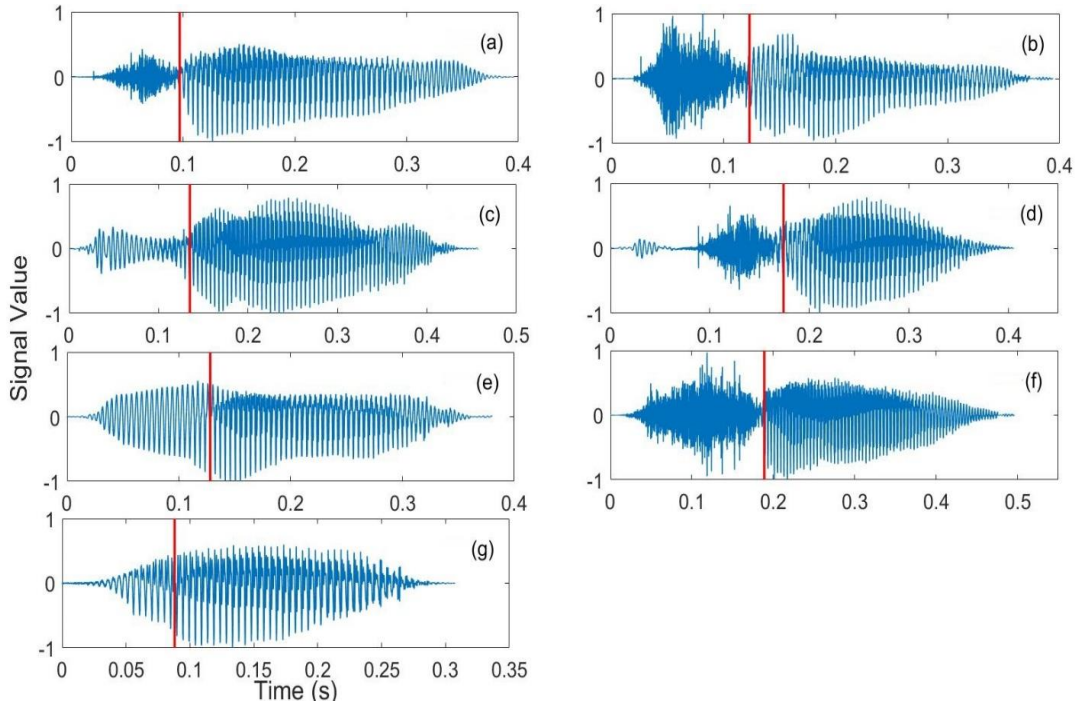


Fig. 3.9 Recorded Palatal Consonant Phonemes: (a) ച് /c/, (b) ഛ് /c^h/, (c) ജ് /ɟ/, (d) ഝ്/$ɟ^h$/, (e) ഞ് /ɲ/, (f) ശ് /ʃ/ and (g) യ് /y/.

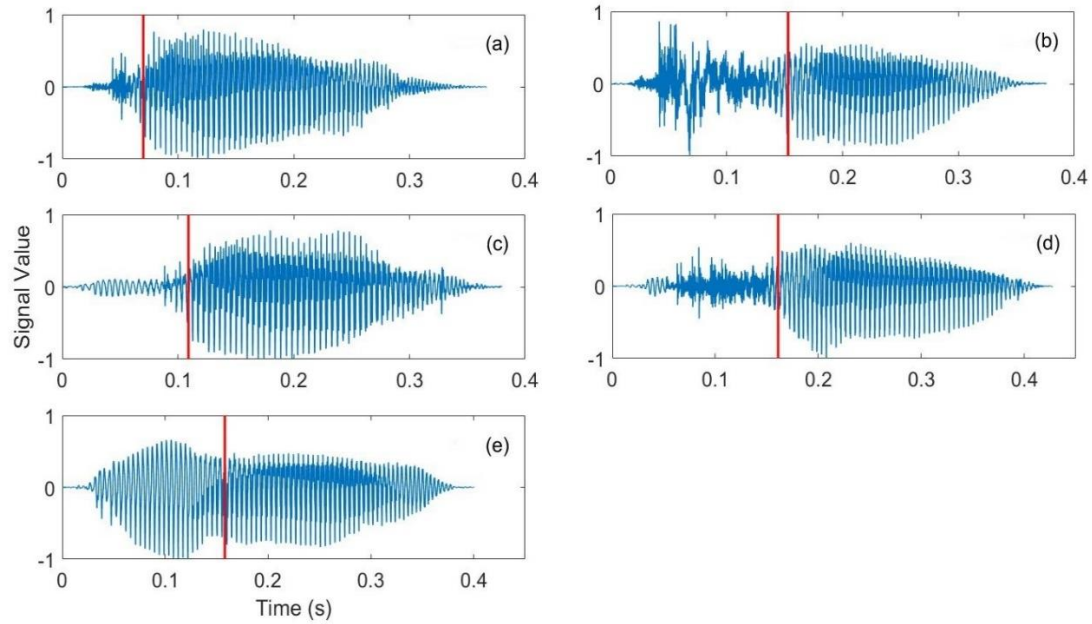


Fig. 3.10 Recorded Velar Consonant Phonemes: (a) ക് /k/, (b) ഖ്/k^h/, (c) ഗ് /g/, (d) ഘ്/g^h/ and (e) ങ് /ŋ/.

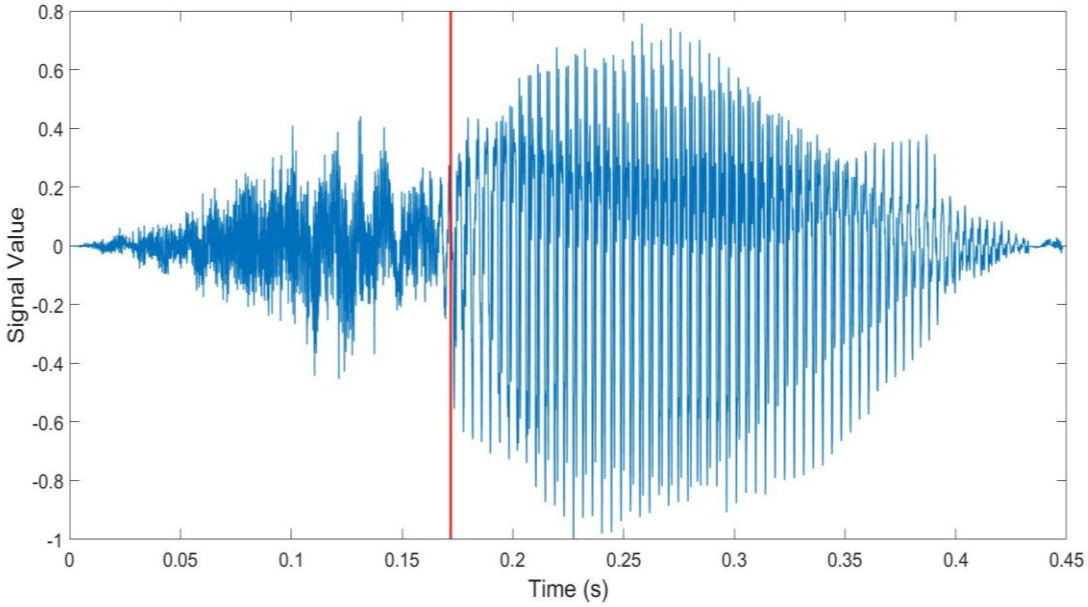


Fig. 3.11 Recorded Glottal Consonant Phoneme: ಹ್ /h/.

3.5 Audio Segmentation and Labelling

After the recording, the most time-consuming task was processing each piece of audio speech data individually and converting it to a standard format for later use. The data processing technique comprises segmenting and labelling audio files for each isolated phoneme and word, which eliminates repetition. The audio signals are segmented and labelled automatically in this study, and the segmented files are carefully examined for noise and inaccurate pronunciations. The separate audio files have the same sound repeated numerous times with sufficient space between them. A spectral subtraction procedure is applied to this audio stream, which removes the noisy background information that occurs while recording. Only audio files captured in a lab environment, where noise fluctuates slowly compared to the spoken signal, are subjected to the spectral subtraction technique. The audio files, which were recorded in an acoustically realistic environment, are kept for future research on the influence of a real-time noisy voice signal.

In the time domain, clean speech and noise are uncorrelated and additive. Furthermore, when it comes to the voice signal, the majority of the additive noise and channel distortion change extremely slowly. Convoluted channel distortion (d(m)) of clean speech (x(m)) and additive noise (n(m)) are commonly used to create a recorded speech signal(y(m)). As a result, the spectral subtraction approach can be used to remove additive noise from a signal.

$$y(m) = x(m) * d(m) + n(m) \tag{3.1}$$

Where 'm' denotes discrete-time index and '*' denotes the convolution operator. The power spectrum of a noisy speech signal is equal to the sum of the clean speech power spectrum X(k) and the noise power spectrum N(k), as shown in Eq. (3.2), where k is the frequency bin index.

$$| Y(k) |^2 = | X(k) |^2 + | N(k) |^2 \tag{3.2}$$

The noise power spectrum is calculated in the recorded signal's silent region. To produce the clean speech power spectrum, the estimated noise power spectrum is subtracted from the power spectrum of the noisy speech signal, as shown in Eq. (3.3).

$$| X(k) |^2 = | Y(k) |^2 - | N(k) |^2 \tag{3.3}$$

By integrating the magnitude of the obtained clear speech power spectrum with the phase information gained from the noisy speech input, the inverse discrete Fourier transform (IDFT) translates it into the time domain.

$$x(m) = \sum_{k=0}^{k=M-1} | X(k)| \mathrm{e}^{\frac{-\mathrm{j}2\pi\mathrm{k}}{\mathrm{M}}} e^{j\theta_Y(k)} \tag{3.4}$$

$\theta_Y(k)$ represents the phase information from the noisy speech signal. Due to the unpredictable nature of noise, spectral subtraction may yield negative results. The recovered signal will be affected by the negative value, which

will be noticeable at low signal-to-noise ratios because the power spectrum is always positive[120]. This problem is imperceptible in the presence of a weak background loud voice signal. Figure 3.12 depicts the spectrum subtraction method's block diagram.

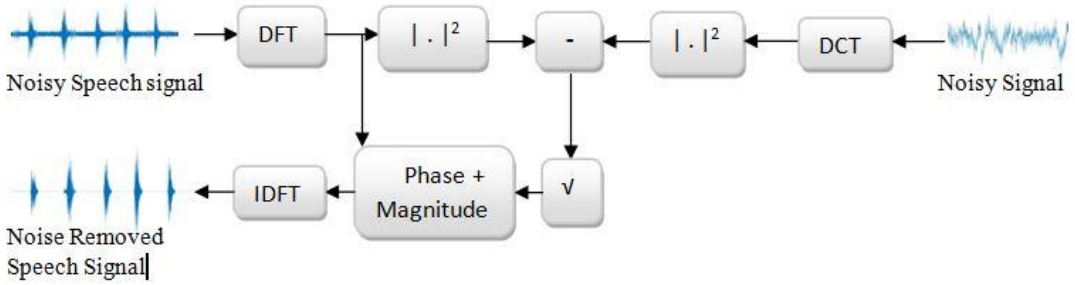


Fig. 3.12 Block Diagram of Spectral Subtraction method.

Following the noise removal stage, the prolonged voice signal (which contains several utterances of single phonemes) is segmented. The signal is divided into 10 ms frames throughout the segmentation procedure. The absolute values of the samples in each frame are compared to the absolute values of the entire signal's maximum. The first frame with a maximum absolute value of the samples in the frame larger than a threshold value (5 percent of the whole signal's maximum absolute value) is found. The second frame, which comes before this one, is called the isolated phoneme's first frame ($frame_{start}$). Then, following $frame_{start}$, the first frame with a maximum amplitude value less than a threshold value (about 5% of the whole signal's maximum absolute values) is detected. Following this frame, the second frame is considered the isolated phoneme's final frame ($frame_{end}$). As the first utterance, all frames between $frame_{start}$ and $frame_{end}$are preserved. The visual counterpart is segmented using the time information of the starting and ending frames. The original signal is stripped of all frames up to $frame_{end}$. This step is repeated until all of the isolated phonemes have been segmented. The steps involved in segmenting acoustic speech sounds are shown in Figure 3.13.

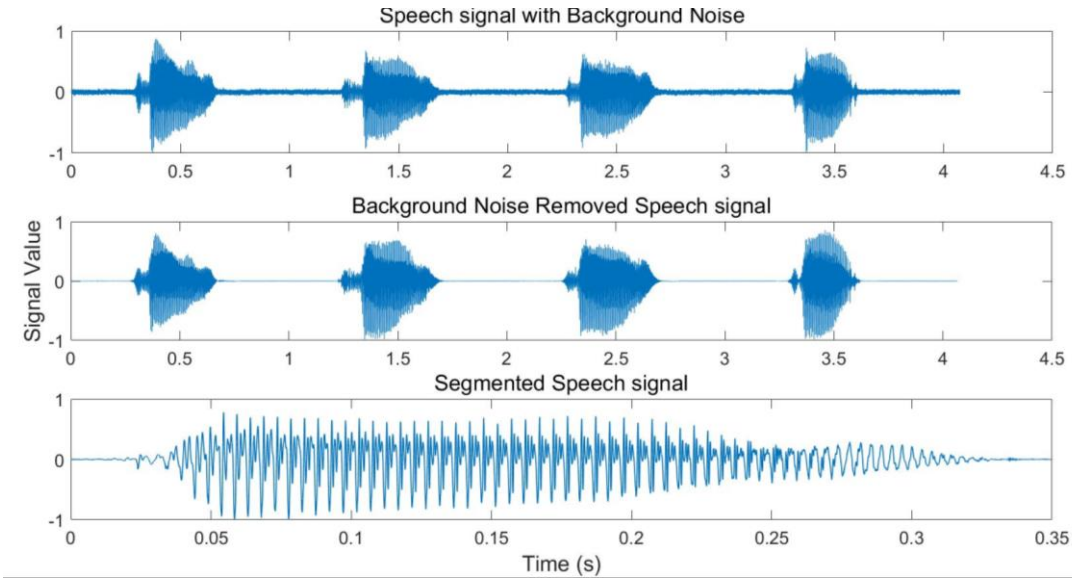


Fig. 3.13 Steps in speech Segmentation Process.

Each segmented speech signal from the same phrase is inspected acoustically and visually after this process, and misspelt and damaged speech signals are manually eliminated. After all segmented speech signals have been scrutinised, they are transferred to the labelling process, which contains all significant information about the signal. In all categories of recording, the labelling procedure was completed fully automatically, reducing the number of errors that occur during manual labelling.

It was chosen to keep each audio file in its own folder under the header of the respective category, obviating the need for category information to be included in the labelling process. The label of the isolated phonemes is represented by the first four characters (the maximum number of characters necessary for labelling is for ൎ/r/ (TTAA)). The extra characters are filled with X (like AXXX for അ /a/) for phonemes with labels that are fewer than four characters long. Allophones are given an extra character (a single digit) to represent the context-based variation of phonemes. The phoneme ു/u/ has the most context-based phonemic variety (7 classes as in table 3.4). Because

some speakers repeated the utterance more than 10 times, the next two characters reflect the repetition of the utterance (from 01 to 99). The speaker's gender ('F' for females and 'M' for males) is represented by the next character (7th for phoneme and 8th for allophone). The age of the speaker is represented by the next two letters. The speaker's nativity is represented by the next character. The initial letter of the city/street represents the speaker's nativity. Because recording requires a large number of speakers, the last two characters reflect the speaker number (each speaker is assigned a unique three-digit number). The phoneme and allophone information are provided by the first six and seven characters, respectively. The speaker's identity is represented by the next seven characters. So, for phonemes and allophones, an audio speech file is titled with 13 and 14 alphanumeric characters, respectively.

Table 3.5 Nomenclature rule of audio Database files

	Sound	**Repetition**	**Gender**	**Age**	**Place**	**Speaker Number**	**Total Length**	**Format**
Phoneme	AXXX	19	M	09	V	017	13	.wav
Allophone	UXXX7	03	F	23	C	009	14	.wav

3.6 Conclusion

The creation and presentation of a new Malayalam audio speech database has been addressed in this chapter. This collection contains 50 isolated Malayalam phonemes and 207 related words that comprise all allophonic variations. The database is created in two modes: closed and open. A clean audio speech database is created with the help of 200 speakers (100 male and 100 female) in a closed environment and noisy speech is developed with the help of 20 speakers (10 male and 10 female) in an open environment. Five to ten years old, twenty to twenty-five years old, and sixty to sixty-five years old are the age categories represented by the speakers. Each speaker was asked to repeat each recording ten times. The database was segmented and

labelled using the spectral subtraction method. The aforementioned database can be utilised to enhance research in a variety of speech-based signal studies. The database is used here to investigate the nonlinearity of the reconstructed hyperspace of the speech production system. For this purpose, the time delay and embedding dimension of the attractor should be determined. The developed database is utilised to optimise time delay and embedding dimension in the following chapter.

CHAPTER 4

OPTIMISATION OF EMBEDDING DIMENSION FOR PHASE SPACE RECONSTRUCTION

4.1 Introduction

To characterize a physiological system's temporal evolution ,the nature of dynamical variables (minimum dimension of the dynamical system) involved in developing the system with time is the most vital element [121]. The methodology most widely used in general non-linear dynamical systems is the method of False Nearest Neighbours (FNN) and Principal Component Analysis (PCA) [122]. The determination of the minimum embedding dimension is highly crucial for constructing the phase space for describing the system [123].

The dynamical factors compose the components of a state vector in phase space. The representation of every dynamical variable as a function of time is the most natural way to characterise any dynamical system. Another effective way to describe the system would be to replace the time independent variable with another dynamical variable of the system. In this case each point in the state space represents a system state in a given instant [121]. Even though three dynamical coordinates would be the highest visual capacity, a hypothetical abstract space with more than three coordinates can represent the dynamical system under study completely. Each representative point resembles the system under investigation and, the swarm of points represent the time evolution of the said system [124]. The attractor geometry offers information about the structural complexity sustaining that dynamics and nature of the underlying physical system.

Several researchers have carried out various studies by reconstructing the phase space in multiple dimensions in line with the time series [125]–[127]. The nature of time series is essential in the embedding and the embedding dimension should be appropriate to acquire the relevant information from time series. The FNN and PCA methods are beneficial in determining the embedding dimension. Two and three- dimensional modelling of the speech production system based on attractor reconstruction using Malayalam vowel sounds has been carried out lately [128]. Still, it gives very less information about the dynamical nature of the system. In this scenario, we carried out dimensionality analysis on Malayalam vowel time series using FNN and PCA. The optimization of dimension may give better understanding about the underlying dynamics of the system.

The following is an overview of the chapter's structure. The approach to phase space reconstruction is described in Section 4.2. The database used for standardisation and analysis is detailed in section 4.3. The acquired results are presented, evaluated, and interpreted in section 4.4 in graphical and tabular form. The chapter is concluded in section 5.5.

4.2 Phase Space Reconstruction

The phase space of a dynamical system can be used to describe the system's time evolution. If the system is deterministic, all future states are determined by the current fixed state. The corresponding phase space points can be used to investigate the system's dynamics. The uniqueness of trajectories is ensured by a set of first order differential equations in phase space, which define dynamical systems. Even non-deterministic systems can be characterised by a set of states and transition rules.

4.2.1 Description of a Dynamical System

In a finite dimensional vector space $\mathbf{R^m}$, a state is specified by a vector x ϵ $\mathbf{R^m}$**.** The dynamics of the system can be described in either of two ways

1. An explicit system of 'm' first order ordinary differential equations

$$\frac{d\,x(t)}{dt} = f\big(x(t)\big),\ \mathrm{t}\ \epsilon\ \mathrm{R} \tag{4.1}$$

2. By a 'm ' dimensional map in phase space

$$X_{n+1}=F(x_n),\ \ n \epsilon Z \tag{4.2}$$

4.2.2 Embedding of a Time Series

In an experiment, we see a time series, most likely only a set of scalar measurements, rather than a phase space object. In the case of a speech production system, the voice signal is the outcome of an experiment that may be sampled to generate a time series. Converting observations into state vectors is a significant problem known as phase space reconstruction. Technically, the method of delays can be used to overcome this problem.

A reconstructed phase space (RPS) can be produced for a measured state variable x_n, n=1, 2, 3, 4 ….N, via the method of delays [129] by creating vectors given by

$$X_n = (x_n, x_{n+\tau}, x_{n+2\tau}, \dots\dots\dots\dots\dots, x_{n+(m-1)\tau}) \tag{4.3}$$

The time difference of number of samples (τ) is the lag or delay time. 'm' stands for embedding dimension. The 'm' dimensional phase space can be obtained from the above delay vectors as

$$X = \begin{bmatrix} x_1 & x_{1+\tau} & x_{1+2\tau} & \cdots & \cdots & x_{1+(m-1)\tau} \\ x_2 & x_{2+\tau} & x_{2+2\tau} & \cdots & \cdots & x_{2+(m-1)\tau} \\ x_3 & x_{3+\tau} & x_{3+2\tau} & \cdots & \cdots & x_{3+(m-1)\tau} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{N-1} & x_{(N-1)+\tau} & x_{(N-1)+2\tau} & \cdots & \cdots & x_{(N-1)+(m-1)\tau} \\ x_N & x_{N+\tau} & x_{N+2\tau} & \cdots & \cdots & x_{N+(m-1)\tau} \end{bmatrix} \quad (4.4)$$

As per Taken's theorem, "Suppose the d-dimensional state vector x(t) evolves according to an unknown but continuous and (crucially) deterministic dynamic. Suppose that the one-dimensional observable y(t) is a smooth function of x(t), and coupled to all the components of x(t). Now at any time we can look not just at the present measurement y(t) , but also at observations made at times removed from us by multiples of some lag τ : $y_{t\tau-}, y_{t-2\tau}$, etc. If we use k lags, we have a k -dimensional vector. One might expect that, as the number of lags is increased, the motion in the lagged space will become more and more predictable, and perhaps in the limit k→∞ would become deterministic. In fact, the dynamics of the lagged vectors become deterministic at a finite dimension; not only that, but the deterministic dynamics are completely equivalent to those of the original state space (More exactly, they are related by a smooth, invertible change of coordinates, or diffeomorphism). The magic embedding dimension is at most 2d+1 , and often less" [126], the phase space should be reconstructed with the correct dimension in order to extract useful information about the system.

The embedding theorem does not consider the time delay between subsequent entries in the delay vectors. It is arbitrary from a statistical point of view. Autocorrelation method and mutual information method are the two popular methods for determining time delay. Here Mutual information

method is used as it gives better predictions as compared to autocorrelation. The false nearest neighbour (FNN) method and principle component analysis (PCA) are the most popular statistical method found in literature. Hence, they are used for optimising embedding dimension.

4.2.3 Time delay by Mutual Information(MI) method

The mutual information, ***I (Z,Y)*** ,which is a measure of the minimum uncertainity in time series ' *z*' (when measurement of series ' *y* ' is given) is

$$I(Z,Y) = \sum_{i,j} p_{yz}\,(y_i, z_i) \log \left[\frac{p_{yz}(y_i,z_i)}{p_y(y_i,z_i)p_z(z_j)}\right] \tag{4.5}$$

Where p_{yz} is thejoint probability mass function of Y and Z. p_y and p_z are the marginal probability of Y and Z respectively.

For a time series x(t), obtained from speech signal, the dependency of the values of x(t+τ) on the values of *x(t)* can be determined by making the assignment [y,z]= [x(t), x(t+τ)]. For speech time series x(t), the average mutual information between x(t) and x(t+τ) can be stated as [125]

$$I(\tau) = \sum_{x(t),x(t+\tau)} p(x(t), x(t+\tau) \log\,[p(x(t), x(t+\tau)/p(x(t)p(x(t+\tau)] \tag{4.6}$$

4.2.4 Embedding Dimension by FNN

Kennel, Brown, and Abarbanel pioneered the notion of FNN (1992). Hegger and Kanz [130], proposed a modified approach, which is applied in this study. The main idea is to look for points in the data set that are neighbours in embedding space, but their future temporal evolution is too different. FNN calculates the percentage of false neighbours based on the distances between samples rebuilt in m-dimensional spaces. The fraction of FNN is given by

$$X_{fnn}(r) = \frac{\sum_{n=1}^{N-m-1}\left\{H\left(\frac{\left|s_n^{m+1}-s_{k(n)}^{m+1}\right|}{\left|s_n^{m}-s_{k(n)}^{m}\right|}-r\right)H(\frac{\sigma}{r}-\left|s_n^{m}-s_{k(n)}^{m}\right|)\right\}}{\sum_{n=1}^{N-m-1}\left\{H(\frac{\sigma}{r}-\left|s_n^{m}-s_{k(n)}^{m}\right|)\right\}} \quad (4.7)$$

Where $s_{k(n)}^{m}$ is the closest neighbour to s_n in 'm' dimensions. H corresponds to step function, 'r' is the threshold distance and σ is the standard deviation. In case the fraction for any given *m* is low, we will obtain a better reconstruction in *m* dimensions. This will help to examine the way the dataset is reorganized and its own information reflected according to the number of dimensions [131]. When the fraction of false neighbour reaches zero, no substantial modification in the distance of the points could be identified. Thus, there is enough information to unfold and comprehend the behaviour of studied system.

4.2.5 Embedding Dimension by PCA

The PCA method recognizes the directions of the m-dimensional coordinate system that shows more significant variances in the data discarding those less essential dimensions. PCA enables a dimensional reduction which helps to simplify the dynamics from an initially very high dimensional space. We can reconstruct the time series to a high dimensional space by taking the delay as obtained from mutual information. From the prominent number of eigenvalues, the embedding dimension of time series can be predicted [121].

PCA is a dimensionality-reduction method commonly used to reduce the dimensionality of large data sets by reducing a vast collection of variables into a smaller set that preserves the majority of the information in the original set. The steps in PCA are

1. Standardization

All the variables will be transformed to the same scale by standardisation. The standardised value is given by

$$z = \frac{value-mean}{standard\ deviation} \quad (4.8)$$

2. Covariance matrix computation

The goal of this stage is to understand how the variables in the input data set differ from the mean in relation to each other, or to discover if there is any link between them. Because variables are sometimes so highly connected that they contain redundant information. So, to find these connections, we compute the covariance matrix. For an N dimensional data set with variables (x_1,x_2,x_3,.....,x_N) the covariance matrix is defined as

$$M = \begin{pmatrix} cov(x_1,x_2) & cov(x_1,x_3) & cov(x_1,x_4) & \cdots & cov(x_1,x_N) \\ cov(x_2,x_1) & cov(x_2,x_2) & cov(x_2,x_3) & \dots & cov(x_2,x_N) \\ cov(x_3,x_1) & cov(x_3,x_2) & cov(x_3,x_3) & \dots & cov(x_3,x_N) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov(x_N,x_1) & cov(x_N,x_2) & cov(x_N,x_3) & \cdots & cov(x_N,x_N) \end{pmatrix} \quad (4.9)$$

3. Computation of eigen values of the covariance matrix

Eigenvectors and eigenvalues are linear algebra concepts that must be computed from the covariance matrix in order to explore the data's principal components. The number of prominent eigen values of the system give the optimized embedding dimension for reconstructing the phase space of the system.

4.3. Database used

For testing the applicability of the methods used two standard nonlinear systems, Lorentz system and Rossler system, are used as model systems. The Short vowel part of the Malayalam speech data base developed (Chapter 3) is used for further investigation.

4.3.1 Model Systems

The Lorentz and Rossler systems, well known nonlinear systems, are used as model systems to compare the FNN and PCA method's effectiveness [124].The Lorenz and Rossler systems consists of a set of three nonlinear ordinary differential equations given by equations 4.10 to 4.12 and 4.13 to 4.15, respectively.

$$\frac{dx}{dt} = \sigma(y - x) \quad (4.10)$$

$$\frac{dy}{dt} = x(\rho - z) - y \quad (4.11)$$

$$\frac{dz}{dt} = xy - \beta z \quad (4.12)$$

$$\frac{dx}{dt} = -y - z \quad (4.13)$$

$$\frac{dy}{dt} = x + ay \quad (4.14)$$

$$\frac{dz}{dt} = b + z(x - c) \quad (4.15)$$

As parameters of initial conditions, we choose σ = 10, ρ = 28 and β = 2.67 for Lorenz system and a=b=0.2 and c=0.78 for Rossler system. 30000 data points sampled from the Lorenz system is used for analysis.

4.3.2. Malayalam Vowel Speech Database

In this work, short vowel phonemes of the Malayalam audio database created is used [119]. The time series corresponding to short vowel phonemes of speakers sampled at 16 kHz, 32 kHz and 44.1 kHz are used for study. The samples include different age groups. The samples used for the analysis is summarised in Table 4.1.

Table 4.1 Malayalam Database used

Age Group	Sampling Frequency	No. of Speakers		Vowel Phonemes IPA	Consonant Phonemes IPA
		Male	Female		
5-10	16 kHz 32 kHz 44.1 kHz	20	20	അ /a/ ഇ /i/ ഏ /e/ ഒ /o/ ഉ /u/	പ/P/ വ/v/ ത/t/ റ/r̲/ ട/ʈ/ ച /c/ ക /k/ ഹ /h/
20-25	16 kHz 32 kHz 44.1 kHz	50	50		
60-65	16 kHz 32 kHz 44.1 kHz	30	30		

4.4. Experiments and Result Analysis

For standardisation, time series generated from Lorenz and Rossler systems were employed. The research is based on Malayalam vowel time series uttered by speakers of various ages and sampled at various frequencies. The Mutual Information (MI) approach is used to determine the embedding delay of each time series. The embedding dimension is optimised using two procedures: FNN and PCA. Both FNN and PCA use the delay value produced from the MI function as an input.

4.4.1 Results of Model Systems

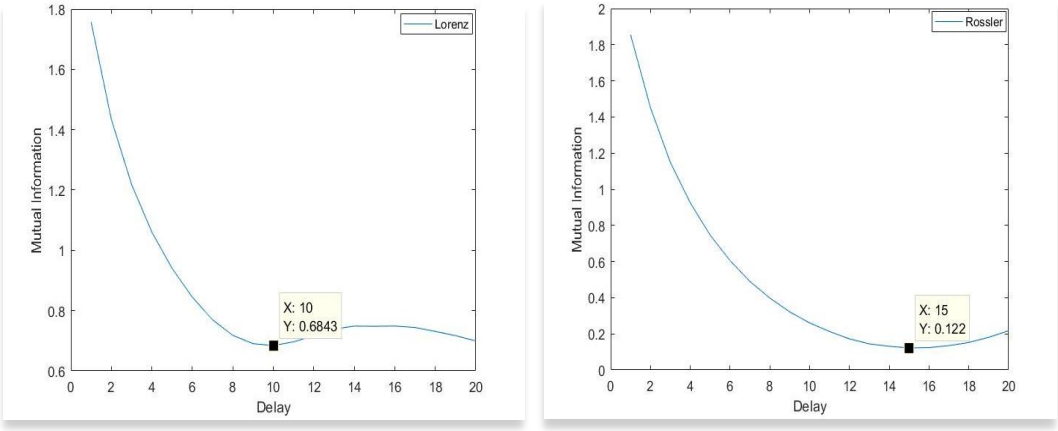


Fig. 4.1 The Variation of Mutual Information with Delay of Lorenz and Rossler systems

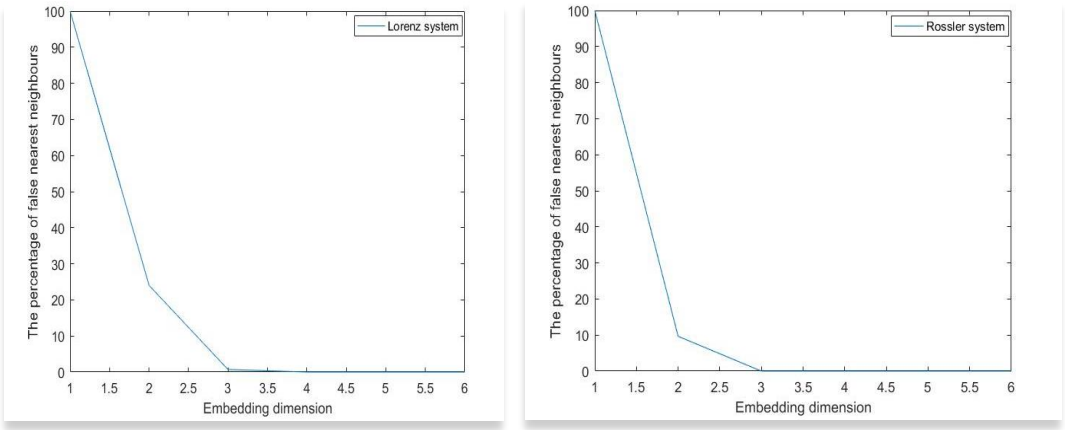


Fig. 4.2 The variation of FNN with embedding dimension of Lorenz and Rossler systems

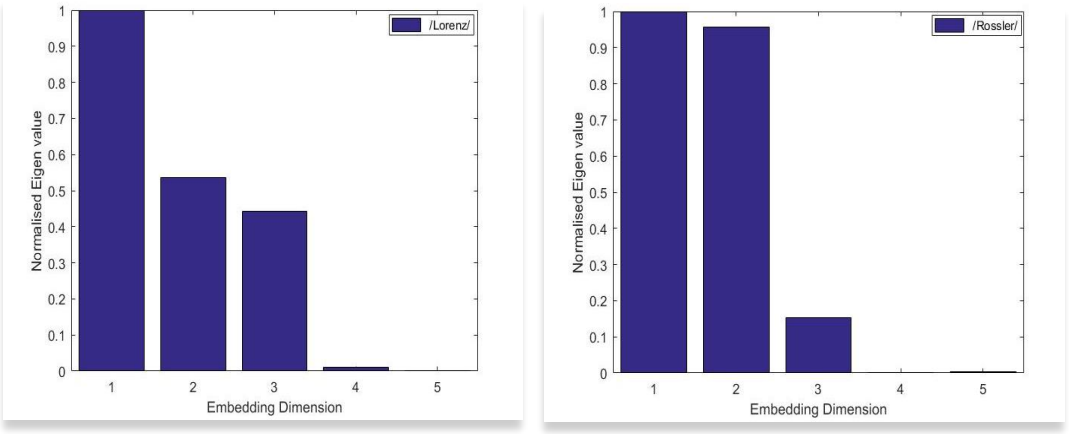


Fig. 4.3 Normalised eigen values of Lorenz and Rossler systems

Fig. 4.1 shows the variation of mutual information with time delay for Lorenz and Rossler systems. The Lorenz and Rossler systems are three-dimensional systems, and FNN drops to zero at embedding dimension three, as shown in Fig 4.2. In the Lorenz and Rossler systems, there are three prominent eigenvalues, as shown in Fig 4. 3.

4.4.2 Time delay for Malayalam Speech Database

(A). Results of MI

The MI technique is used to determine the time delay for all speech samples as given in Table 1. Fig 4.4 to Fig 4.9 show the variation of MI with time delay of short vowel phonemes uttered by speakers of different age and gender sampled at different frequencies. It was found that delay varies depending on the speaker, the speech, and the sampling frequency.

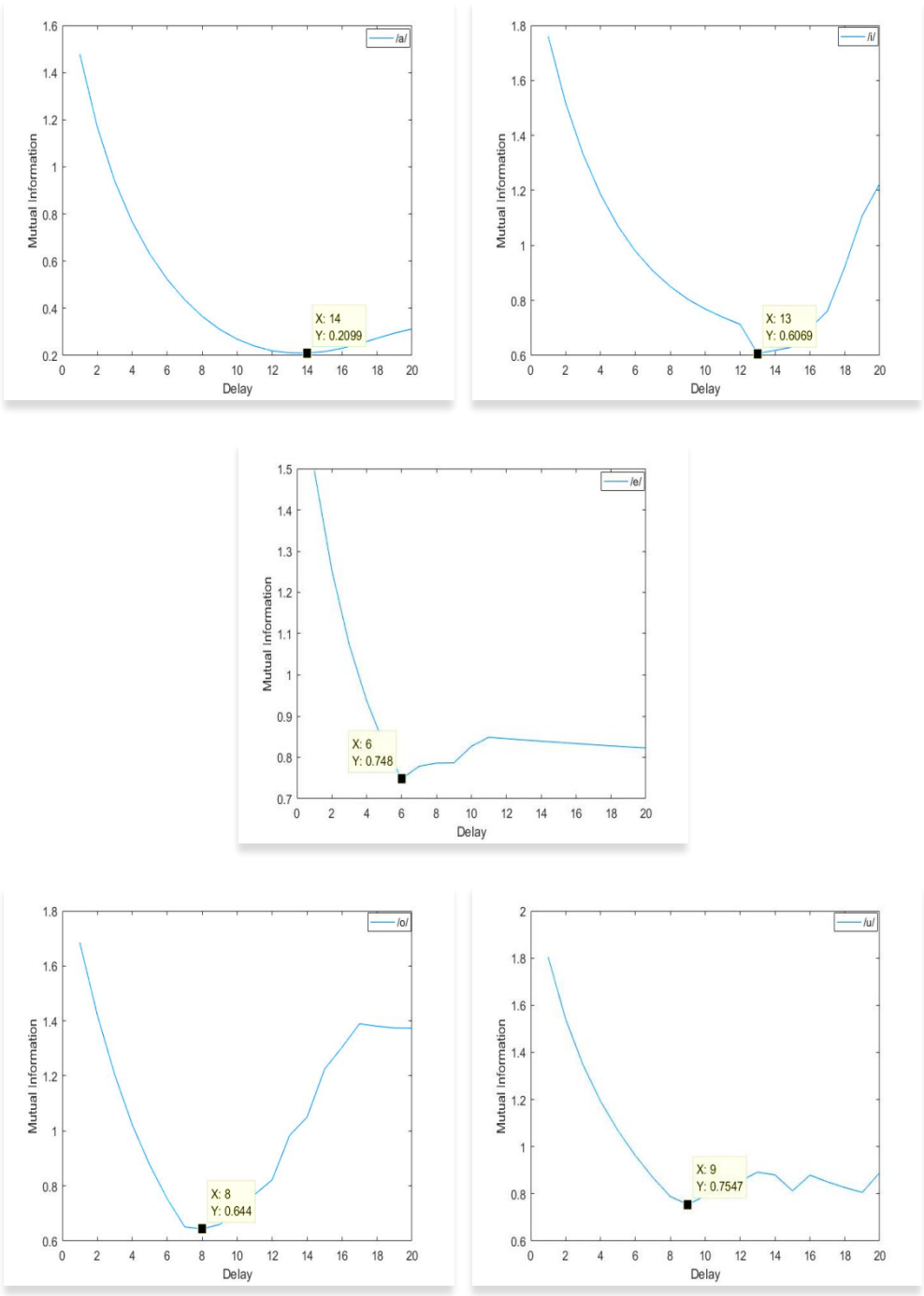


Fig. 4.4 The variation of MI with Embedding dimension for female speaker of age 5-10 sampled at frequency 16 kHz for Malayalam vowel (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/.

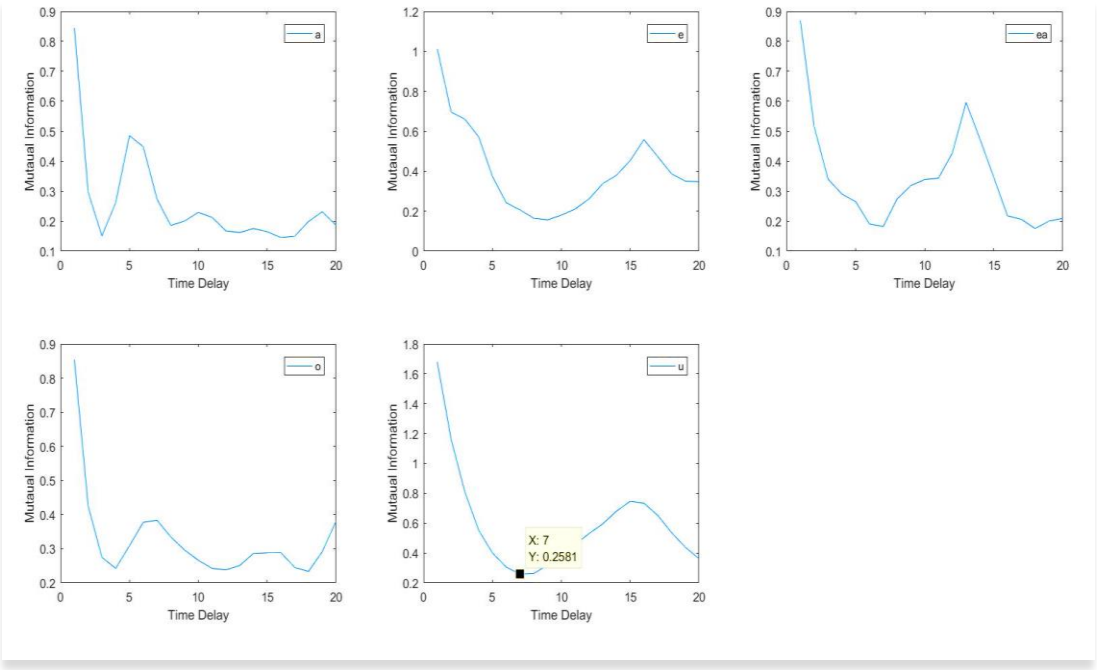


Fig. 4.5 The variation of MI with Embedding dimension for female speaker of age 20-25 sampled at frequency 16 kHz for Malayalam vowel (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/.

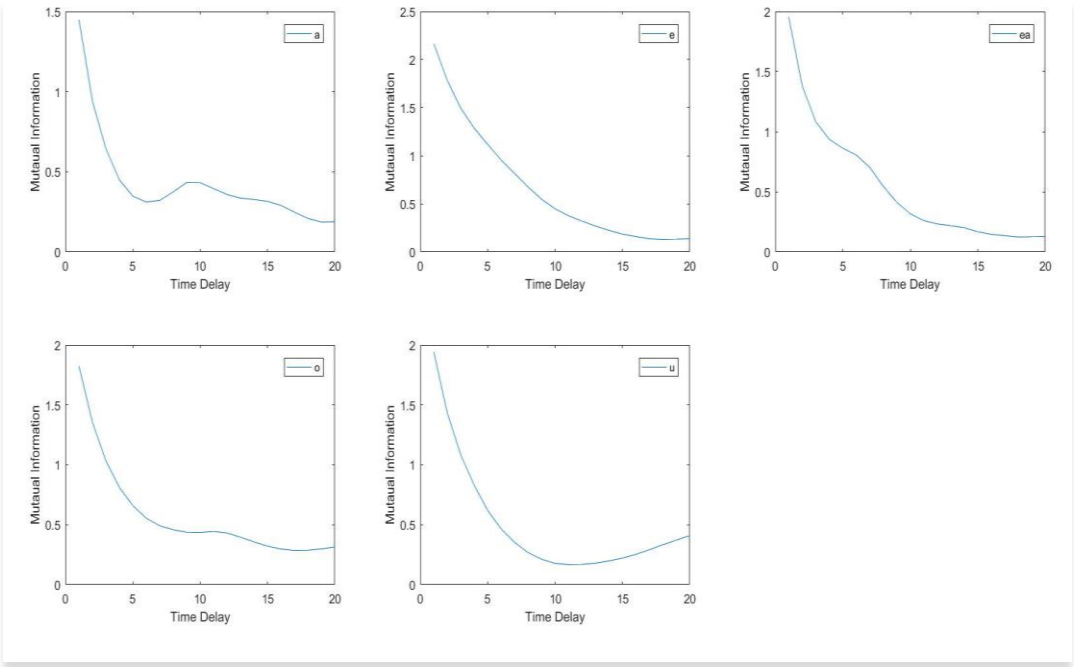


Fig. 4.6 The variation of MI with Embedding dimension for female speaker of age 60-65 sampled at frequency 16 kHz for Malayalam vowel (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/.

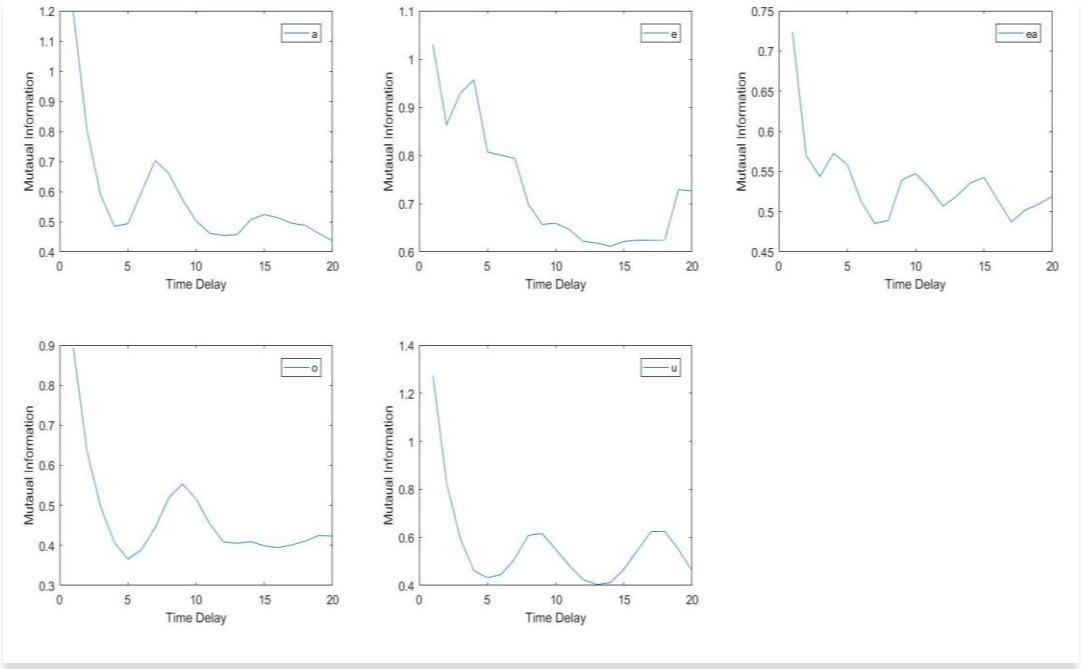


Fig. 4.7 The variation of MI with Embedding dimension for male speaker of age 20-25 sampled at frequency 16 kHz for Malayalam vowel (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/.

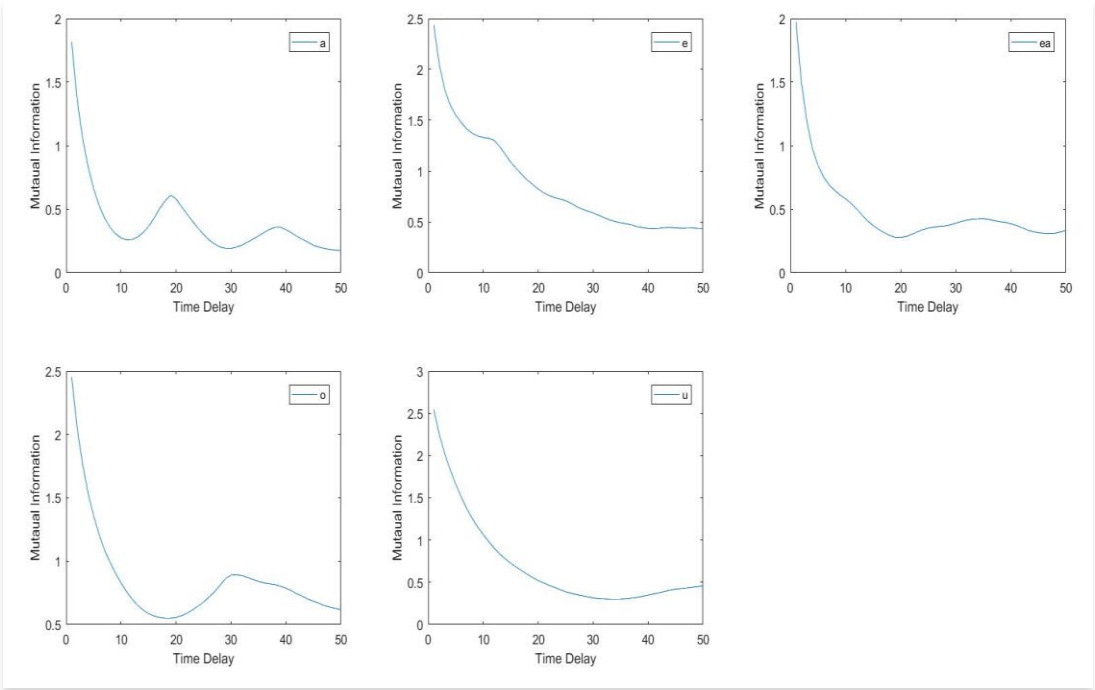


Fig. 4.8 The variation of MI with Embedding dimension for female speaker of age 20-25 sampled at frequency 32 kHz for Malayalam (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/.

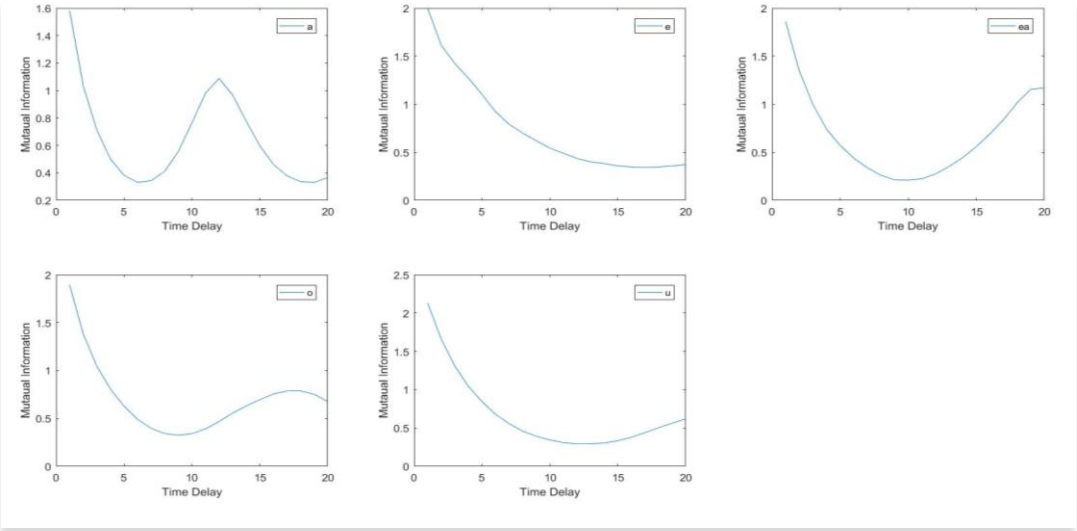


Fig. 4.9 The variation of MI with Embedding dimension for female speaker of age 20-25 sampled at frequency 44.1 kHz for Malayalam vowel (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/.

(B) Analysis of results

Figures 4.10 (male sound) and 4.11 (female sound) show the statistical distribution functions for the samples studied. The mean (μ) and standard deviation (σ) was found to be high for both male and female sounds. Table 4.2 shows the mean and standard deviation of time delay from the average value for various sampling frequencies and age groups. Because of the large standard deviation, it is impossible to generalise the delay. As a result, the time delay for each sample should be determined at the time of analysis.

Table 4.2 Mean (μ) and Standard Deviation (σ) of Time Delay for Malayalam database

Age Group	Sampling Frequency	μ (σ) (Male)	μ (σ) (Female)
5-10	16kHz	4 (2.32)	4 (2.44)
	32kHz	5 (2.55)	5 (2.65)
	44.1kHz	8 (2.92)	7 (3.12)
20-25	16kHz	5 (2.51)	6 (2.48)
	32kHz	6 (2.63)	7 (2.71)
	44.1kHz	9 (3.01)	9 (2.99)
60-65	16kHz	6 (2.67)	5 (2.61)
	32kHz	8 (3.02)	8 (2.99)
	44.1kHz	10 (3.46)	10 (3.33)

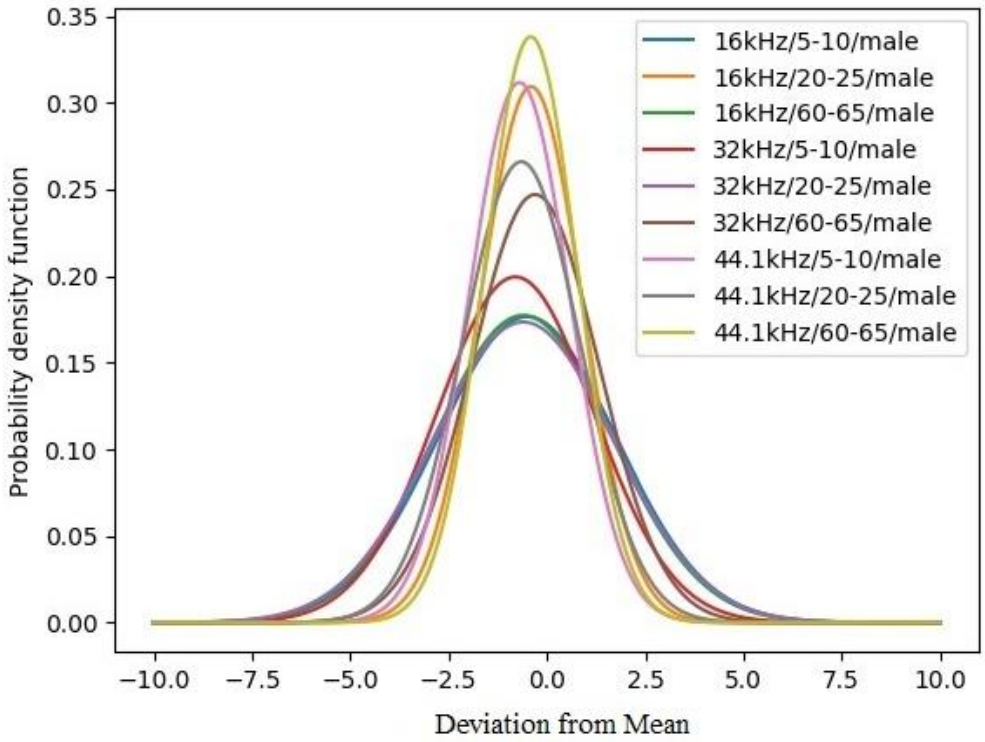


Fig.4.10 Probability distribution of Time delay (male sound)

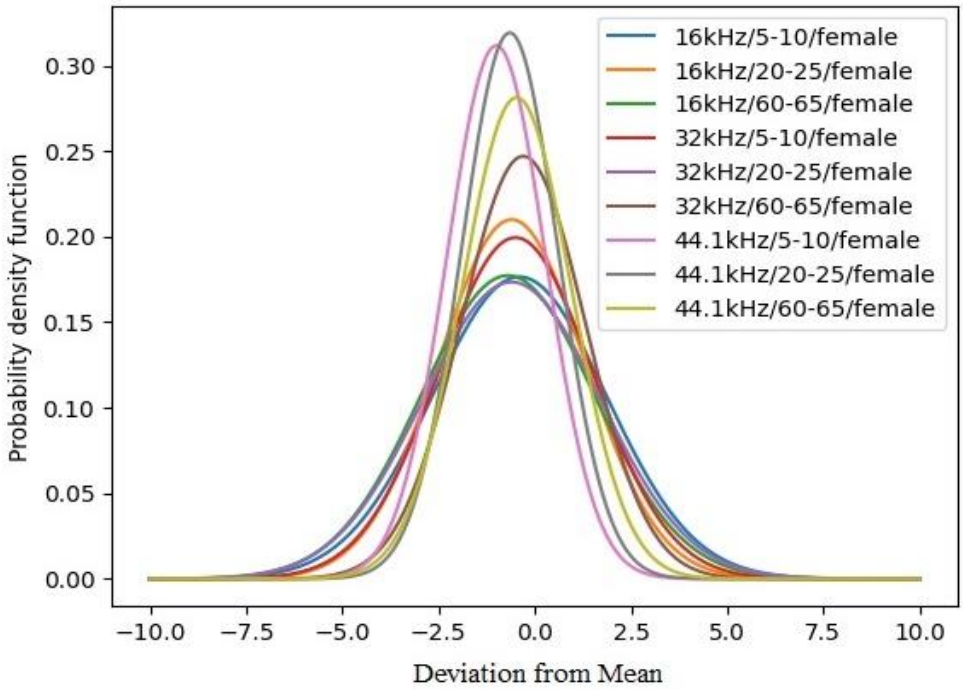


Fig.4.11 Probability distribution of Time delay (female sound)

Time delay is found to be dependent on utterance, speaker, age, gender, and sampling frequency. The delay rises with sampling frequency and age, as seen by the probability distribution. In none of the samples, there is a distinct variance with the varied utterances. As a result, standardising the embedding delay is extremely challenging. When analysing the samples, the sampling frequency must be carefully chosen.

4.4.3 Embedding Dimension for Malayalam Speech Database

(A) Results of FNN

FNN analysed speech samples with various gender, age groups, utterances, and sampling frequencies which can be seen in Table 4.1 are examined. Separately and collectively, vowel and consonant phonemes are studied. Figures 4.12 to 4.14 demonstrate the variation of FNN with Embedding dimension for Malayalam vowels vowel (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/ said by female speakers of age groups 5-10, 20-25, and 60-65 sampled at frequency 16 kHz. The same is shown in Figure 4.15 for male speakers between the ages of 20 and 25. The corresponding analysis of sampling frequencies 32 kHz and 44.1 kHz are represented in Figures 4.16 and 4.17, respectively. Figure 4.18 shows the variation of FNN for two female speakers uttering eight consonants at 16 kHz: Bilabial പ/P/, labiodental വ/v/, dental ഹ /h/, alveolar ത/t/, retroflex ർ/r̲/, palatal ട/ṭ/, velar ച /c/, and glottal ക /k/. For almost all samples, FNN false to zero at an embedding dimension of six. The mean value for all of the samples examined is close to six, with a mode value of six. The detailed statisical analysis is given in section 4.4.3.(C).

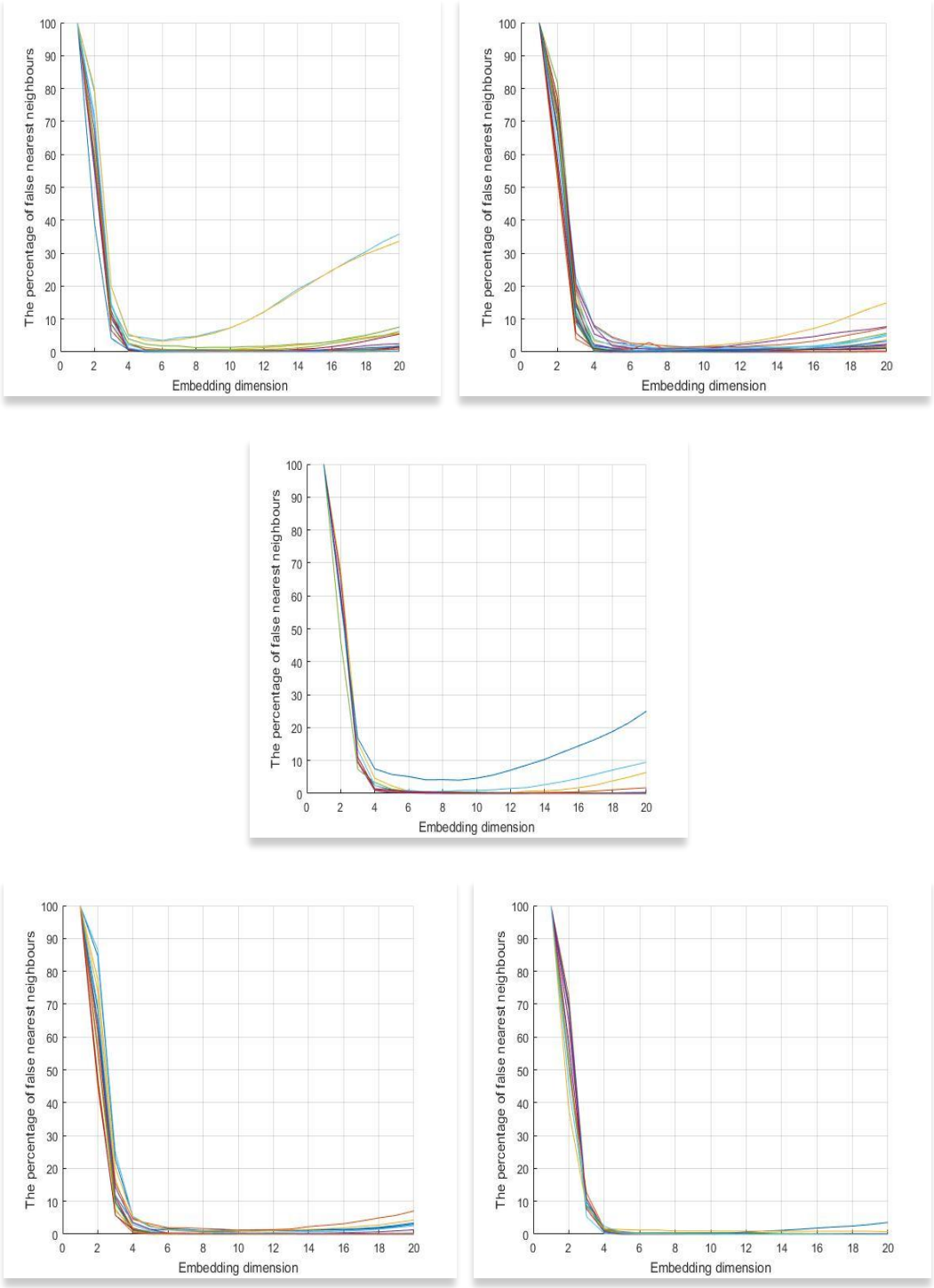


Fig. 4.12 The variation of FNN with Embedding dimension for 20 different female speakers of age 5-10 sampled at frequency 16 kHz for Malayalam vowel (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/.

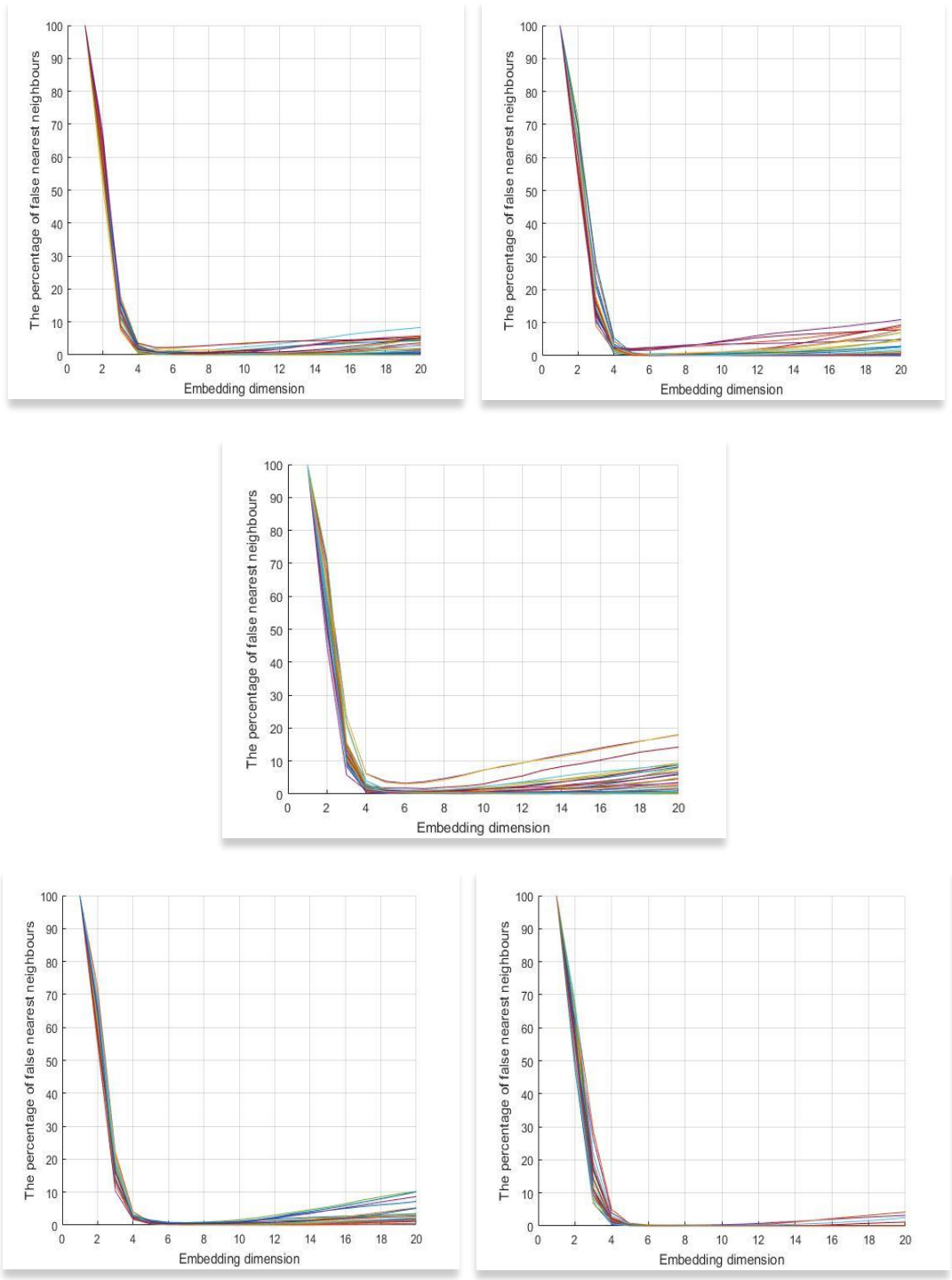


Fig. 4.13 The variation of FNN with Embedding dimension for 20 different female speakers of age 20-25 sampled at frequency 16 kHz for Malayalam vowel (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/.

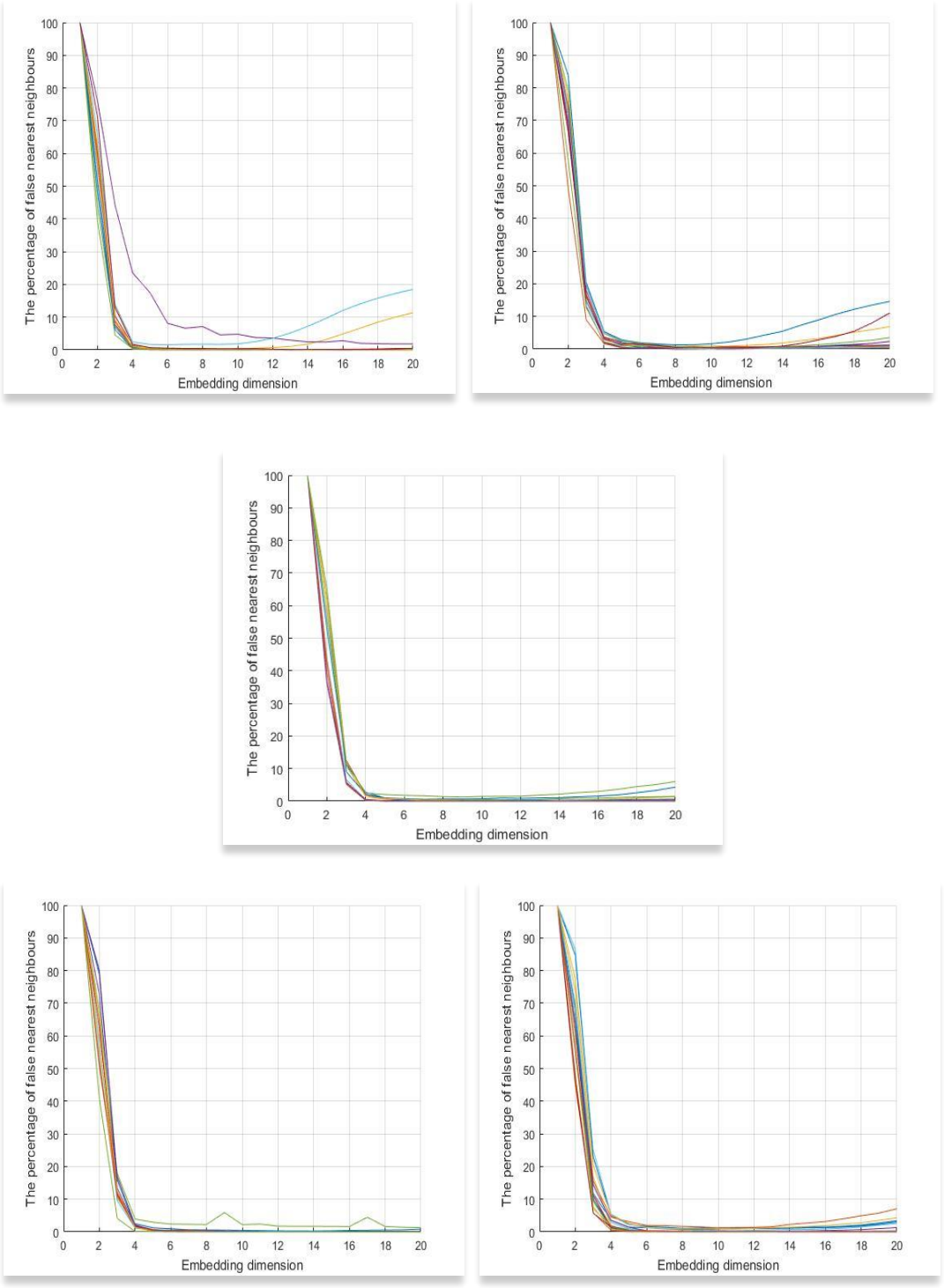


Fig. 4.14 The variation of FNN with Embedding dimension for 20 different female speakers of age 60-65 sampled at frequency 16 kHz for Malayalam vowel (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/.

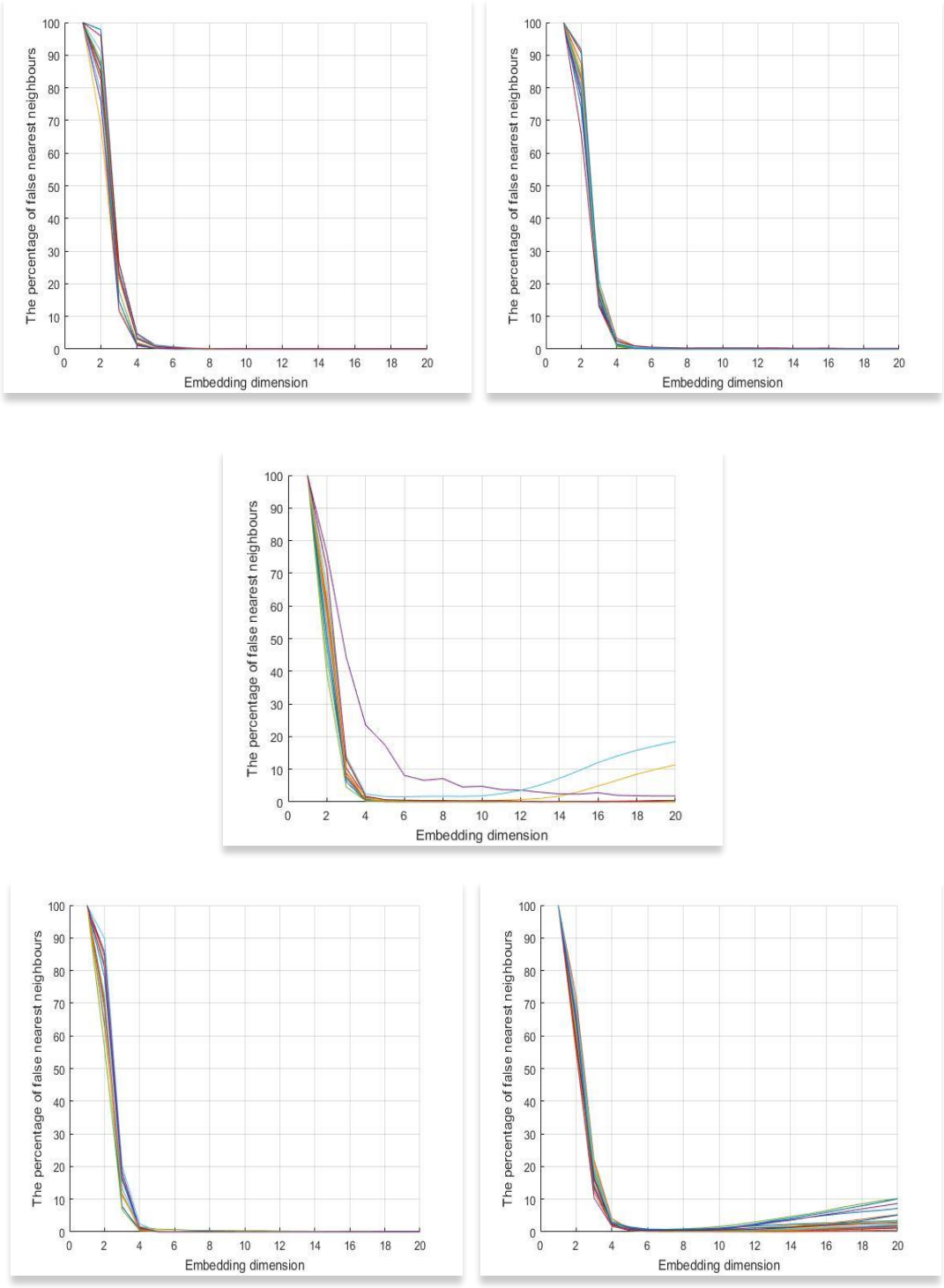


Fig. 4.15 The variation of FNN with Embedding dimension for 20 different male speakers of age 20-25 sampled at frequency 44.1 kHz for Malayalam vowel (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/.

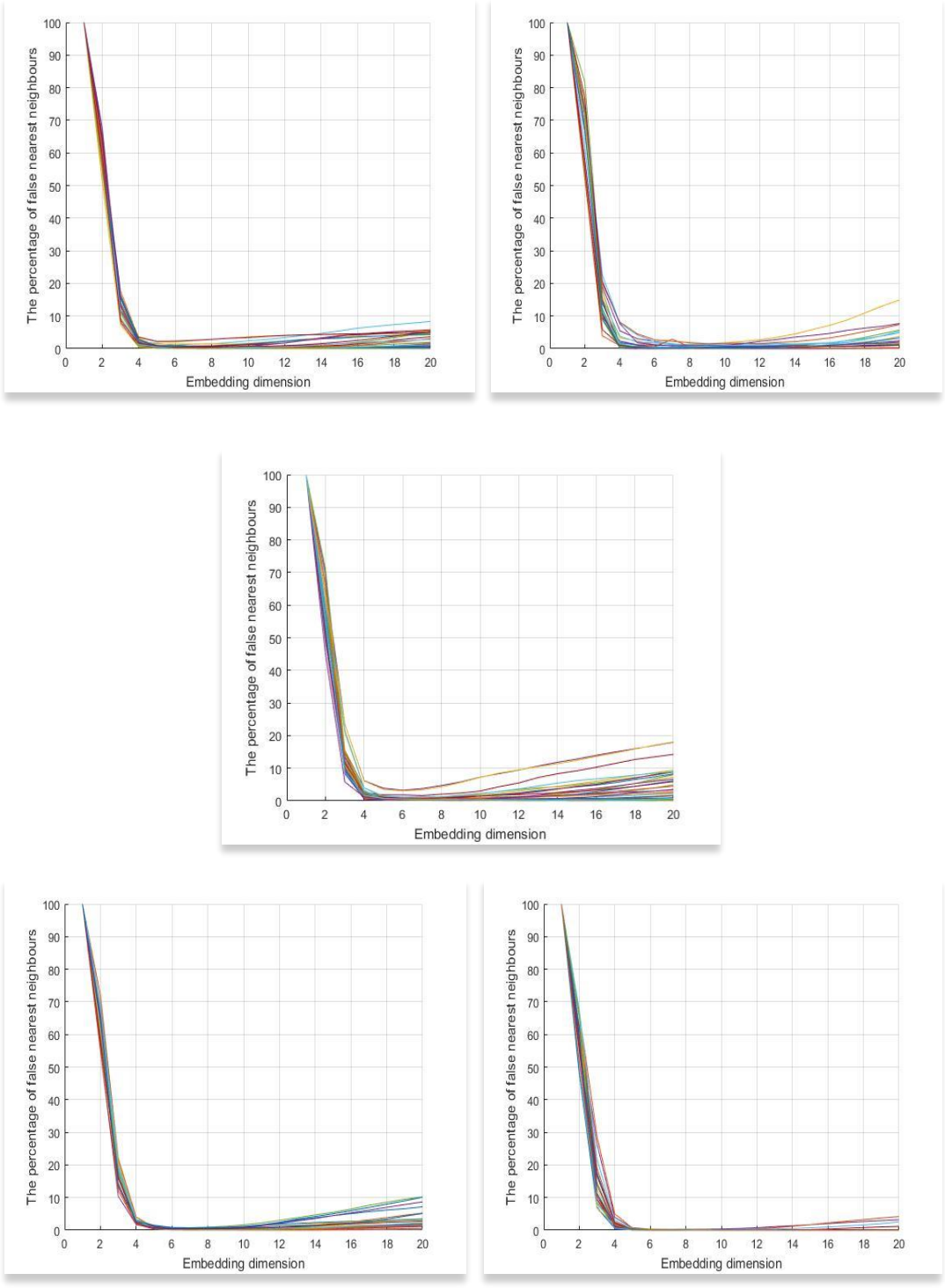


Fig. 4.16 The variation of FNN with Embedding dimension for 20 different female speakers of age 20-25 sampled at frequency 32 kHz for Malayalam vowel (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/.

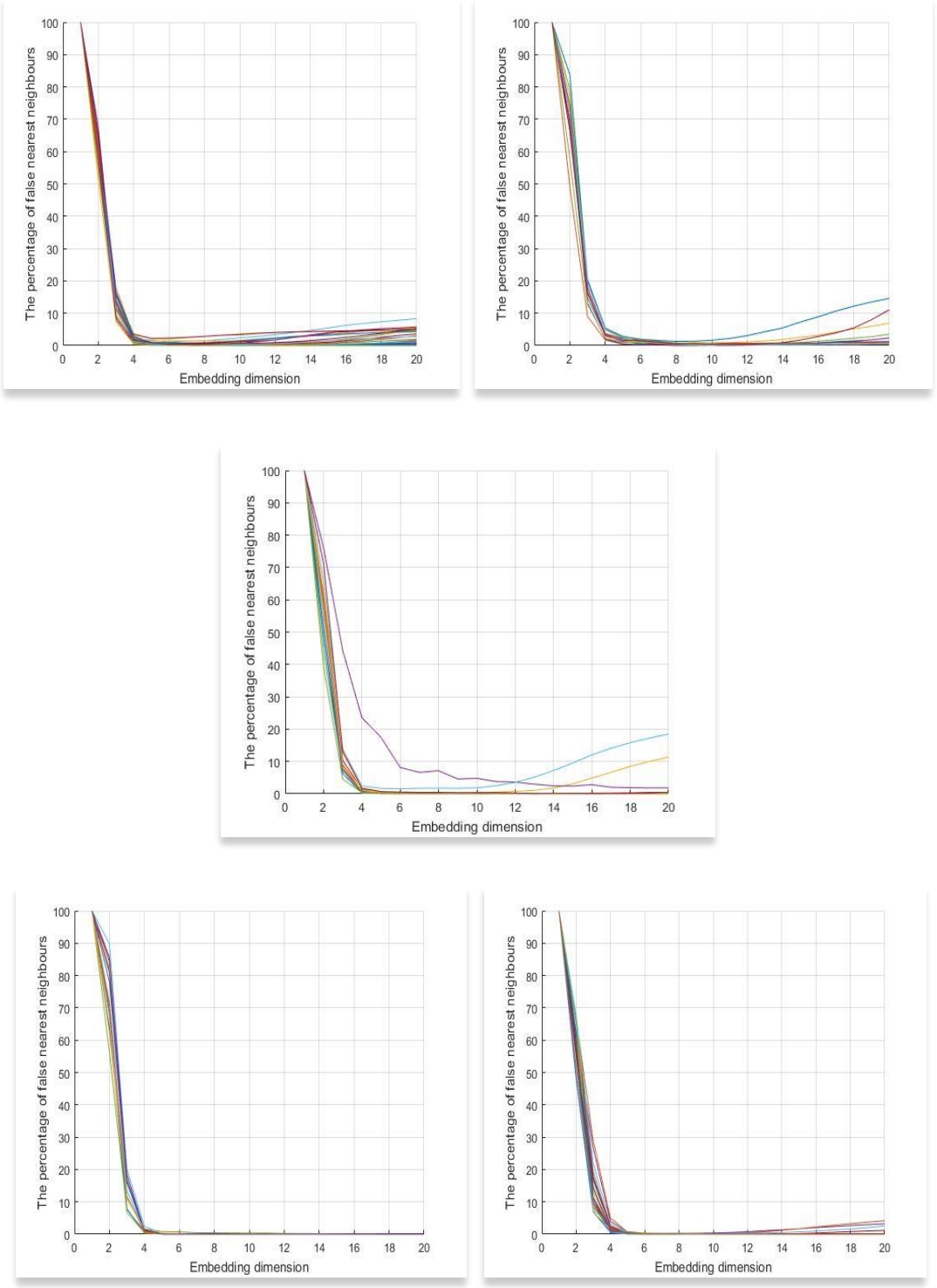


Fig. 4.17 The variation of FNN with Embedding dimension for 20 different male speakers of age 20-25 sampled at frequency 44.1 kHz for Malayalam vowel (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/.

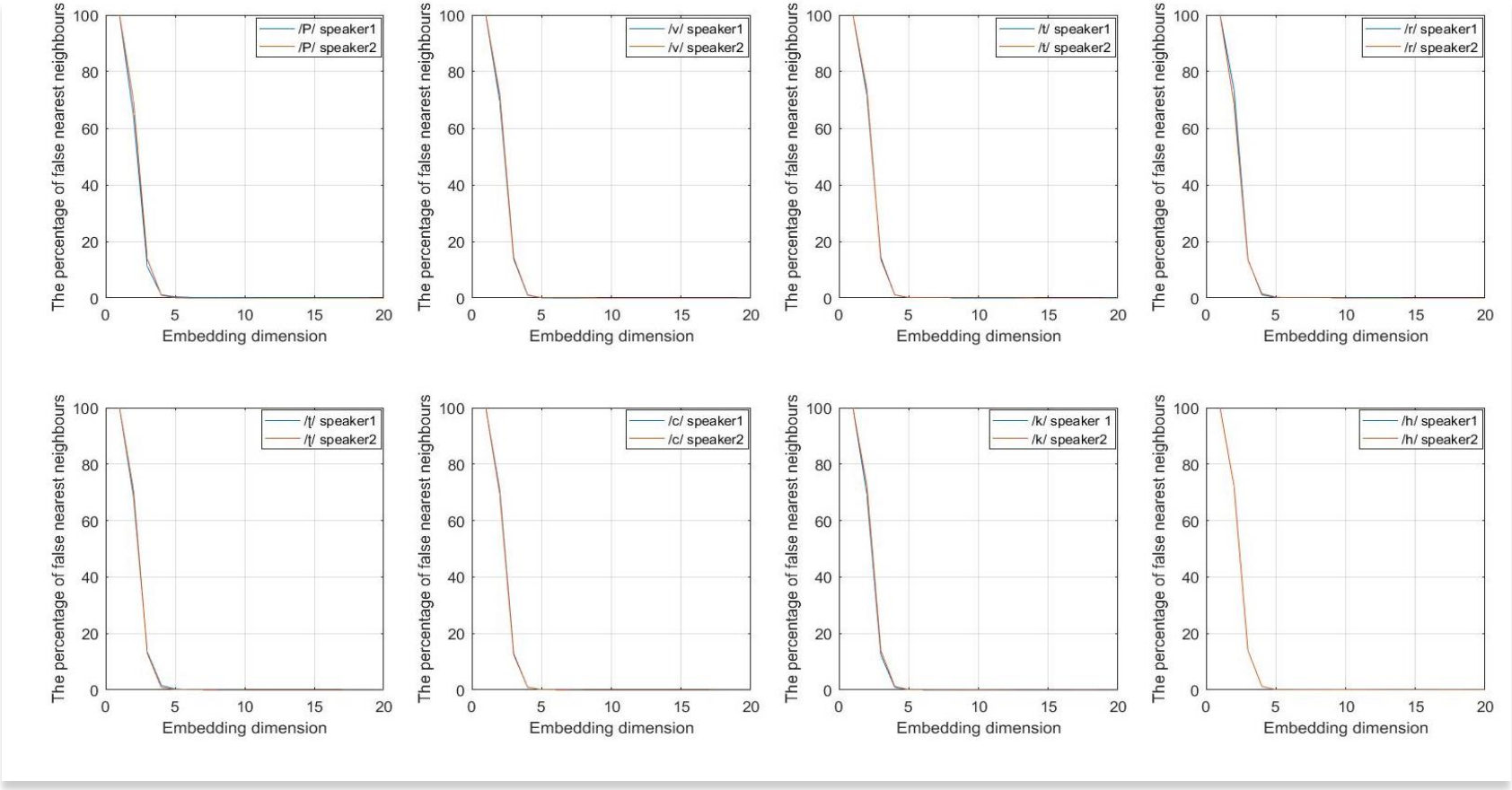


Fig. 4. 18 The variation of FNN with Embedding dimension for two female speakers of age 20-25 sampled at frequency 16 kHz for eight Malayalam consonants

(B) Results of PCA

PCA has done for five Malayalam short vowel time series obtained from speakers of different age group (both genders) at different sampling frequencies and the results are analysed. Speech samples with various gender, age groups, utterances, and sampling frequencies (Table 4.1) are analysed. Vowel and consonant phonemes are examined separately and collectively. Figures 4.19 to 4.21 demonstrate the mean normalized eigen values for Malayalam vowels (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/ uttered by female speakers of age groups 5-10, 20-25, and 60-65 sampled at frequency 16 kHz. The same is shown in Figure 4.22 for male speakers between the ages of 20 and 25. The analysis corresponding to sampling frequencies 32 kHz and 44.1 kHz are represented in Figures 4.23 and 4.24, respectively. It has been observed that the covariance matrix has six prominent eigen values for almost all the samples. The statistical analysis of the result is discussed in session 4.4.3 (C).

The results of PCA Malayalam phoneme time series taken from speakers of various ages (both genders) at various sample frequencies are analysed. Separately and jointly, vowel and consonant phonemes are studied. The mean normalised eigen values for Malayalam vowels (1) അ/a/, (2) ഇ/i/, (3) എ/e/, (4) ഒ/o/ and (5) ഉ/u/pronounced by female speakers of age groups 5-10, 20-25, and 60-65 recorded at frequency 16 kHz are shown in Figures 4.19 to 4.21. Figure 4.22 shows the same for male speakers between the ages of 20 and 25. Fig. 4.23 and Fig. 4.24 show the results of the analysis at sampling frequencies of 32 kHz and 44.1 kHz, respectively. Almost all of the samples have six notable eigen values, according to the covariance matrix. In session 4.3.3, the statistical analysis of the results is discussed.

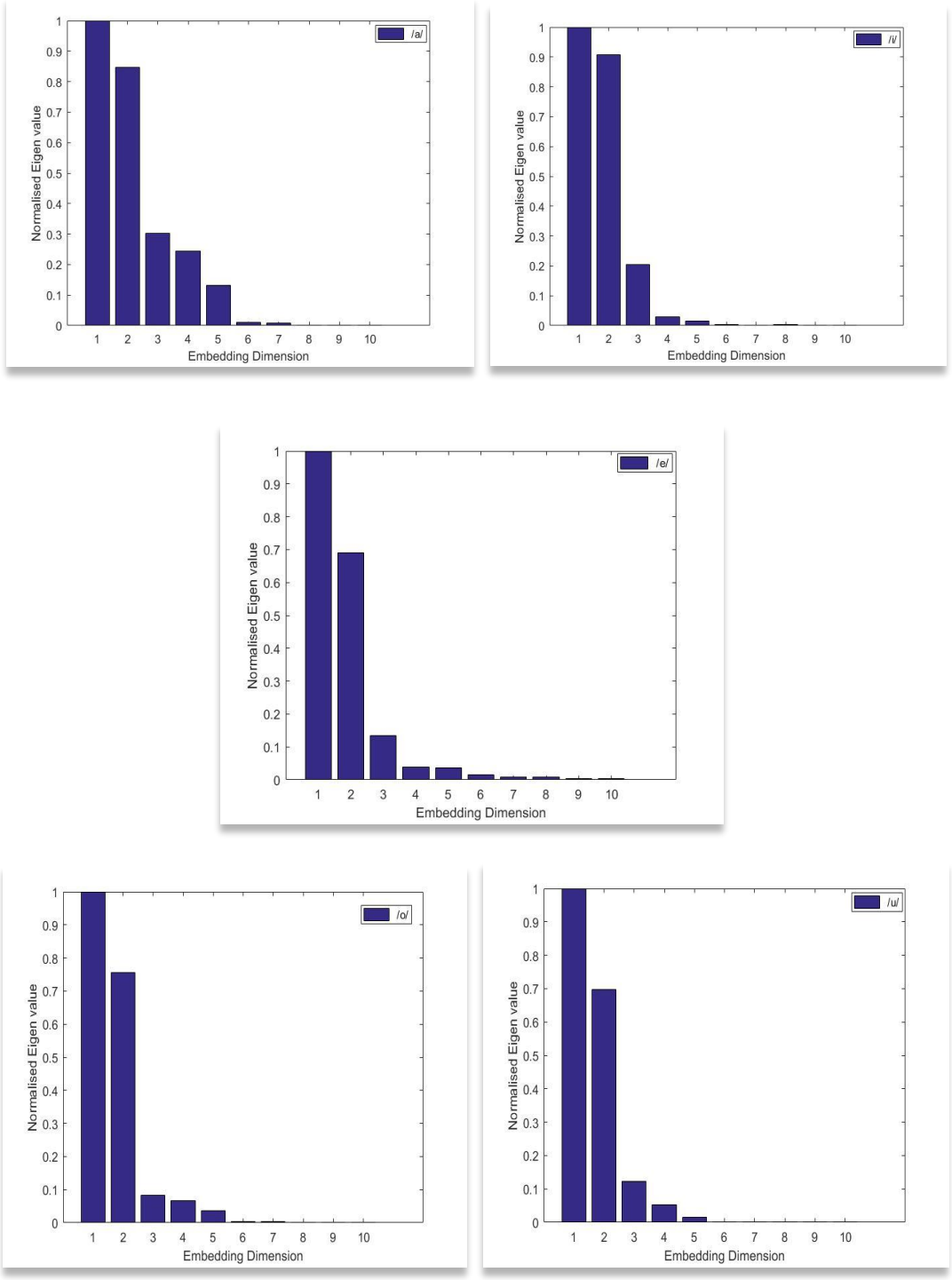


Fig. 4.19 The variation of Mean Normalised Eigen value with Embedding dimension for 20 different female speakers of age 5-10 sampled at frequency 16 kHz for Malayalam vowels.

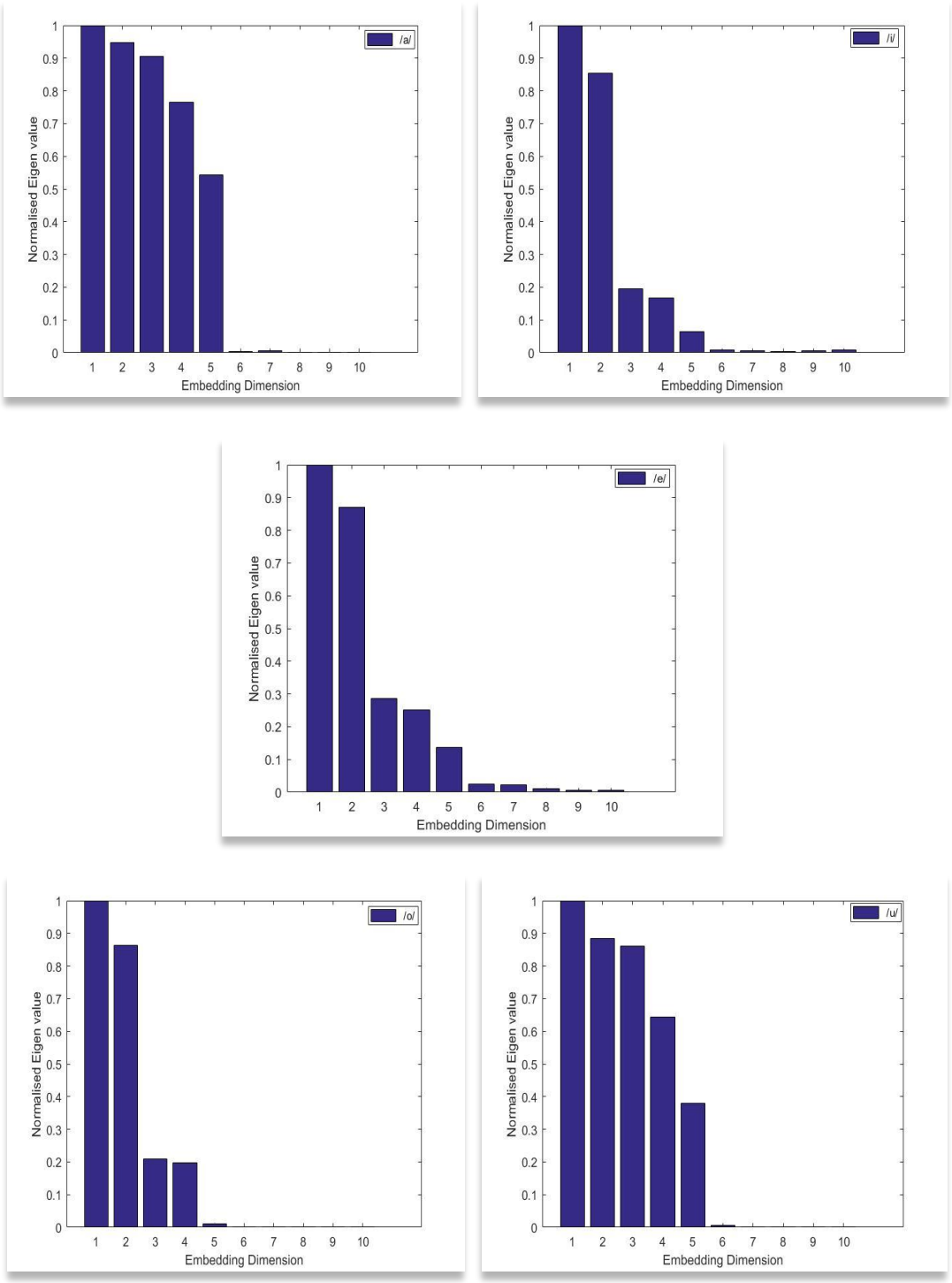


Fig. 4.20 The variation of Mean Normalised Eigen value with Embedding dimension for 50 different female speakers of age 20-25 sampled at frequency 16 kHz for Malayalam vowels.

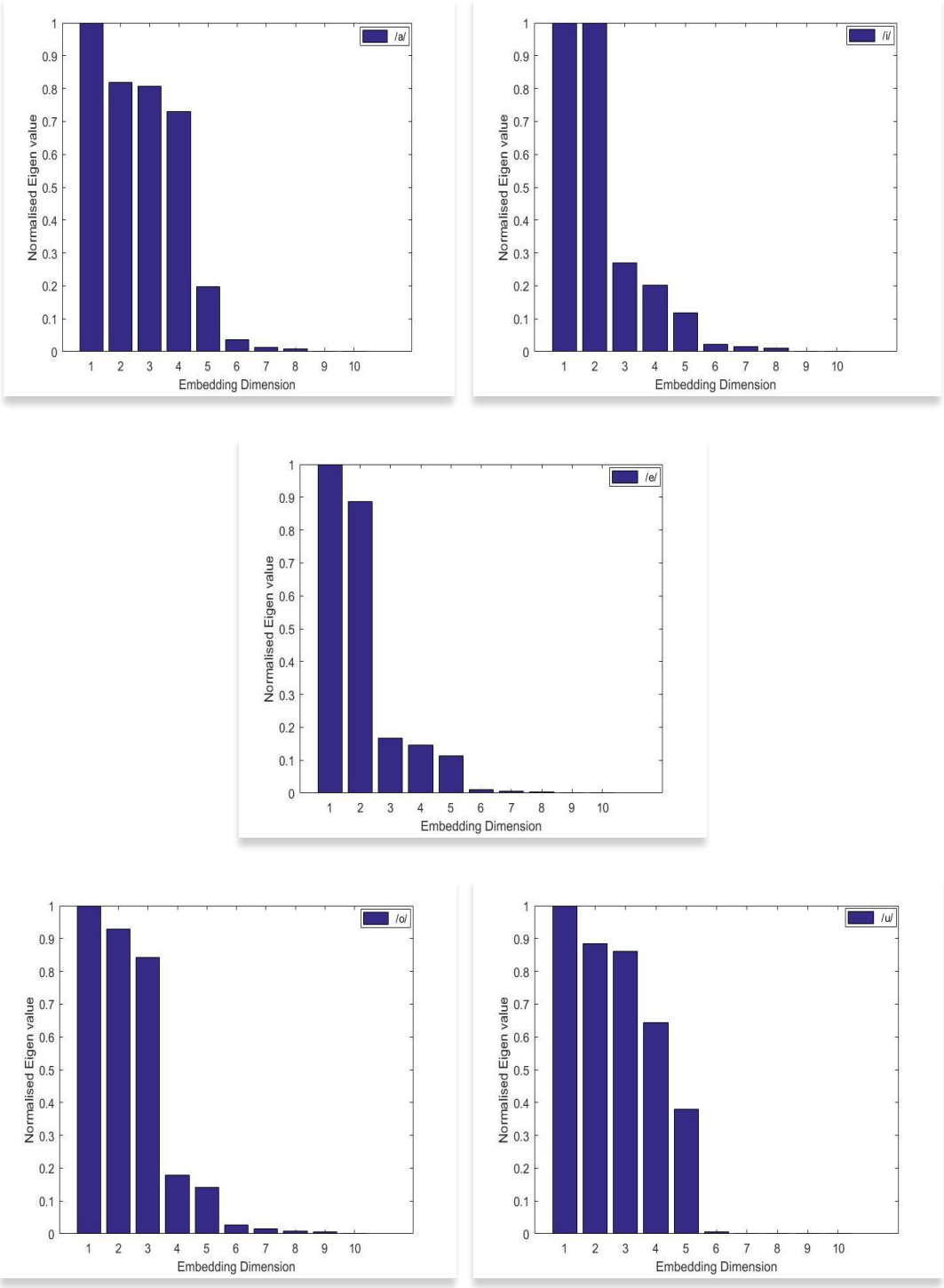


Fig. 4.21 The variation of Mean Normalised Eigen value with Embedding dimension for 30 different female speakers of age 60-65 sampled at frequency 16 kHz for Malayalam vowels.

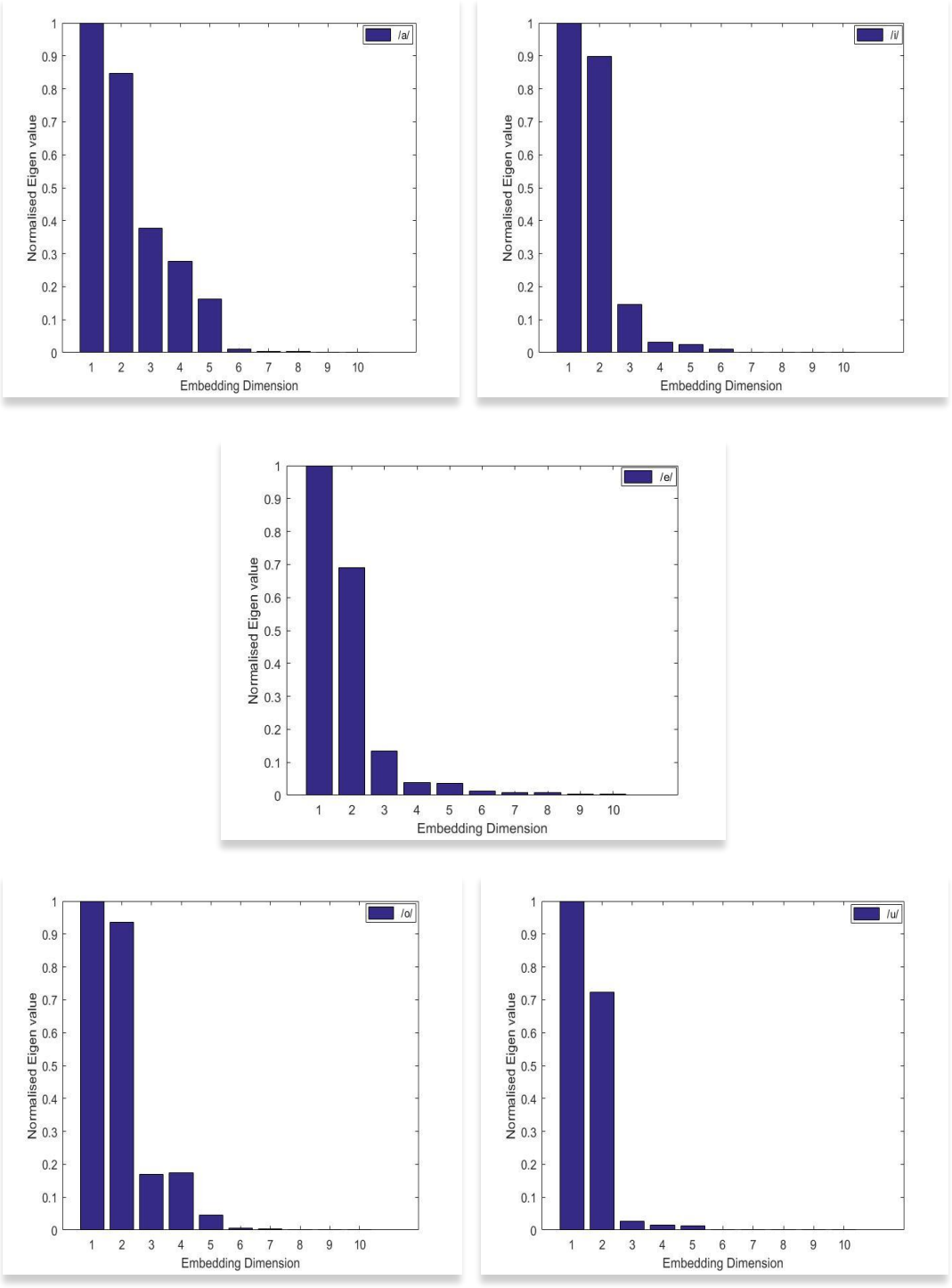


Fig. 4.22 The variation of Mean Normalised Eigen value with Embedding dimension for 50 different male speakers of age 20-25 sampled at frequency 16 kHz for Malayalam vowels.

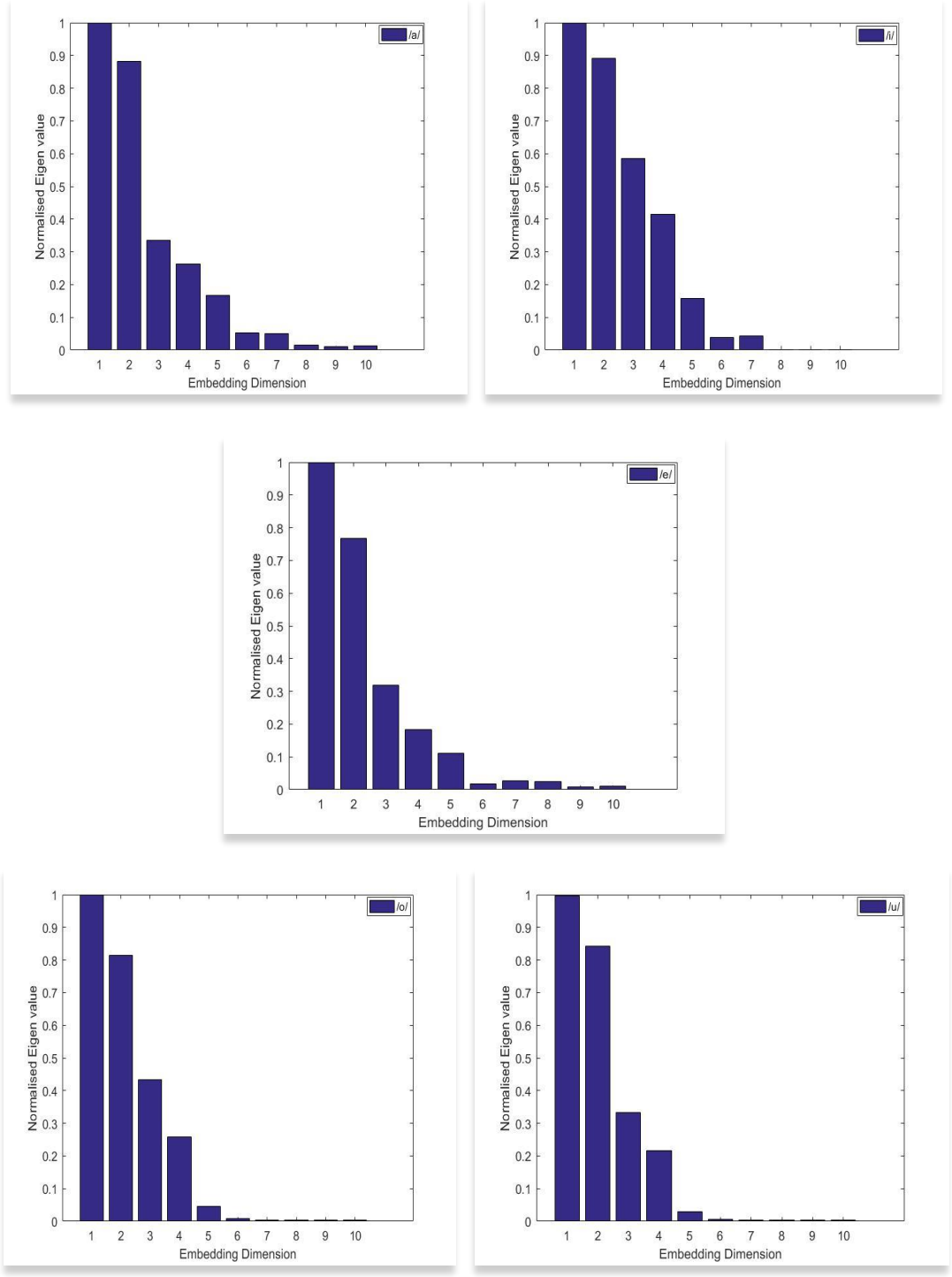


Fig. 4.23 The variation of Mean Normalised Eigen value with Embedding dimension for 50 different female speakers of age 20-25 sampled at frequency 32 kHz for Malayalam vowels.

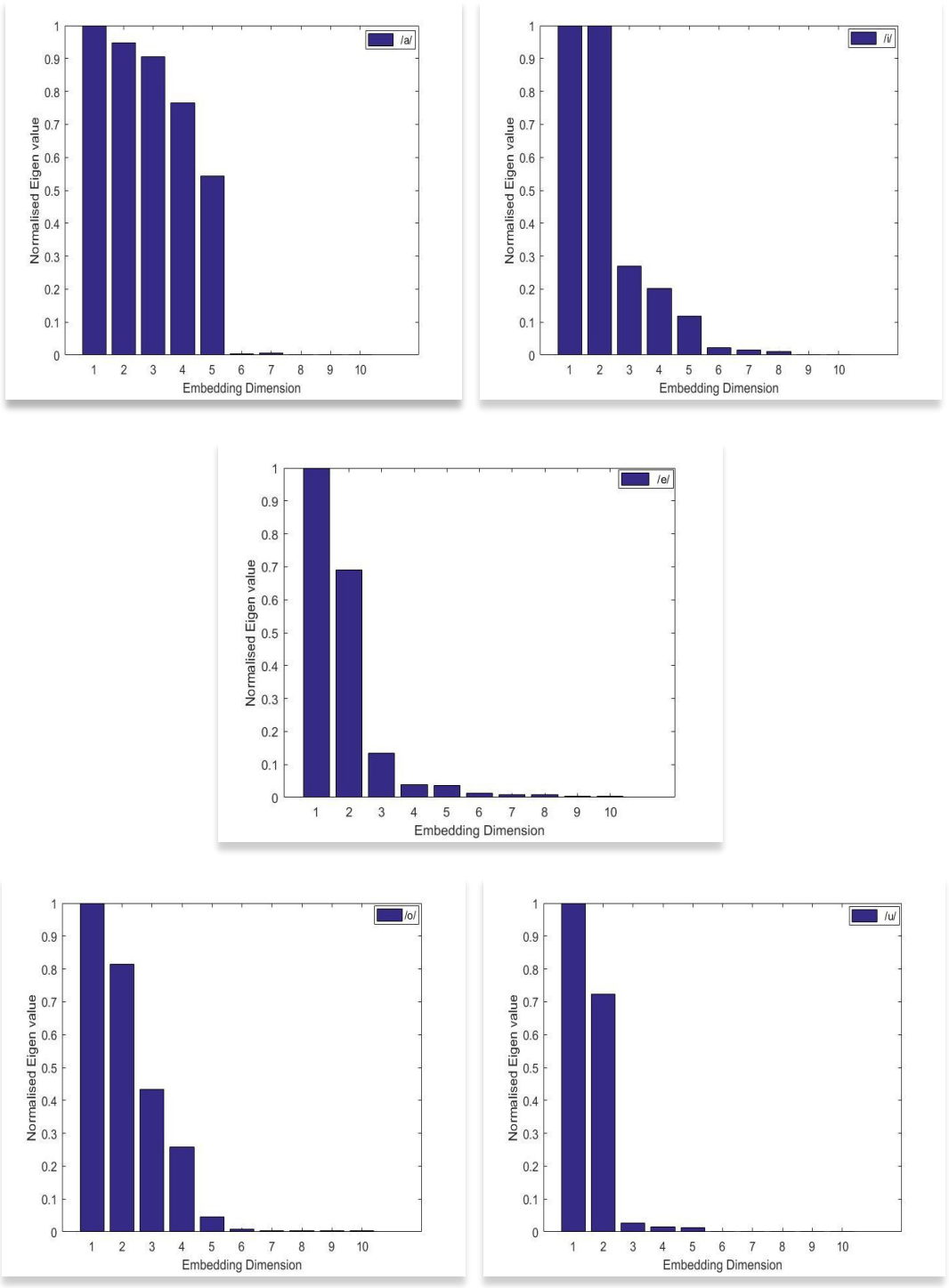


Fig. 4.24 The variation of Mean Normalised Eigen value with Embedding dimension for 50 different female speakers of age 20-25 sampled at frequency 44.1 kHz for Malayalam vowels.

(C) Result Analysis(FNN and PCA)

The outcomes from FNN and PCA are pooled for each sample, and the probability distribution function for each analysed sample is shown. For this purpose, all of the phonemes from a certain class are combined into a single unit. The distribution plot for male voices in various age groups and frequencies is shown in Fig.4.25. Figure 4.26 depicts the same for female voices. Figure 4.27 shows the results from all male and female utterance classes. Fig.4.28 shows the resultant analysis from all of the samples that were examined. Table 4.3 summarises the result.

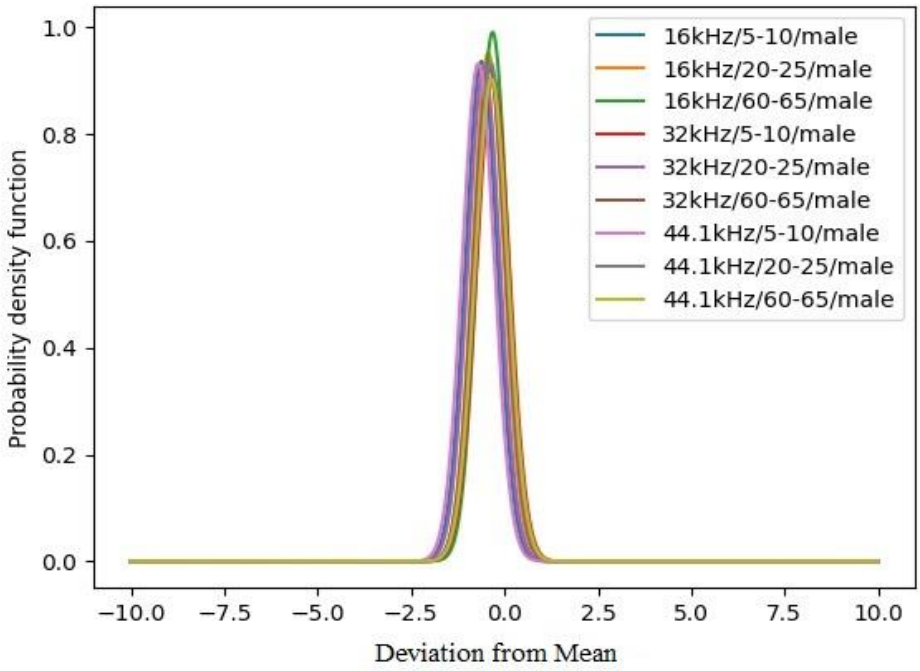


Fig. 4.25 Probability distribution of Embedding dimension (male sound).

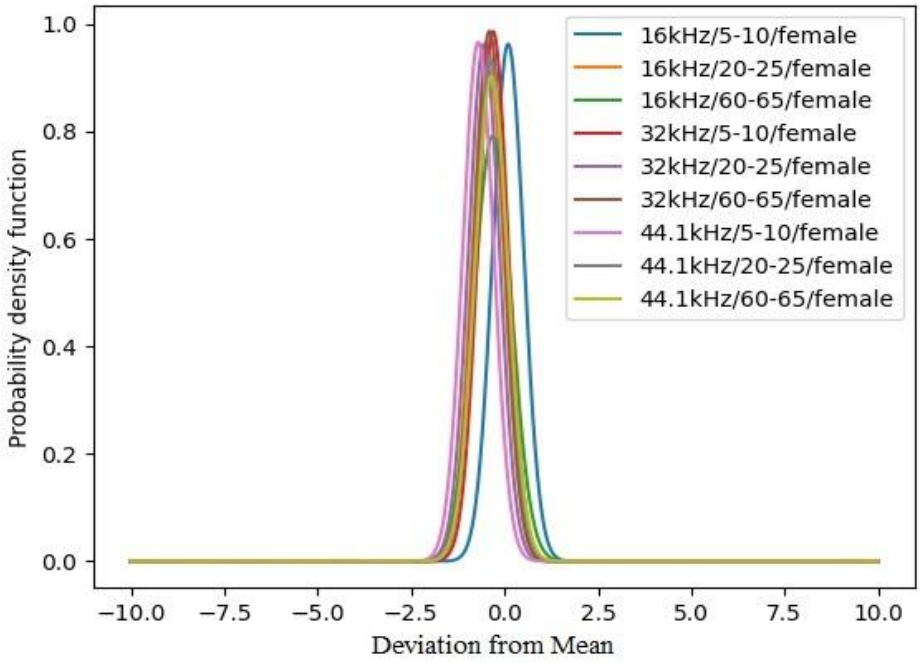


Fig. 4.26 Probability distribution of Embedding dimension (female sound).

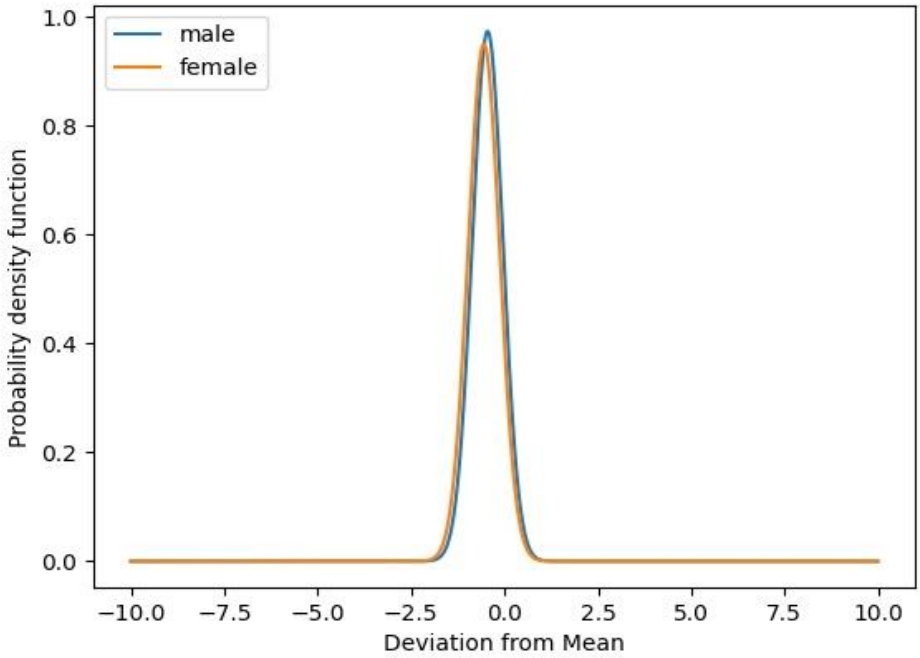


Fig. 4.27 Probability distribution of Embedding dimension (male and female sound separately).

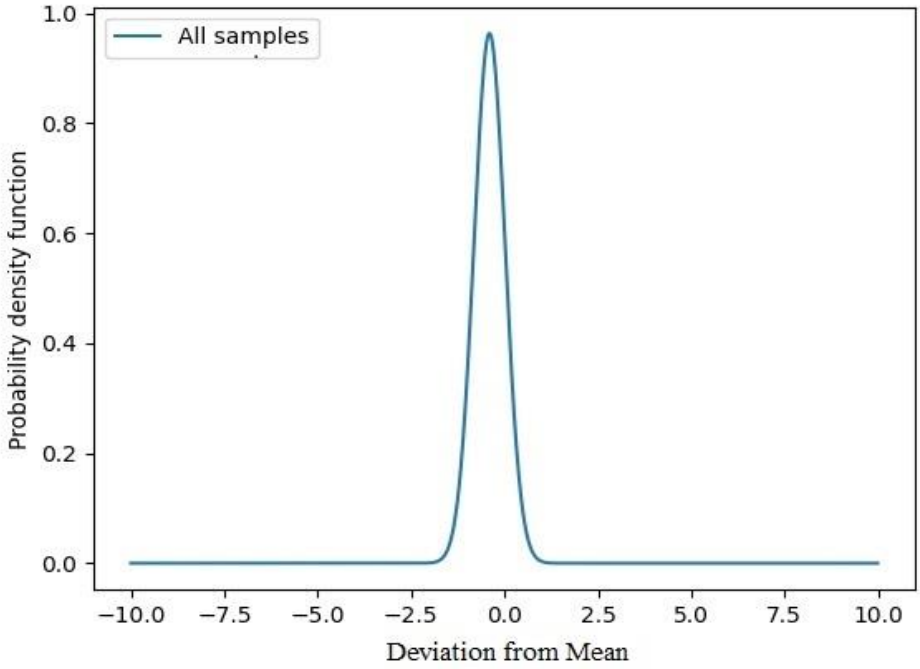


Fig. 4.28 Probability distribution of Embedding dimension (All samples).

Table 4.3 The Mean Standard Deviation (σ) of Embedding dimension for Malayalam database

Age Group	Sampling Frequency	Mean	Mode	σ (male) FNN&PCA	Mean	Mode	σ (female) FNN&PCA
5-10	16kHz	5.98	6	0.28	5.92	6	0.30
	32kHz	5.97	6	0.32	5.96	6	0.41
	44.1kHz	5.99	6	0.43	6.03	6	0.40
20-25	16kHz	5.95	6	0.41	5.95	6	0.35
	32kHz	5.98	6	0.29	5.93	6	0.34
	44.1kHz	5.93	6	0.33	6.12	6	0.37
60-65	16kHz	5.90	6	0.34	5.92	6	0.38
	32kHz	5.88	6	0.36	5.97	6	0.41
	44.1kHz	6.02	6	0.41	6.03	6	0.36

Table 4.3 shows that all of the studied samples' standard deviations are minimal (by both FNN and PCA), and the 'mode' for all data samples is six, with the mean being close to it. Male samples have an average standard deviation of 0.35, while female samples have an average standard deviation of

0.33. With a 'mode' value of six, the overall fluctuation is found to be 0.34. Hence a six-dimensional hyperspace is enough to reconstruct the phase space of the speech production system from the speech signal.

4.5. Conclusion

The hypothetical abstract space that reflects the system under study will aid the detailed analysis of the system's dynamical behaviour. Using Lorenz and Rossler systems as model systems and a speech production system as the system under study, the problem of optimising the minimum embedding dimension is addressed. The mutual information method was used to determine the time delay of embedding, and it was observed that it varies from sample to sample. The applicability of the methodologies FNN and PCA is confirmed by calculating the embedding dimension for Lorenz and Rossler systems, which is three. When dealing with Malayalam phoneme time series, it was found that the embedding dimension of the speech production system is unaffected by age, gender, or sampling frequency. The mode value of the tested samples is six, with a mean value close to it. The mode value can be used as the minimal embedding dimension as the standard deviation is so small. According to the data analysis, a six-dimensional hyperspace reconstruction can be used to analyse and model the speech production system. It should be reconstructed, and the phase space features evaluated in order to study the dynamics of underlying dynamics in the phase space. Before beginning, some tests, such as surrogate data analysis, should be used to ensure that the signal is nonlinear. Surrogate data analysis is discussed in the following chapter.

CHAPTER 5

DETECTING NONLINEARITY IN SPEECH: SURROGATE DATA ANALYSIS

5.1 Introduction

With the advancement of nonlinear dynamical studies, many investigators have striven to learn whether a nonlinear model can represent the speech production system [132], [133]. Recently, many authors [52], [134] have stressed the importance of nonlinearity in the study of vocal tract system and suggested that the nonlinearity measures like Correlation dimension (D_2) and Correlation entropy (K_2) could provide a new observational window into the complex mechanism of sound production. The usage of the methods of chaotic signal analysis with a short time series, like the phoneme time series in speech utterance, is addressed in [133]. The nonlinear measures in speech are used for various tasks such as the analysis of pathological voices, detection of polyps in the larynx, classification of speech signals, linguistic categorisation of speech and for the study of dynamics of the physical system.

The evaluation of time series arising out of dynamical systems often requires the estimation of a few nonlinear measures to describe the behaviour of the underlying dynamics. Without proper statistical testing, we cannot guarantee the reliability of a single value obtained with these measures [135]. While working with limited data sets, like speech phoneme time series, a specific hypothesis can be tested using statistical methods like surrogate data to assure the reliability of the result. The surrogate method for nonlinearity testing in a time series gained immense popularity after the work of [136]. It was emphasised that the use of nonlinear analysis approaches has to be justified by testing nonlinearity in the time series. Every time a nonlinear

measure, like D_2 or K_2, is implemented to original time series along with the assortment of surrogates and if the outcome from the time series deviates significantly from the surrogate data, then the time series is likely to have a nonlinear origin. The typically used surrogates, for testing nonlinearity in time series, are Fourier Transform (FT) surrogates, Amplitude Adjusted Fourier Transform(AAFT)surrogates and Iterative Amplitude Adjusted Fourier Transform (IAAFT) surrogates. They have been widely studied and utilized in a diverse range of applications [135]. Theiler and co-workers originally proposed the AAFT method used to generate surrogate data [136]. But a more consistent method IAAFT was proposed by [137]. The most widely accepted method used to calculate the D_2 and K_2 is Graaberger-Proccatia (GP) algorithm. K.P.Harikrishnan suggested a modified algorithm for the calculation of correlation dimension [138] and correlation entropy [139].

Though a large number of studies are reported on the use of nonlinear discriminating measures like D_2, Lyapunov exponent and K_2, the surrogate data analysis of speech data is not yet widely studied. The surrogate data analysis of EEG signals was conducted based on Lyapunov exponent, and D_2 was analysed in [140]. [141] studied the nonlinear dynamics of standard vowels by the spike and wave surrogate analysis which describes Wayland translation error as the nonlinear discriminating statistic. Since the nonlinear parameters may have the dependency on linguistic parameters, vocal fold structures and emotional aspects, surrogate data analysis is highly essential before using the nonlinear parameters as discriminating measures in applications for almost any data set in a particular language.

This work aims to test the nonlinearity in time series of Malayalam speech database. The time series obtained from short Malayalam vowel utterance of 50 same-aged female speakers with 100 surrogates of each sample is used for the analysis. The standard nonlinear systems, Lorenz

system and Rossler system, are used as model systems.D_2 and K_2 at the minimum embedding dimension are used as nonlinear discriminating measures. IAAFT surrogates generated from the Python package is used for the analysis. The minimum embedding dimension of the speech time series is obtained using the False Nearest Neighbour (FNN) method [123]. For calculating D_2 and K_2, the algorithm suggested [138], [142] is used. D_2 and K_2 at minimum embedding dimension is used as the nonlinear measure for detecting the underlying nonlinearity in the speech time series.

This chapter is organized as follows. Section 5.2 describes the generalised fractal dimension and entropy which are the discriminating measures used. The theory and algorithm of surrogate data generation is explained in section 5.3.The results of surrogate analysis using D_2 and K_2 is explained and analysed in section 5.4.Section 5.5 concludes the chapter.

5.2 Generalised Fractal Dimension and Entropy

Fractals are geometric shapes that are complex geometric shapes with fine structure at arbitrarily small scales. The self-similarity is one of the attributes of fractals. In most cases, the similarity is only approximate or statistical. Fractals are intriguing because of their exquisite combination of beauty, complexity, and unbounded structure. In a way that traditional shapes like cones and squares can't, they evoke natural objects like mountains, clouds, coastlines, and blood vessel networks. They've also shown to be useful in a variety of scientific applications, including computer graphics and image compression, as well as crack structural mechanics and viscous fingering fluid mechanics.

The capacity and Hausdorff dimensions of a fractal are purely geometric, making no mention about the measure of the attractor. These two measures are not enough to characterise the behaviour of the dynamics of a complicated system in phase space. Generalised dimensions are an attempt to

solve this issue and are highly useful in characterizing a dynamical system. Two formulations of the generalized dimensions, box counting approach and partition function approach, are discussed here.

5.2.1 Box counting approach

Cover the attractor with boxes of size 'R' and define a probability of finding a point in the i^{th} box.

$$P_i = \frac{N_i}{N} \tag{5.1}$$

Where N_i is the number of points in the i^{th} box and Nis the total number of points. The q^{th} generalized dimension is defined as

$$D_q = \lim_{R\to 0}\left\{\frac{1}{q-1}\frac{log\sum_i P_i^q}{\log R}\right\} \tag{5.2}$$

5.2.2 Partition function approach

In this approach the set is covered with 'm' boxes of size 'l_i' which is less than R. The partition function is defined as

For q≤ 1 and τ≤0

$$\Gamma(q,\tau,R) = inf\sum_i \frac{P_i^q}{l_i^\tau} \tag{5.3}$$

For q ≥1 and τ≥0

$$\Gamma(q,\tau,R) = sup\sum_i \frac{P_i^q}{l_i^\tau} \tag{5.4}$$

And there exist a τ(q) such that $\lim_{R\to 0}\Gamma(q,\tau,R)$ is zero for τ<τ(q) and infinity for τ>τ(q)

The generalized dimension is defined as

$$D_q^H = \frac{\tau(q)}{q-1} \tag{5.5}$$

5.2.3 Correlation Dimension (D_2)

When q=0 generalized dimension reduces to Capacity dimension for box counting algorithm and Hausdorff dimension for partition function algorithm. For q=1,the generalized dimension reduces to information dimension (D_1)

$$D_1 = \lim_{q\to 1}\left\{\lim_{R\to 0}\left[\frac{1}{q-1}\frac{log\sum_i P_i P_i^{q-1}}{log\ R}\right]\right\} \tag{5.6}$$

$$D_1 = \lim_{q\to 1}\left\{\lim_{R\to 0}\left[\frac{1}{q-1}\frac{log\sum_i P_i(1+(q-1)logP_i)}{log\ R}\right]\right\} \tag{5.7}$$

$$D_1 = \lim_{R\to 0}\left\{\frac{-\sum_i P_i logP_i}{-log\ R}\right\} \tag{5.8}$$

For q=2

$$D_2 = \lim_{R\to 0}\left\{\frac{log\sum_i P_i^2}{log\ R}\right\} \tag{5.9}$$

Where D_2 is called correlation dimension, which is the most robust dimension among the fractal dimensions. $\sum_i P_i^2$ is the probability that two points lie within a cell of length 'R' and which is defined by the pair correlation sum integral(C_m(R)).

$$C_m(R) = \lim_{N\to\infty}\left\{\frac{1}{N^2}\sum_{i,j}^{N} H(\boldsymbol{R} - \left|\boldsymbol{X_i} - \boldsymbol{X_j}\right|\right\} \tag{5.10}$$

Where X_i and X_j are delay vectors given by equation 4.3 with n=i, j. N is the total number of reconstructed vectors, H stands for Heaviside step function and 'm' indicates embedding dimension. The Correlation dimension (D_2) is the scaling index of the variation of C_m(R) with R as R tends to zero.

$$D_2 = \lim_{R\to 0}\left\{\frac{log\ (C_m(R)}{log(R)}\right\} \tag{5.11}$$

5.2.4 Correlation Entropy (K_2)

Entropy is a thermodynamic quantity that describes how disordered a system is. The term entropy can be used to describe the amount of information stored in a wider range of probability distributions. The information theory is a useful tool for analysing time series data. A system's observation can be viewed as a source of data, a stream of numbers that can be viewed as a transmitted message.

The numerical value of a time series' entropy is interesting for two reasons. To begin with, its inverse is the time scale that is relevant to the system's predictability. It also provides topological information about the folding process. Entropies are difficult to extract from time series because they require more data points than dimensions to compute. Although generalised entropies for various 'q' values can be defined, the Correlation entropy(K_2), which can be defined from the correlation sum, is the most efficient and useful entropy term.

The computation of Correlation function entails the number of trajectory points in the embedding space remain within the space R of each other. Hence as the embedding dimension increases Correlation function decreases for a fixed value of R [142].The relation defines the correlation entropy (K_2) is

$$C_M(R)\ \alpha\ \ e^{-mK_2\Delta t} \tag{5.12}$$

Where Δt is the time step between consicutive values in the time series, which can be obtained from the sampling frequency. From this K_2 can be written as

$$K_2 = \frac{1}{\Delta t}\lim_{R\to 0} log\left\{\frac{C_M(R)}{C_{M+1}(R)}\right\} \tag{5.13}$$

5.3 Surrogate Analysis

In some cases, we may want to classify or characterise our time series using only a few numbers (dimension, entropy etc). They can be used as indicators of some underlying concealed process that evolves at a slower rate than observed dynamics. Sometimes the estimated measures are just a bunch of numbers derived from algorithmic calculations with no physical meaning. We must ensure that the nonlinear metrics used in the analysis are reliable. If we don't have strong evidence that we're measuring dynamical variables, we can use statistical hypothesis testing to assess the likelihood that our measurements are reflecting physical reality. Surrogate data analysis is one of the most popular methods for ensuring nonlinearity.

The rationale behind the surrogate analysis is to formulate a null hypothesis that a stationary linear stochastic process has created the data, and then to attempt to reject it by comparing a suitable measure for the data with proper realisations of any surrogate data. The main goal of surrogate analysis is to compare the distribution of a nonlinear metric estimated from available time series to the distribution of the same metric obtained from a large number of time series that satisfy a null hypothesis. It’s plausible that the time series are generated by a linear stochastic process as the null hypothesis. Then we can calculate the possibility that our metric is real or purely coincidental. We can reject the null hypothesis if the probability of that being true is high. The number of surrogate signals used will determine the confidence level at which the null hypothesis can be rejected.

In order for this method to work, our surrogate data must have some properties in common in order to establish the null hypothesis. We usually require some of the linear properties (mean, autocorrelation, power spectrum, etc.) of the test time series and the surrogates to be the same as we are primarily interested in testing nonlinear measures. The null hypothesis that

the time series are generated by an autoregressive moving average (ARMA) model would be supported by this.

In this work, Iterated amplitude adjusted Fourier transformed (IAAFT) surrogates are used. These surrogates preserve all the linear statistical and spectral properties in the surrogates (mean, variance, probability distribution, autocorrelation, power spectrum etc). However, any nonlinear structures in the data will be eliminated by the randomization procedure. The different steps in constructing the IAAFT surrogates are discussed below.

1. Sort the original time series x_n (n=0,1,2....,N-1) by its magnitude is ascending order in to y_n and sorting it with the corresponding Fourier transform magnitudes.

$$X_n = \frac{1}{N}\left|\sum_{k=0}^{N-1} x_k\, e^{j2\pi nk/N}\right| \tag{5.14}$$

2.The iterative process is initiated by randomly reshuffling x_n to get $\{r_n^{(0)}\}$

3.Take Digital Fast Fourier Transform(DFFT) of $\{r_n^i\}$,where i varies from 0 to

N-1

$$R_n^{(i)} = \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1} r_k^{(i)} e^{j2\pi nk/N} \tag{5.15}$$

4.Now replace the magnitude $\left|R_n^{(i)}\right|$ with X_n, but keep the phase $\left(\Phi_n^{(i)}\right)$ same and take Inverse Digital Fast Fourier Transform (IDFFT) to get

$$x_n^{(i)} = \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1} X_k e^{j\Phi_k^{(i)}} e^{-j2\pi nk/N} \tag{5.16}$$

5. Finally transform $\{x_n^{(i)}\}$ in to $\{r_n^{i+1}\}$ by rank ordering according to the stored y_n

The rejection of the null hypothesis is done by a parameter called statistical significance level (S)[139]given by

$$S = \frac{|f - \langle f \rangle_{surr}|}{\sigma_{surr}} \tag{5.17}$$

Where f is a nonlinear measure of the data which is suitable for analysis. $\langle f \rangle_{surr}$ denotes the average value of the distribution fora number of surrogates and σ_{surr} is its standard deviation for surrogates.

5.4 Experiments and Results

For the comparison with the Malayalam phoneme database, time series derived from the Lorenz and Rossler systems (section 3.1) are used. The vowel phonemes and consonant phonemes (one from each group) of 50 male and female speakers in the age group of 20-25, sampled at 16 kHz, are used from the database discussed in section 3.2. For each sample, 100 surrogates are generated, and the significance level is determined for each sample and averaged among all samples. Python was used to implement the IAAFT algorithm for surrogate generation. The preservation of the linearity properties of the surrogates is ensured by analysing the autocorrelation and power spectrum. The Correlation dimension at the minimum embedding dimension (D_{2m}) and correlation entropy at the minimum embedding dimension (K_{2m})are used as nonlinear discriminating measures. Because it was optimised in chapter 3, the minimum embedding dimension was chosen to be six.

5.4.1 Results of Surrogate Analysis using D_{2m}

When D_{2m} is used as the nonlinear discriminating measure, the significance level will be written as

$$S = \frac{|D_{2m} - \langle D_{2m} \rangle_{surr}|}{\sigma_{surr}} \tag{5.18}$$

Where $\langle D_{2m} \rangle_{surr}$ is the average of the D_{2m} for a number of surrogates and σ_{surr} is the standard deviation of D_{2m} distribution for surrogates. D_2 values are determined for each of samples for *m* from 1 to 5 for Lorenz and Rossler system and from 1 to 7 for Malayalam speech time series. Fig. 5.1 and 5.2, respectively show the variation of D_2 with embedding dimension with the histogram of D_{2m} for surrogates for Lorenz and Rossler system.

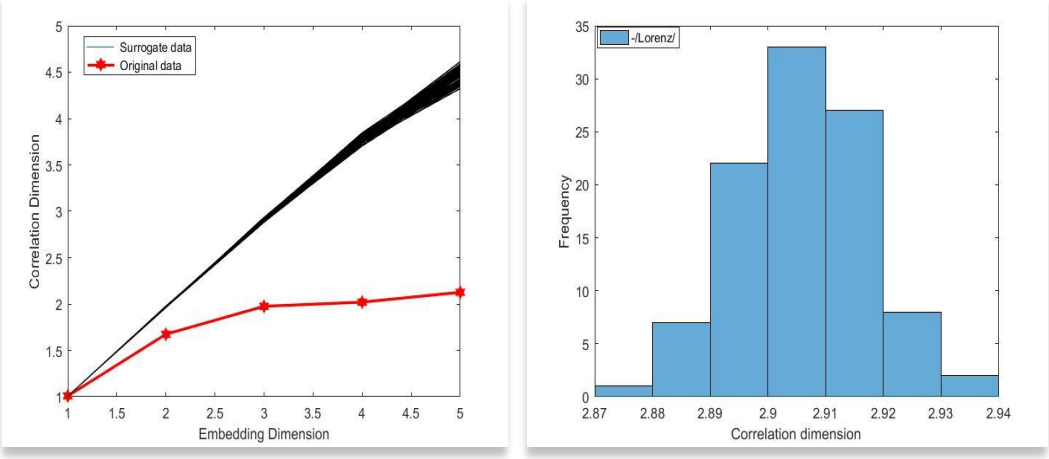


Fig. 5.1 (a) Variation of $\mathbf{D_2}$ with ***m*** for original time series and 100 surrogates for Lorenz system. (b) Histogram of D_{2m} for 100 surrogates

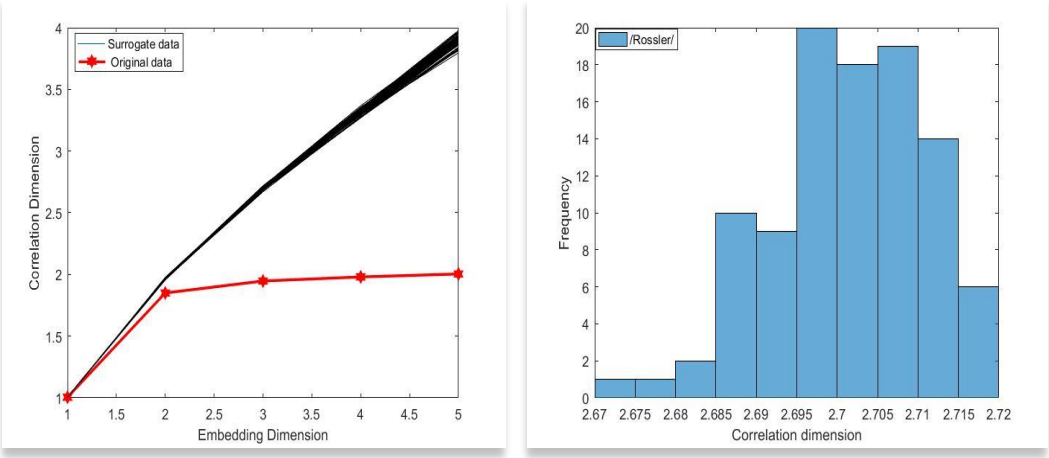


Fig. 5.2 (a) Variation of $\mathbf{D_2}$ with ***m*** for original time series and 100 surrogates for Rossler system. (b) Histogram of D_{2m} for 100 surrogates

The correlation sum (C(R)) for different samples (Malayalam vowels /a/, /i/, /e/, /o/, and /u/) is calculated and plotted against R using equation 5.10. Fig. 5.3 shows the behaviour of the correlation sum at different dimensions for Malayalam short vowel speech utterances.

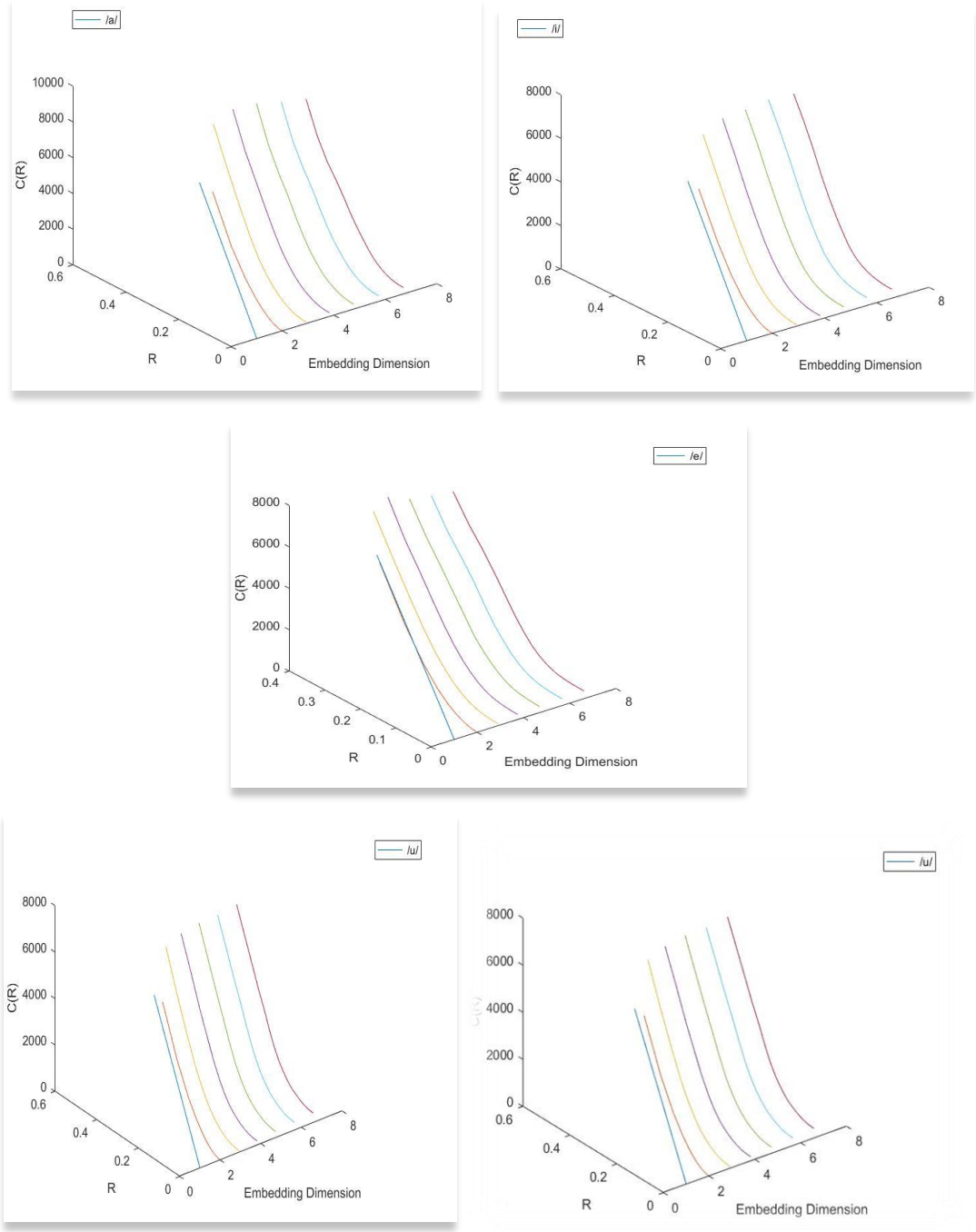


Fig. 5.3 Variation of Correlation sum (C(R)) with R at different embedding dimensions

From the correlation sum, D_2 has been calculated for Malayalam vowels അ /a/, ഇ /i /, എ /e/, ഒ /o/ and ഉ /u/ and consonants പ/P/,വ/v,/ത/t/,റ/r̲/,ട/ṭ/,ച /c/,ക /k/,ഹ /h/,(for all speakers from the sample) at embedding dimensions from 1 to7. The average D_2 at the embedding dimension six (D_{2m}) for Lorenz system, Rossler system and Malayalam vowel utterance are tabulated in Table 5.1. The Figures 5.4 to 5.8 show the variation of D_2 with embedding dimension with the histogram of D_{2m} for surrogates (for a specific speaker from the sample).

Table 5.1 The average D_{2m} values for Lorenz, Rossler and Malayalam vowels.

	Lorenz	**Rossler**	**അ/a/**	**ഇ/i/**	**എ/e/**	**ഒ/o/**	**ഉ/u/**
Original	1.975	1.947	2.627	2.960	2.482	2.287	2.857
Surrogate1	2.916	2.704	4.481	4.674	4.388	3.976	4.138
2	2.910	2.695	4.387	4.634	4.532	4.099	3.905
3	2.917	2.702	4.479	4.634	4.420	3.712	4.129
4	2.920	2.699	4.432	4.560	4.537	3.929	3.914
5	2.881	2.706	4.427	4.503	4.546	3.842	4.105
6	2.894	2.698	4.598	4.605	4.506	3.803	3.954
7	2.904	2.702	4.513	4.699	4.458	3.852	3.907
8	2.908	2.704	4.392	4.577	4.334	3.919	4.195
9	2.901	2.701	4.517	4.469	4.430	3.549	3.981
10	2.890	2.707	4.453	4.541	4.588	3.512	3.974
11	2.890	2.708	4.431	4.728	4.458	4.041	4.031
12	2.914	2.695	4.363	4.709	4.434	3.876	4.114
13	2.891	2.697	4.465	4.768	4.551	3.658	4.030
14	2.907	2.692	4.432	4.440	4.573	3.482	4.170
15	2.924	2.690	4.445	4.712	4.446	3.289	3.561
16	2.911	2.699	4.508	4.539	4.579	3.482	4.180
17	2.933	2.705	4.540	4.654	4.464	4.045	4.094
18	2.905	2.701	4.434	4.598	4.445	3.802	3.939
19	2.891	2.709	4.488	4.722	4.460	3.843	3.892
20	2.905	2.716	4.523	4.762	4.364	3.552	4.120
21	2.879	2.703	4.526	4.611	4.302	4.108	4.369
22	2.917	2.706	4.513	4.744	4.363	3.745	4.029

23	2.901	2.691	4.571	4.513	4.405	3.477	3.988
24	2.905	2.698	4.512	4.295	4.601	3.943	4.089
25	2.920	2.692	4.400	4.733	4.190	3.692	3.985
26	2.904	2.689	4.558	4.698	4.479	3.653	4.197
27	2.918	2.695	4.528	4.647	4.413	3.413	3.963
28	2.913	2.714	4.502	4.712	4.452	3.900	4.293
29	2.897	2.701	4.481	4.563	4.555	3.990	4.105
30	2.909	2.687	4.442	4.815	4.514	3.260	3.804
31	2.907	2.670	4.443	4.587	4.320	3.247	3.911
32	2.888	2.711	4.510	4.789	4.482	3.820	3.920
33	2.910	2.684	4.461	4.567	4.592	3.327	3.750
34	2.891	2.709	4.511	4.803	4.543	3.983	3.529
35	2.907	2.693	4.528	4.649	4.522	3.561	3.978
36	2.909	2.714	4.591	4.650	4.565	3.333	3.927
37	2.904	2.707	4.522	4.558	4.590	3.999	4.169
38	2.915	2.695	4.460	4.641	4.606	3.774	3.877
39	2.915	2.685	4.382	4.465	4.404	3.839	4.287
40	2.906	2.706	4.507	4.648	4.588	4.019	3.925
41	2.918	2.709	4.542	4.764	4.422	3.867	3.917
42	2.909	2.711	4.399	4.778	4.614	3.841	4.001
43	2.904	2.702	4.492	4.755	4.517	3.832	4.036
44	2.908	2.704	4.553	4.606	4.614	3.911	4.129
45	2.907	2.696	4.536	4.436	4.462	3.936	3.893
46	2.896	2.696	4.432	4.593	4.344	3.630	3.895
47	2.904	2.701	4.517	4.559	4.378	3.743	4.173
48	2.916	2.716	4.507	4.533	4.590	3.445	4.066
49	2.898	2.696	4.500	4.572	4.508	3.490	3.910
50	2.889	2.686	4.329	4.574	4.393	3.828	4.137
51	2.902	2.695	4.500	4.284	4.528	3.594	4.185
52	2.906	2.710	4.561	4.650	4.465	3.550	4.277
53	2.911	2.706	4.410	4.768	4.472	3.696	4.186
54	2.899	2.688	4.465	4.701	4.546	3.893	3.940
55	2.915	2.697	4.483	4.341	4.259	3.823	4.068
56	2.902	2.712	4.425	4.605	4.518	3.993	4.020
57	2.916	2.705	4.437	4.273	4.535	3.476	3.953
58	2.896	2.711	4.413	4.233	4.504	3.948	4.282
59	2.898	2.713	4.562	4.730	4.323	3.426	3.981
60	2.890	2.710	4.464	4.433	4.419	3.774	4.173
61	2.914	2.697	4.448	4.675	4.513	3.766	4.056

62	2.919	2.691	4.527	4.741	4.618	3.679	3.638
63	2.904	2.686	4.407	4.745	4.516	3.576	3.791
64	2.914	2.702	4.485	4.475	4.565	4.014	3.905
65	2.906	2.716	4.493	4.850	4.586	3.704	3.956
66	2.898	2.700	4.462	4.685	4.400	3.839	3.251
67	2.906	2.696	4.579	4.572	4.326	3.774	4.242
68	2.892	2.687	4.546	4.625	4.319	3.747	4.242
69	2.890	2.682	4.458	4.436	4.326	3.595	4.310
70	2.883	2.704	4.389	4.737	4.420	3.899	4.227
71	2.910	2.695	4.358	4.666	4.620	3.634	3.918
72	2.896	2.706	4.514	4.711	4.490	3.694	3.893
73	2.914	2.715	4.356	4.828	4.584	3.753	4.140
74	2.888	2.694	4.564	4.492	4.400	4.006	3.609
75	2.925	2.714	4.440	4.789	4.482	3.920	4.095
76	2.897	2.703	4.458	4.392	4.569	3.700	4.100
77	2.918	2.693	4.456	4.574	4.458	3.776	3.763
78	2.912	2.699	4.515	4.637	4.385	3.868	3.953
79	2.887	2.708	4.511	4.713	4.698	3.616	3.990
80	2.924	2.719	4.476	4.635	4.105	3.882	3.961
81	2.925	2.709	4.408	4.686	4.399	3.629	3.841
82	2.937	2.710	4.622	4.683	4.321	3.811	4.245
83	2.913	2.679	4.490	4.613	4.523	3.598	4.073
84	2.903	2.706	4.472	4.667	4.453	3.723	4.005
85	2.890	2.689	4.417	4.663	4.625	3.847	4.067
86	2.897	2.699	4.435	4.614	4.477	3.529	4.110
87	2.915	2.706	4.431	4.733	4.500	3.988	4.072
88	2.905	2.715	4.473	4.642	4.593	3.854	3.894
89	2.918	2.708	4.497	4.592	4.465	3.746	3.668
90	2.907	2.713	4.434	4.765	4.576	3.820	3.773
91	2.921	2.703	4.548	4.712	4.434	4.021	3.846
92	2.897	2.703	4.429	4.560	4.540	3.816	4.196
93	2.899	2.699	4.470	4.820	4.451	3.779	4.008
94	2.902	2.711	4.395	4.690	4.386	3.765	3.845
95	2.911	2.706	4.443	4.567	4.451	3.671	3.836
96	2.905	2.689	4.520	4.547	4.602	4.177	4.057
97	2.905	2.686	4.515	4.702	4.426	3.902	3.846
98	2.908	2.691	4.462	4.525	4.577	3.617	3.980
99	2.887	2.713	4.449	4.715	4.181	3.800	4.001
100	2.922	2.700	4.469	4.704	4.461	3.625	4.228

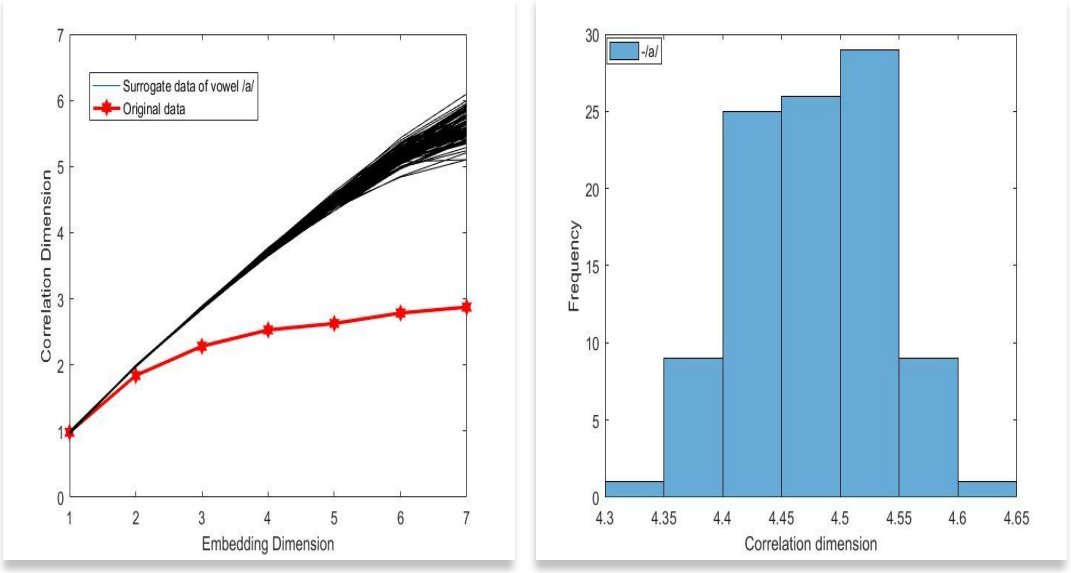


Fig. 5.4 (a) Variation of $\mathbf{D_2}$ with ***m*** for original time series (Malayalam Vowel ആ/a/) and 100 surrogates. (b) Histogram of D_{2m} for 100 surrogates

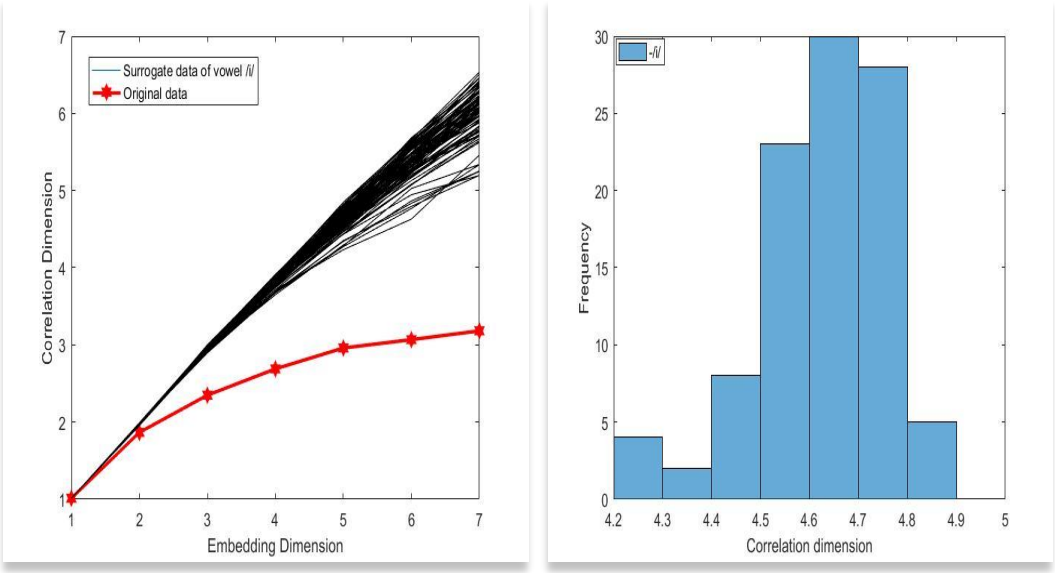


Fig. 5.5 (a) Variation of $\mathbf{D_2}$ with ***m*** for original time series (Malayalam Vowel ഇ /i/) and 100 surrogates. (b) Histogram of D_{2m} for 100 surrogates

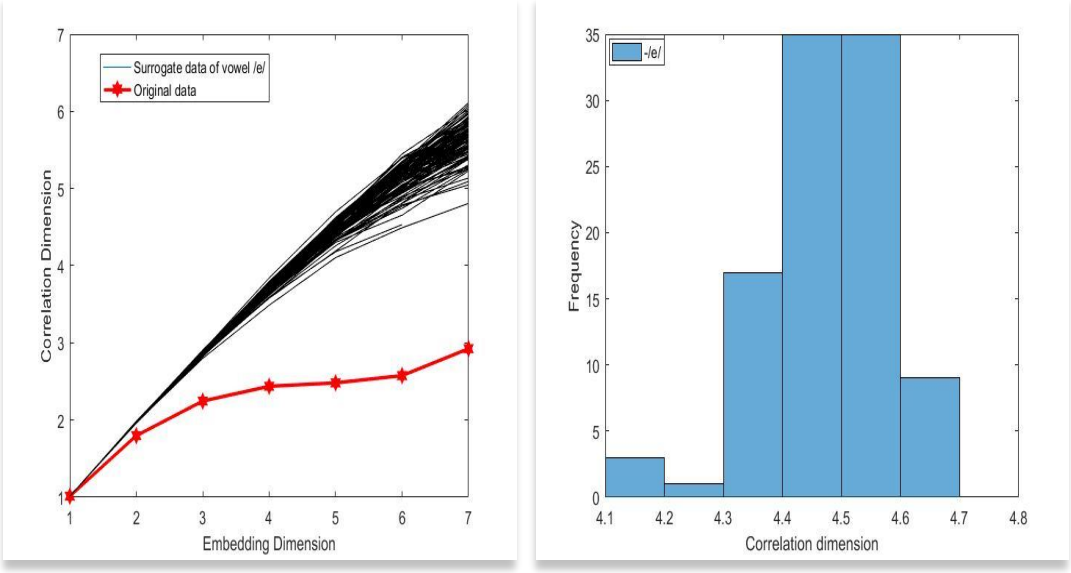


Fig. 5.6 (a) Variation of $\mathbf{D_2}$ with ***m*** for original time series (Malayalam Vowel എ/e/) and 100 surrogates. (b) Histogram of D_{2m} for 100 surrogates

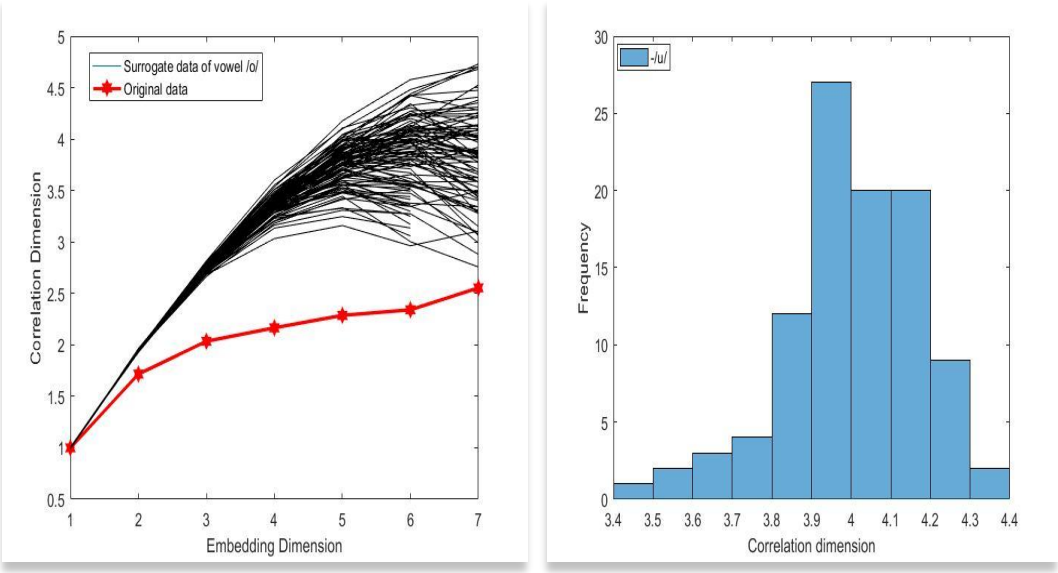


Fig. 5.7 (a) Variation of $\mathbf{D_2}$ with ***m*** for original time series (Malayalam Vowel ഓ/o/) and 100 surrogates. (b) Histogram of D_{2m} for 100 surrogates

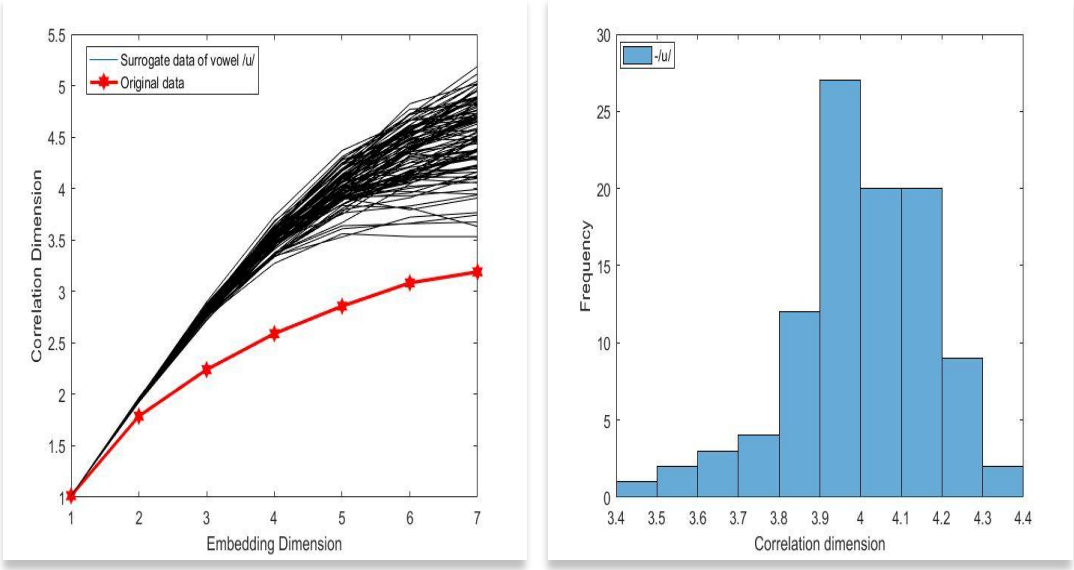


Fig. 5.8 (a) Variation of $\mathbf{D_2}$ with ***m*** for original time series (Malayalam Vowel ഉ/u/) and 100 surrogates. (b) Histogram of D_{2m} for 100 surrogates

Table.5.2 shows the statistical significance level(S) of Lorenz system,Rossler system and single female speaker vowel utterances.The range of D_{2m} values and significance level for 100 speakers (All analysed short vowel samples) are listed in Table 5.3. The range of D_{2m} values and significance level for 100 speakers (All analysed consonant samples) are listed in Table 5.4.

Table 5.2 D_{2m} Significance level (S) comparison of time series-Lorenz and Rossler system vs Single speaker utterance.

Signal	$\langle D_{2m} \rangle$	$\langle D_{2m} \rangle_{surr}$	σ_{surr}	Significance level (S)
Lorenz	1.97± 0.20	2.91±0.20	0.012	78.30 ±0.40
Rossler	1.95±0.20	2.70±0.20	0.010	75.00 ±0.40
അ/a/	2.63±0.20	4.48±0.20	0.058	31.90 ±0.40
ഇ/i/	2.96±0.20	4.63±0.20	0.127	13.20 ±0.40
ഏ/e/	2.48±0.20	4.47±0.20	0.106	18.80 ±0.40
ഒ/o/	2.29±0.20	3.75±0.20	0.202	7.20 ±0.40
ഉ/u/	2.86±0.20	4.00±0.20	0.175	6.50 ±0.40

Table 5.3 Range of $< D_{2m} >$, $< D_{2m} >_{surr}$ and significance level for 100 speakers (vowels)

Signal	$<D_{2m}>$		$<D_{2m}>_{surr}$		Significance level(S)	
	Min	Max	Min	Max	Min	Max
അ/**a**/	2.48±0.20	3.42±0.20	4.33±0.20	5.22±0.20	31.20±0.40	32.40±0.40
ഇ/**i**/	2.52±0.20	3.64±0.20	4.29±0.20	5.35±0.20	13.00±0.40	13.90±0.40
ഏ/**e**/	2.20±0.20	3.35±0.20	4.21±0.20	5.26±0.20	18.00±0.40	19.60±0.40
ഒ/**o**/	2.08±0.20	2.98±0.20	3.70±0.20	4.39±0.20	7.00±0.40	8.10±0.40
ഉ/**u**/	2.41±0.20	3.32±0.20	3.63±0.20	4.67±0.20	6.30±0.40	7.70±0.40

Table 5.4 Range of $< D_{2m} >$, $< D_{2m} >_{surr}$ and significance level for 100 speakers (consonants)

Signal	$<D_{2m}>$		$<D_{2m}>_{surr}$		Significance level(S)	
	Min	Max	Min	Max	Min	Max
പ/**P**/	2.18±0.20	3.52±0.20	4.13±0.20	4.82±0.20	41.31±0.40	43.40±0.40
വ/**v**/	2.42±0.20	3.32±0.20	4.29± 0.20	5.05±0.20	29.00±0.40	31.26±0.40
ത/**t**/	2.10±0.20	3.65±0.20	4.11±0.20	4.96±0.20	32.00±0.40	35.03±0.40
ർ/**r̲**/	2.30±0.20	3.18±0.20	3.80±0.20	4.19±0.20	25.00±0.40	26.10±0.40
ട/**ʈ**/	2.43±0.20	3.12±0.20	3.93±0.20	4.37±0.20	32.30±0.40	34.75 ±0.40
ച /**c**/	2.52±0.20	3.02±0.20	4.23±0.20	4.87±0.20	29.20±0.40	31.46±0.40
ക /**k**/	2.71±0.20	3.16±0.20	4.93±0.20	5.26±0.20	36.56±0.40	39.08±0.40
ഹ/**h**/	2.63±0.20	3.35±0.20	3.73±0.20	4.29±0.20	36.33±0.40	38.21±0.40

From Table 5.3 and 5.4 it is clear that the significance levels for different phonemes are different. This difference is the indication of the presence of time-variant nonlinearity in the speech production mechanism. The system has 'S' value comparable with Lorenz and Rossler systems. The significance level is high for consonants as compared to vowels. Even though the value of D_{2m} varies with the speaker ,as shown in Table 5.3 and 5.4, the significance level for a particular vowel and consonant seemed to be almost constant with little fluctuations. Almost the same 'S' value for a typical vowel for different speakers show some resemblance of different systems while uttering a particular vowel.

5.4.2 Results of Surrogate Analysis using K_{2m}

When K_{2m} is used as the nonlinear discriminating measure, the significance level will be written as

$$S = \frac{|K_{2m} - \langle K_{2m} \rangle_{surr}|}{\sigma_{surr}} \tag{5.19}$$

Where $\langle K_{2m} \rangle_{surr}$ is the average of the K_{2m} for surrogates and σ_{surr} is the standard deviation of K_{2m} distribution for surrogates. K_2 has been calculated for Malayalam vowels അ /a/, ഇ /i /, എ /e/, ഒ /o/ and ഉ /u/ and consonants പ/P/, വ/v, /ത/t/, റ/r̲/, ട/ṭ/, ച /c/, ക /k/, ഹ /h/ (for all speakers from the sample) at embedding dimensions from 1 to7.The average value of K_2 at the embedding dimension six (K_{2m}) for Lorenz system, Rossler system and Malayalam vowel utterance are tabulated in Table 5.4. The Figures 5.9 and 5.10, respectively for Lorenz system and Rossler system, show the variation of K_2 with embedding dimension with the histogram of K_{2m} for surrogates (for a specific speaker from the sample). The same for Malayalam vowelsഅ /a/, ഇ /i /, എ /e/, ഒ /o/ and ഉ /u/ respectively are shown in Fig 5.11 to 5.15.

Table 5.5 AverageK_{2m} values for Lorenz, Rossler and Malayalam vowels.

	Lorenz	**Rossler**	**അ/a/**	**ഇ/i/**	**എ/e/**	**ഒ/o/**	**ഉ/u/**
Original data	0.914	0.392	0.157	0.358	0.110	0.175	0.289
Surrogate 1	1.651	1.061	0.848	0.681	0.878	0.695	0.734
2	1.650	1.063	0.973	0.621	0.828	0.709	0.654
3	1.643	1.069	0.936	0.665	0.960	0.833	0.613
4	1.644	1.073	0.912	0.615	0.831	0.777	0.731
5	1.589	1.048	0.912	0.614	0.819	0.713	0.523
6	1.644	1.043	0.892	0.636	0.745	0.632	0.568
7	1.663	1.062	0.929	0.634	0.877	0.741	0.691
8	1.646	1.012	0.875	0.686	0.871	0.848	0.603
9	1.597	1.062	0.904	0.569	0.804	0.703	0.654
10	1.648	1.043	0.921	0.521	0.918	0.769	0.668
11	1.640	1.043	0.922	0.720	0.924	0.923	0.595
12	1.643	1.030	0.933	0.750	0.821	0.624	0.742

	Lorenz	**Rossler**	**അ/a/**	**ഇ/i/**	**എ/e/**	**ഒ/o/**	**ഉ/u/**
13	1.651	1.038	0.933	0.749	0.856	0.918	0.626
14	1.646	1.053	0.857	0.575	0.767	0.834	0.540
15	1.646	1.075	0.834	0.732	0.872	0.862	0.479
16	1.607	1.033	0.864	0.651	0.866	0.811	0.801
17	1.609	1.054	0.822	0.602	0.911	0.854	0.619
18	1.650	1.066	0.751	0.679	0.951	0.740	0.652
19	1.655	1.062	0.877	0.689	0.813	0.807	0.690
20	1.653	1.046	0.902	0.703	0.808	0.778	0.698
21	1.655	1.034	0.848	0.658	0.822	0.836	0.567
22	1.663	1.028	0.924	0.696	0.832	0.844	0.601
23	1.596	1.015	0.836	0.598	0.948	0.781	0.611
24	1.653	1.044	0.844	0.621	0.814	0.742	0.536
25	1.653	1.078	0.913	0.690	0.893	0.847	0.540
26	1.647	1.056	0.919	0.684	0.791	0.781	0.599
27	1.612	1.038	0.824	0.675	0.865	0.795	0.586
28	1.597	1.080	0.913	0.689	0.845	0.924	0.566
29	1.650	1.084	0.895	0.592	0.865	0.797	0.565
30	1.612	1.037	0.867	0.787	0.784	0.827	0.712
31	1.640	1.048	0.921	0.709	0.896	0.832	0.626
32	1.643	1.031	0.900	0.798	0.860	0.665	0.619
33	1.641	1.074	0.923	0.722	0.918	0.814	0.680
34	1.650	1.085	0.821	0.796	0.905	0.775	0.734
35	1.655	1.040	0.843	0.644	0.899	0.862	0.467
36	1.657	1.052	0.830	0.617	0.849	0.783	0.675
37	1.655	1.063	0.918	0.609	0.770	0.794	0.792
38	1.657	1.064	0.887	0.730	0.852	0.785	0.661
39	1.655	1.057	0.906	0.671	0.856	0.832	0.545
40	1.605	1.073	0.858	0.760	0.795	0.841	0.599
41	1.646	1.066	0.884	0.711	0.728	0.870	0.604
42	1.636	1.027	0.914	0.710	0.923	0.744	0.706
43	1.654	1.053	0.853	0.757	0.832	0.801	0.598
44	1.648	1.037	0.870	0.670	0.860	0.849	0.689
45	1.649	1.052	0.877	0.581	0.893	0.782	0.736
46	1.652	1.057	0.881	0.708	0.923	0.862	0.732
47	1.648	1.046	0.919	0.728	0.835	0.770	0.520
48	1.660	1.069	0.846	0.591	0.860	0.921	0.698
49	1.657	1.067	0.876	0.628	0.784	0.691	0.736
50	1.615	1.046	0.958	0.647	0.924	0.858	0.659
51	1.664	1.082	0.905	0.508	0.868	0.740	0.675
52	1.597	1.047	0.849	0.670	0.826	0.694	0.749
53	1.649	1.051	0.907	0.742	0.892	0.741	0.624
54	1.609	1.065	0.849	0.684	0.946	0.768	0.682
55	1.646	1.061	0.819	0.625	0.810	0.735	0.745
56	1.648	1.048	0.890	0.643	0.718	0.791	0.536
57	1.612	1.082	0.854	0.579	0.884	0.790	0.673
58	1.655	1.036	0.913	0.651	0.783	0.810	0.648

	Lorenz	**Rossler**	**അ/a/**	**ഇ/i/**	**എ/e/**	**ഒ/o/**	**ഉ/u/**
59	1.661	1.067	0.875	0.509	0.867	0.722	0.698
60	1.642	1.073	0.908	0.648	0.842	0.944	0.665
61	1.657	1.055	0.894	0.710	0.912	0.784	0.696
62	1.645	1.059	0.844	0.699	0.811	0.799	0.658
63	1.655	1.065	0.898	0.642	0.903	0.742	0.647
64	1.620	1.050	0.871	0.614	0.673	0.707	0.751
65	1.606	1.049	0.905	0.611	0.885	0.697	0.766
66	1.651	1.098	0.889	0.694	0.892	0.772	0.701
67	1.643	1.037	0.887	0.695	0.823	0.811	0.549
68	1.603	1.056	0.876	0.722	0.861	0.751	0.590
69	1.665	1.057	0.895	0.594	0.845	0.619	0.516
70	1.612	1.070	0.872	0.679	0.886	0.862	0.490
71	1.658	1.030	0.889	0.684	0.913	0.828	0.610
72	1.643	1.060	0.846	0.713	0.925	0.896	0.777
73	1.663	1.057	0.873	0.693	0.949	0.858	0.789
74	1.641	1.056	0.870	0.568	0.941	0.897	0.718
75	1.652	1.061	0.843	0.698	0.904	0.751	0.546
76	1.642	1.043	0.901	0.541	0.905	0.800	0.664
77	1.603	1.059	0.847	0.706	0.909	0.827	0.714
78	1.659	1.048	0.860	0.680	0.940	0.826	0.613
79	1.654	1.055	0.851	0.715	0.789	0.863	0.664
80	1.660	1.055	0.894	0.712	0.973	0.660	0.683
81	1.645	1.073	0.864	0.648	0.714	0.680	0.515
82	1.619	1.043	0.883	0.586	0.861	0.817	0.605
83	1.620	1.053	0.927	0.713	0.904	0.826	0.610
84	1.649	1.058	0.900	0.661	0.852	0.699	0.646
85	1.607	1.069	0.865	0.644	0.919	0.829	0.721
86	1.653	1.067	0.881	0.680	0.867	0.774	0.679
87	1.609	1.061	0.881	0.632	0.775	0.627	0.716
88	1.614	1.062	0.972	0.632	0.905	0.807	0.574
89	1.630	1.034	0.909	0.653	0.874	0.822	0.774
90	1.646	1.082	0.820	0.705	0.867	0.823	0.721
91	1.646	1.078	0.910	0.649	0.734	0.788	0.600
92	1.649	1.035	0.815	0.769	0.935	0.843	0.579
93	1.605	1.103	0.899	0.665	0.976	0.871	0.576
94	1.649	1.054	0.853	0.653	0.890	0.657	0.566
95	1.657	1.012	0.920	0.664	0.902	0.857	0.596
96	1.650	1.066	0.822	0.654	0.673	0.694	0.567
97	1.650	1.038	0.891	0.653	0.810	0.737	0.626
98	1.591	1.052	0.893	0.610	0.824	0.870	0.599
99	1.664	1.068	0.805	0.656	0.804	0.769	0.736
100	1.601	1.067	0.836	0.637	0.850	0.858	0.665

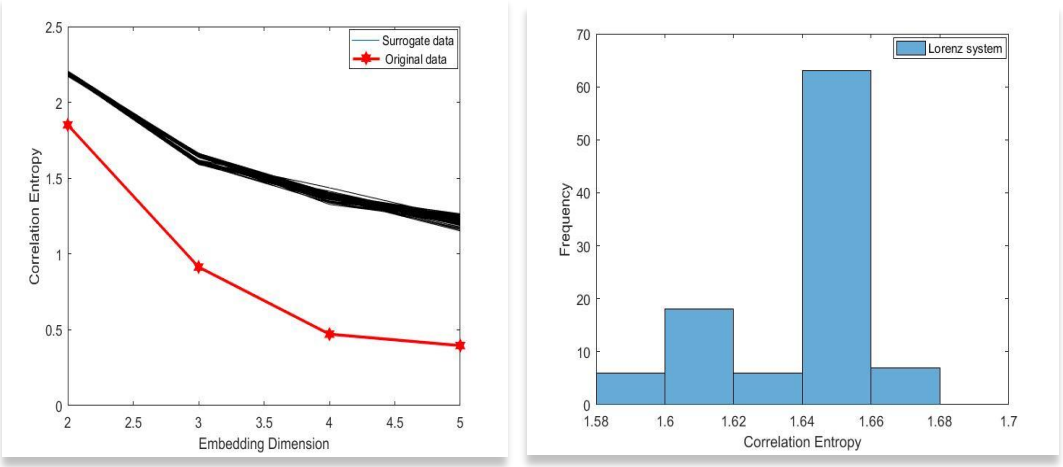


Fig.5.9 (a) Variation of $\mathbf{K_2}$ with $\boldsymbol{m}$ for original time series and 100 surrogates for Lorenz system. (b) Histogram of K_{2m} for 100 surrogates

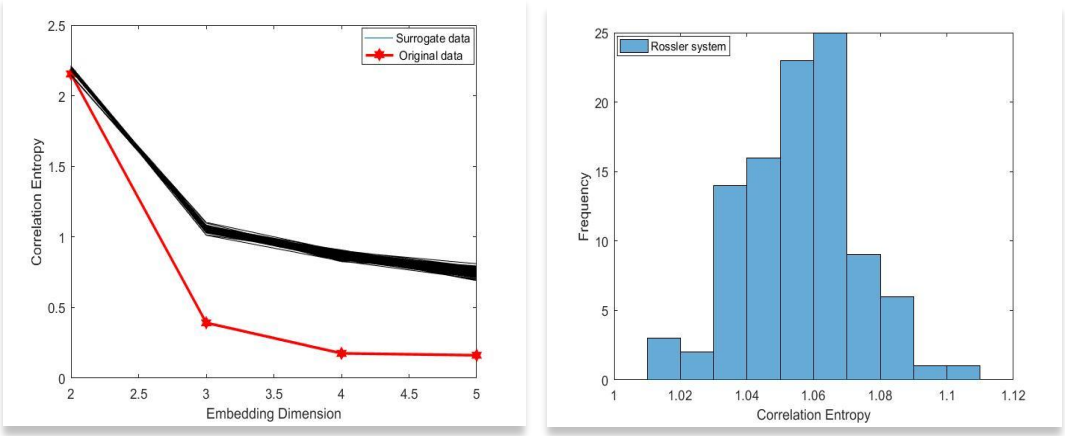


Fig. 5.10 (a) Variation of $\mathbf{K_2}$ with $\boldsymbol{m}$ for original time series and 100 surrogates for Rossler system. (b) Histogram of K_{2m} for 100 surrogates

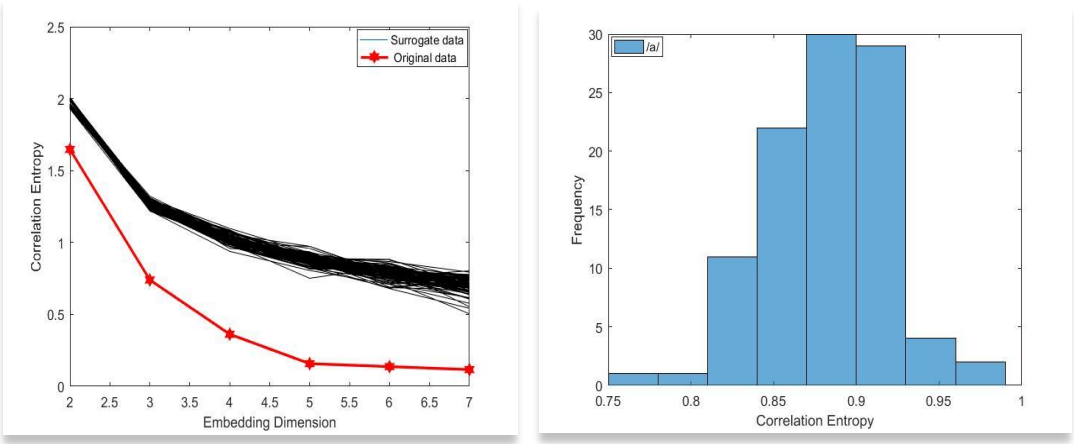


Fig. 5.11 (a) Variation of $\mathbf{K_2}$ with $\boldsymbol{m}$ for original time series and 100 surrogates for Malayalam vowel ൿ/a/. (b) Histogram of K_{2m} for 100 surrogates

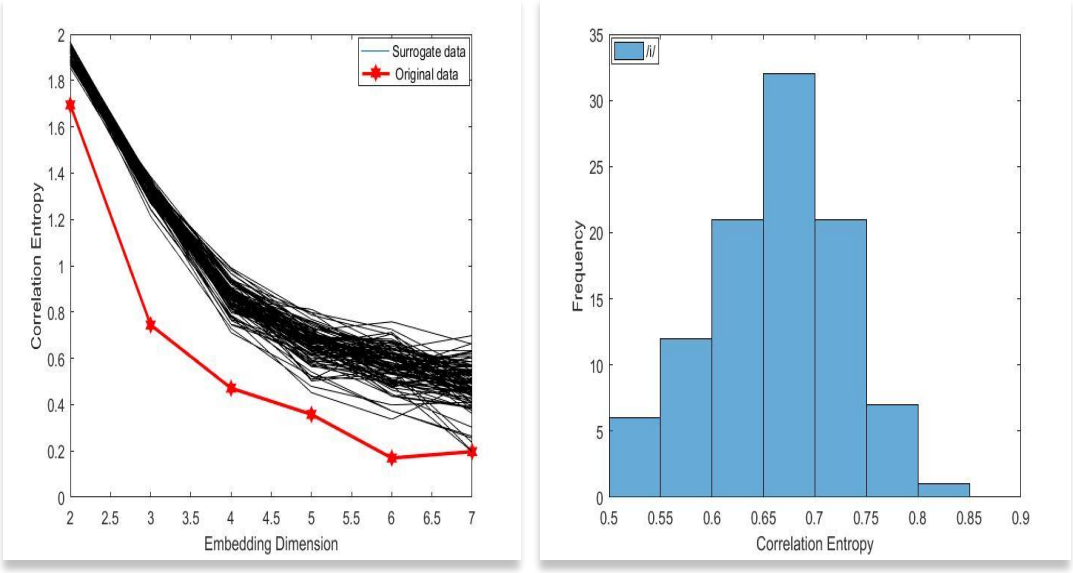


Fig. 5.12 (a) Variation of $\mathbf{K_2}$ with ***m*** for original time series and 100 surrogates for Malayalam vowel ഇ/i/. (b) Histogram of K_{2m} for 100 surrogates

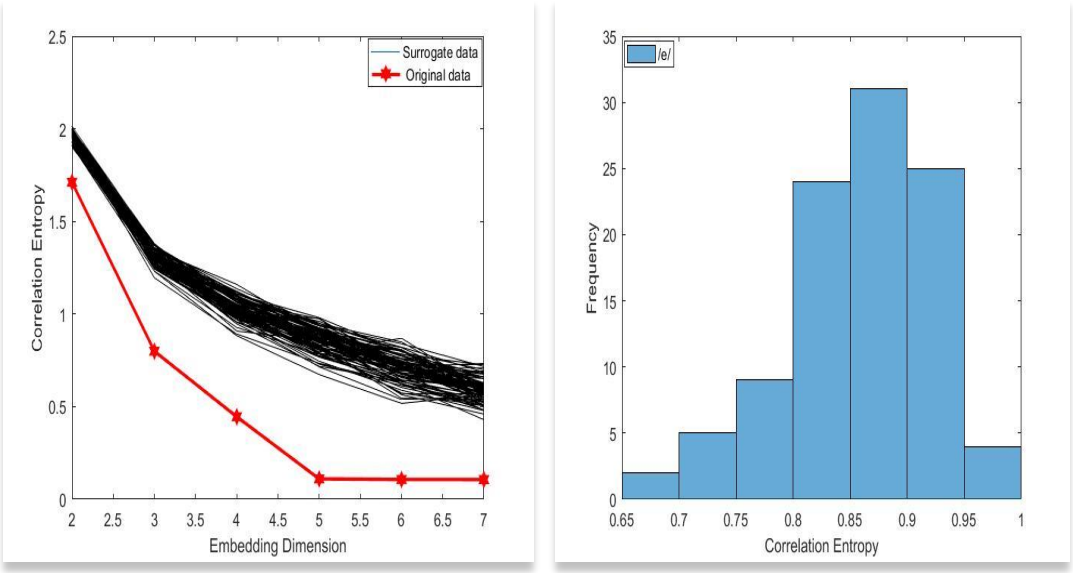


Fig. 5.13 (a) Variation of $\mathbf{K_2}$ with ***m*** for original time series and 100 surrogates for Malayalam vowel എ/e/. (b) Histogram of K_{2m} for 100 surrogates

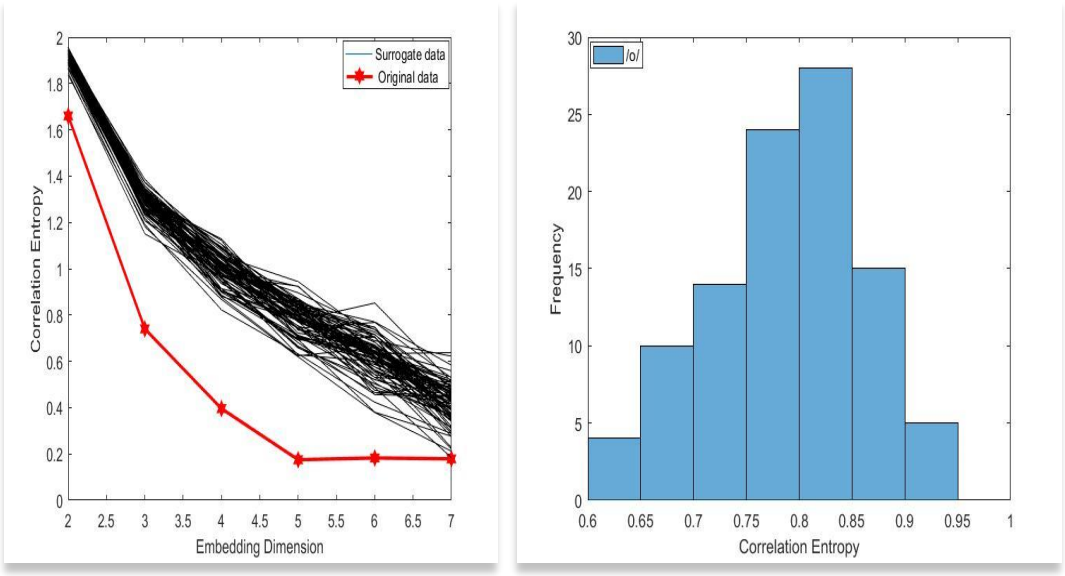


Fig. 5.14 (a) Variation of $\mathbf{K_2}$ with ***m*** for original time series and 100 surrogates for Malayalam vowel ഓ/o/. (b) Histogram of K_{2m} for 100 surrogates

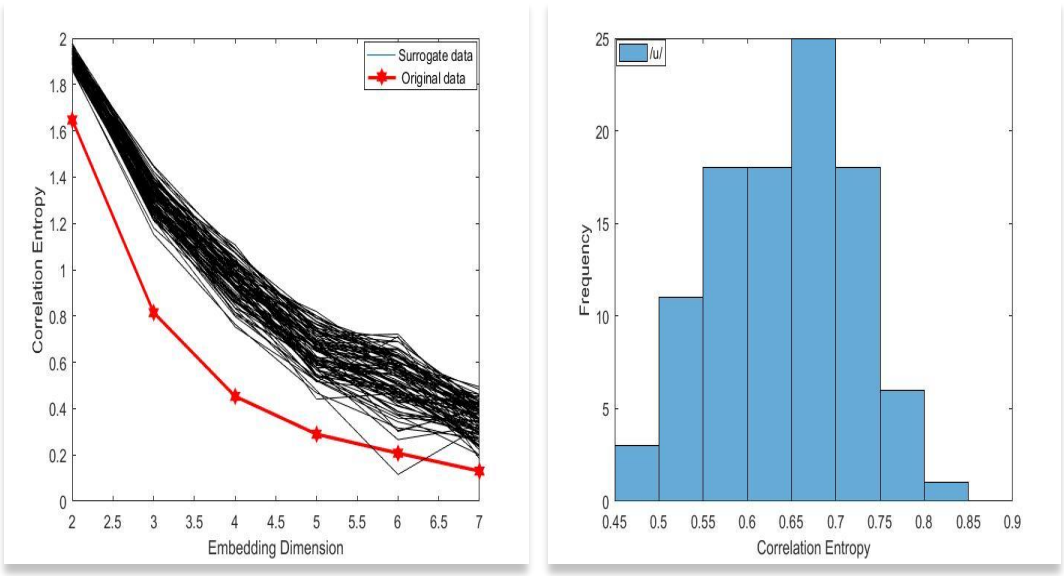


Fig. 5.15 (a) Variation of $\mathbf{K_2}$ with ***m*** for original time series and 100 surrogates for Malayalam vowel ഉ/u/. (b) Histogram of K_{2m} for 100 surrogates

Table.5.6 shows the statistical significance level(S) of Lorenz system,Rossler system and single female speaker utterances (K_{2m}). The range of K_{2m} values and significance level for 100 speakers are listed in Table 5.7 (All analysed vowel samples) and Table 5.8 (All analysed consonant samples).

Table 5.6 K_{2m} Significance level (S) comparison of time series-Lorenz and Rossler system vs Single speaker utterance.

System	$\langle K_{2m} \rangle$	$\langle K_{2m} \rangle_{surr}$	σ_{surr}	Significance level (S)
Lorenz	0.91±0.01	1.64±0.01	0.021	34.52±0.02
Rossler	0.39±0.01	1.06±0.01	0.017	39.00±0.02
അ/**a**/	0.16±0.01	0.88±0.01	0.038	19.05±0.02
ഇ/**i**/	0.36±0.01	0.66±0.01	0.059	5.18±0.02
ഏ/**e**/	0.11±0.01	0.86±0.01	0.063	11.85±0.02
ഒ/**o**/	0.18±0.01	0.79±0.01	0.072	8.54±0.02
ഉ/u/	0.29±0.01	0.64±0.01	0.077	4.61±0.02

Table 5.7 Range of $\langle K_{2m} \rangle$,$\langle K_{2m} \rangle_{surr}$ and significance level for 100 speakers

System	$\langle K_{2m} \rangle$		$\langle K_{2m} \rangle_{surr}$		Significance level(S)	
	Min	Max	Min	Max	Min	Max
അ/**a**/	0.12±0.01	0.17±0.01	0.87±0.01	0.93±0.01	19.12 ±0.02	20.34 ±0.02
ഇ/**i**/	0.30±0.01	0.39±0.01	0.62±0.01	0.74±0.01	15.16 ±0.02	16.30 ±0.02
ഏ/**e**/	0.10±0.01	0.16±0.01	0.86±0.01	0.90±0.01	11.54 ±0.02	12.81 ±0.02
ഒ/**o**/	0.15±0.01	0.23±0.01	0.79±0.01	0.77±0.01	8.53 ± 0.02	9.26± 0.02
ഉ/**u**/	0.22±0.01	0.46±0.01	0.60±0.01	0.84±0.01	4.21 ± 0.02	4.95± 0.02

Table 5.8 Range of $< K_{2m} >$,$<K_{2m}>_{surr}$ and significance level for 100 speakers (consonants)

Signal	$<K_{2m}>$		$<K_{2m}>_{surr}$		Significance level(S)	
	Min	Max	Min	Max	Min	Max
പ/P/	0.13±0.01	0.16±0.01	0.88±0.01	0.90±0.01	39.20±0.02	42.30±0.02
വ/v/	0.24±0.01	0.26±0.01	0.77±0.01	0.82±0.01	28.51±0.02	32.60 ±0.02
ത/t/	0.33±0.01	0.37±0.01	0.97±0.01	0.99±0.01	40.02±0.02	42.36 ±0.02
റ/r̲/	0.22±0.01	0.28±0.01	0.97±0.01	1.02±0.01	32.66±0.02	34.87±0.02
ട/ʈ/	0.29±0.01	0.17±0.01	0.73±0.01	0.82±0.01	36.01±0.02	38.32±0.02
ച /c/	0.16±0.01	0.21±0.01	1.00±0.01	1.12±0.01	38.54±0.02	40.43±0.02
ക /k/	0.25±0.01	0.31±0.01	0.99±0.01	1.08±0.01	37.10 ±0.02	39.93±0.02
ഹ /h/	0.18±0.01	0.24±0.01	0.87±0.01	0.98±0.01	32.16 ±0.02	36.45±0.02

From Table.5.7 and 5.8, it is clear that the significance level for different phonemes are different. The speech production system behaves like a nonlinear system. It has an S value comparable to the Lorenz and Rossler systems. For utterances ഒ /o/, and ഉ /u/, the S value is comparatively low. For those utterances, the system has more resemblance to linear models. It is understood that the 'S' values fall within a minimal range for a particular phoneme, even though K_{2m} values vary with the speaker.

5.5 Conclusion

From the comparison of the statistical significance level of Malayalam phoneme time series with standard Lorenz and Rossler systems, the significance level of different phonemes was found to be different and comparable. The significance level for D_{2m} and K_{2m} analysis shows that the values are closer to those of standard systems for vowels/a/,/i/,/e/ and all the analysed syllables. The levels for vowels ഒ /o/, and ഉ /u/ are comparatively smaller and it is better to avoid these sounds in the nonlinearity studies of the system. It can be concluded that inherent nonlinearity exists in speech

production, and the system is time-variant. The amount of nonlinearity is different while uttering different phonemes, and the system's nonlinearity is greater while uttering syllables. Thus, D_{2m} and K_{2m} can be used as better tools for the study of nonlinear dynamical structures in the speech production system, emotion recognition, and pathological analysis in place of saturated values of D_2 and K_2. In the next three chapters, these nonlinear features are used for analysing pathological, noisy, and emotional speech signals.

CHAPTER 6

ANALYSIS OF PATHOLOGICAL VOICES USING NONLINEAR FEATURES

6.1 Introduction

To assess the pathologies of the speech production system, an immediate inspection of the vocal folds are required [72].To analyse pathological voice signals, digital signal processing techniques have been used as an additional tool. The use of linear and nonlinear measures obtained from the recorded voice signals helps quantify the voice pathology and document the patient evolution. The traditional techniques for the discrimination can be improvised by the use of these automatic and quick techniques [143].

Objective measures of audio quality have been studied in the past decades. Fundamental frequency [144], amplitude perturbation, pitch perturbation and many other parameters [145], [146], [147] are used as objective measures in the existing literature. These studies report efficiency percentage from 75% to 95% accuracy for different pathological subjects with the combined use of feature vectors [47], [148]–[150]. Since each study has been evaluated with various databases, the comparison of the results is laborious.

To extract information about the speech production system from the speech signal, nonlinear time series analysis has been adopted recently. The common nonlinear characteristics studied include Lyapunov exponents (LE) [124], Correlation dimension (D_2) [72], [132], [143] and Correlation entropy (K_2) [143]. Recently nonlinear dynamical analysis methods have been popularly used in the investigation of normal and diseased vocal tract systems [72], [150]–[153]. The LLE, D_2 and K_2 have been shown to discriminate

normal subjects from pathological voices. Even D_2 and K_2 at minimum embedding dimension [52] (D_{2m} & K_{2m}) and the various combinations of these parameters have been proved as valid tools for the discrimination of pathological voices from normal voices. But the reported efficiencies are not comparable.

This chapter attempts to utilize the parameters of the singularity spectrum ($f(\alpha)$) of strange attractors in the phase space along with D_{2m} and K_{2m} for the discrimination of signals. The four nonlinear parameters which realize the $f(\alpha)$ spectrum γ_1, γ_2, α_{min} and α_{max} together with D_{2m} and K_{2m} are used for the analysis of speech samples. The said features of 50 normal samples and 50 pathological samples with Hyperkinetic dysphonia (from the VOICED database) [154] are examined. With the help of Iterated Amplitude Adjusted Fourier Transform (IAAFT) surrogates generated from the TISEAN package [130], surrogate analysis of datasets has been done, and the significance level was found to be high enough to ensure the nonlinearity of the data. For calculating D_{2m}, K_{2m} and $f(\alpha)$ spectrum, the modified algorithm suggested by Harikrishnan et al. [138], [142], [155] is utilised. The outcomes of the Support Vector Machine (SVM) classifier is also incorporated in work.

The chapter is arranged as follows. Section 6.2 discuss the peculiarities of the database used for research. Section 6.3 addresses the nonlinear parameterisation in which the nonlinear feature vectors and the surrogate analysis is discussed. The Support Vector Machine (SVM) classifier is discussed in section 6.4. The results and its discussion are done in section 6.5. The paper is concluded in Section 6.6, and future directions are mentioned.

6.2 Database used

Databases may be required for research activities, particularly those related to the development of smart healthcare systems. Three critical features must be examined to produce high-quality research results. To begin, there is

the database quality, which ensures that the findings are accurate and generally applicable. Second, the database size, which ensures that enough data is available to train and test tools. Finally, data availability is critical because it allows research to begin while also helping to improve the state of the art.

The VoiceICarfEDerico (VOICED) database, containing 208 healthy and pathological voices, is used in this study. VOICED is accessible on https://physionet.org/ and is freely available. The recordings contain the vowel /a/ of five seconds in length, as required by the clinical protocol of the diagnosis of the main voice pathologies. It also includes information regarding medical diagnosis of speakers. Since the database provides information regarding patient's life habits and previous diseases in connection with vocal disorders, the pathological samples can be easily identified [154].

The database includes recordings of people with three different types of pathologies. (1) Hyperkinetic dysphonia: This is a common clinical pathology characterised by muscular hyper contraction of the pneumo-phonic apparatus. The high glottic resilience to the exhalation of air stream makes phonation more exhausting and causes respiratory dynamics to change. Vocal fold nodules, Reinke's edima, rigid vocal folds, polyps, and prolapse are some of the diseases that fall into this category. (2) Hypo kinetic dysphonia: This condition is marked by a decreased adduction of the vocal folds during the respiratory cycle, resulting in an airflow obstruction at the level of the larynx. Dysphonia of the chordal groove, adduction deficit, presbiphonia, and vocal fold paralysis are all examples of hypo kinetic dysphonia. (3) Reflux laryngitis: It is a laryngeal inflammation caused by stomach acid backing up into the oesophagus. Table 6.1 summarises the study population available in the VOICED database.

Table 6.1 Study population of VOICE database

Age Group	Normal voice		Hyperkinetic dysphonia		Hypokinetic dysphonia		Reflux laryngitis	
	Male	Female	Male	Female	Male	Female	Male	Female
18-34	7	22	7	10	2	9	2	1
35-49	8	9	7	16	2	10	8	9
≥50	6	6	9	21	5	13	10	9
Total	21	37	23	47	9	32	20	19

Fifty healthy voices and 50 pathological voices (20 male and 30 female) of individuals suffering from Hyperkinetic dysphonia is used in the study. The samples are of different age category, sex and the habitual behaviour of patients are not considered. [154].

6.3. Nonlinear Parameterisation

The reconstruction of phase space is the foundation of the nonlinear parameterization of the system. The topological structures of the system's dynamics can be extracted from the FNN-generated hypothetical phase space. To ensure nonlinearity, surrogate analysis of both normal and pathological voices was performed with D_2 and K_2. The analysis is based on the nonlinear features D_2, K_2 and parameters of the $f(\alpha)$ spectrum of strange attractor. D_2 and K_2 are already explained in sections 5.2.3 and 5.2.4.

6.3.1 Phase space reconstruction

The time delay is calculated using the Mutual information method, and the embedding dimension is optimised using FNN. Figures 6.1 and 6.2 show the variation of MI with time delay for healthy and pathological voices, respectively. In Fig 6.3 and Fig 6.4, the delay obtained is used in FNN, and the variation of FNN with embedding dimension for normal and pathological samples is shown.

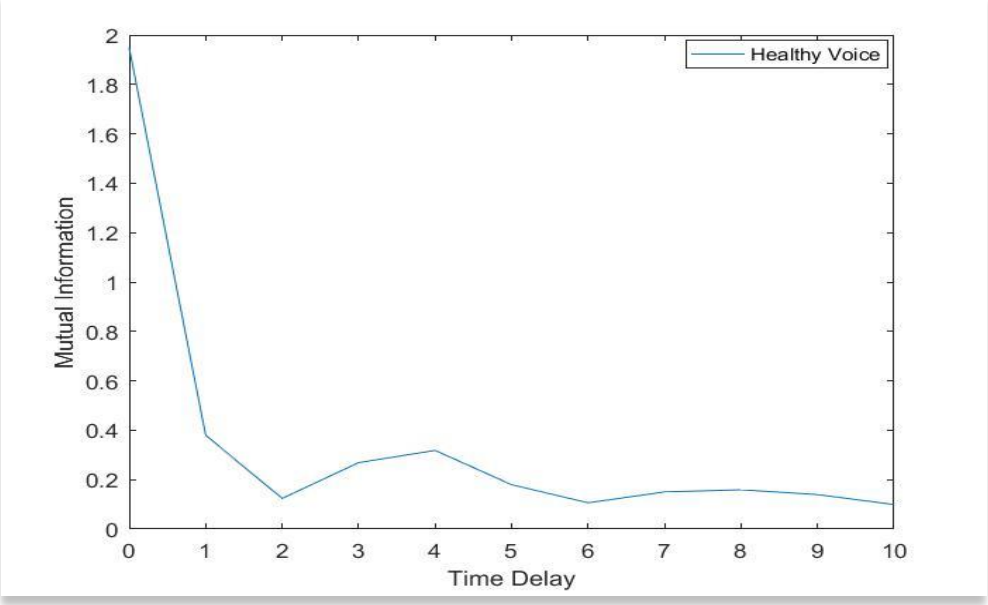


Fig. 6.1 Embedding delay of healthy voice signal

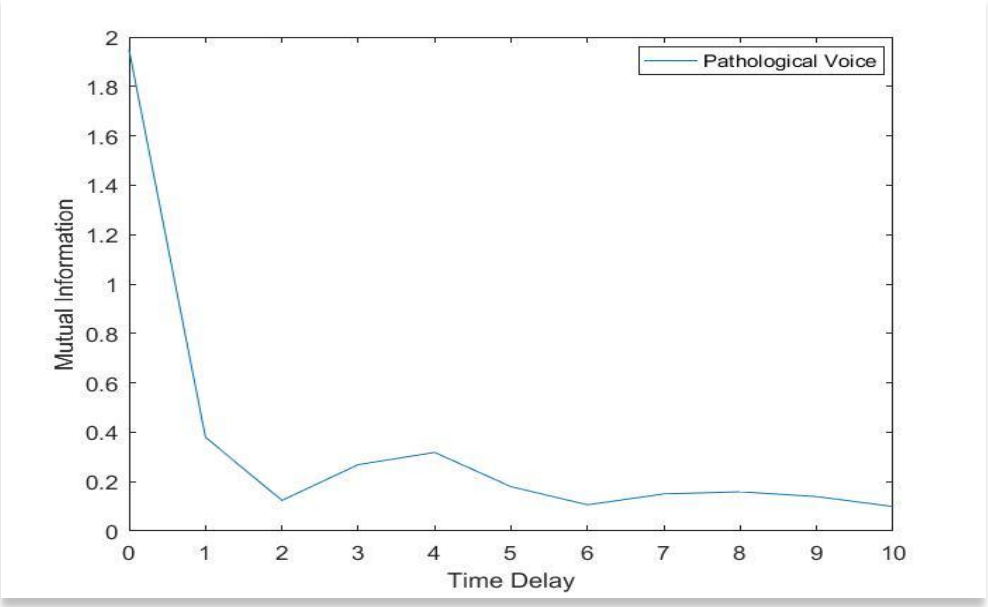


Fig. 6.2 Embedding delay of pathological voice signal

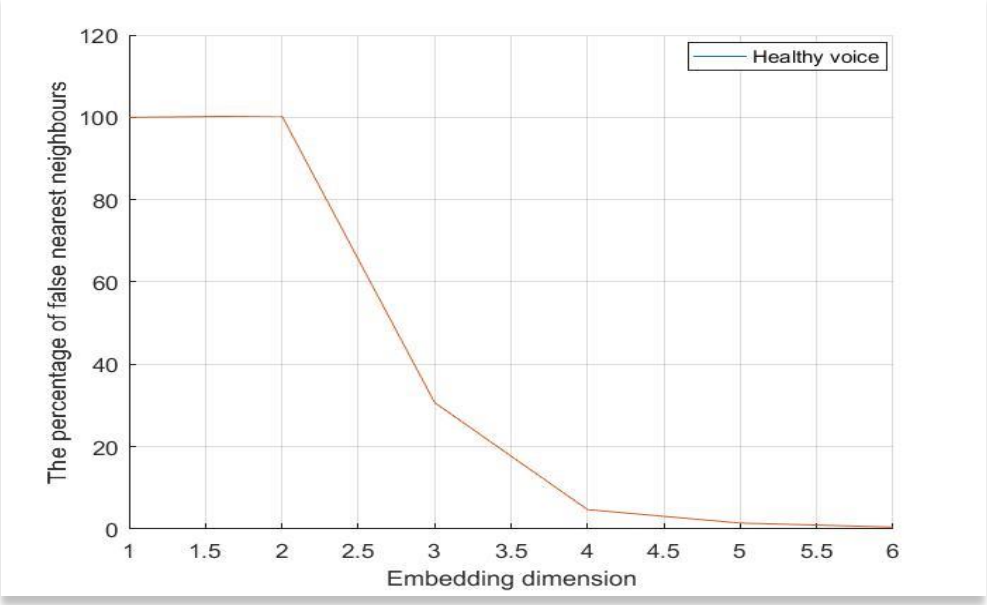


Fig. 6.3 Embedding dimension of healthy voice signal

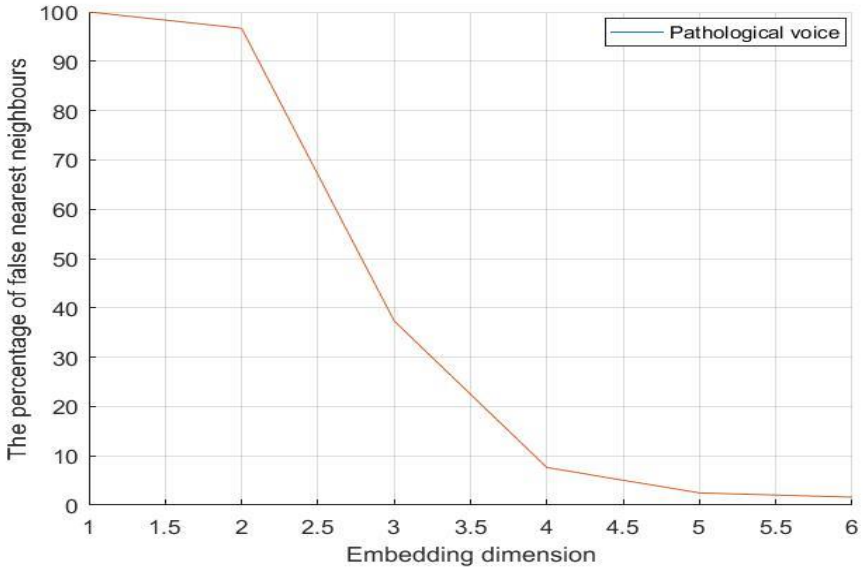


Fig. 6.4 Embedding dimension of pathological voice signal

From Fig 6.3 and 6.4 a six-dimensional hyperspace is enough to discuss the signal characteristics in phase space.

6.3.2 Surrogate analysis

To describe the behaviour of the inherent dynamics of a system the estimation of a couple of nonlinear measures of the time series is required. Before using these measures for system description, the reliability of these measures should be verified by surrogate data analysis as discussed in chapter 5 [135]. While working with a limited data base, like VOICED, a particular hypothesis could be tested using statistical methods like surrogate data to assure the reliability of the outcome. The surrogate analysis has been performed on the healthy and pathological voice signals by taking D_{2m} and K_{2m} as nonlinear discriminating measures by constructing 100 IAAFT surrogates. The embedding dimension is taken as six. Fig 6.5 shows the surrogate analysis of normal signal with 10 surrogates of D_{2m} and K_{2m}. The same for pathological signal is given in Fig 6.6.

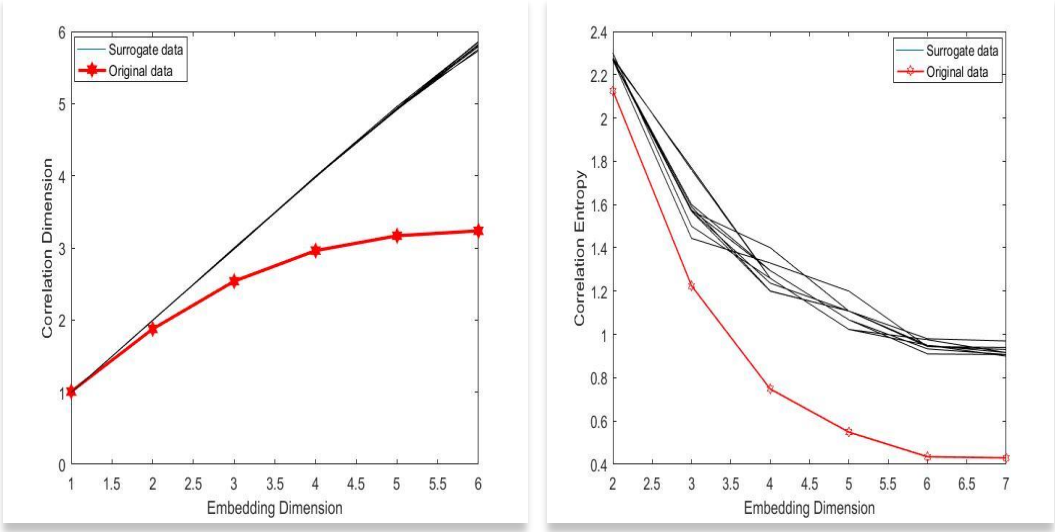


Fig. 6.5 Surrogate analysis of healthy signal

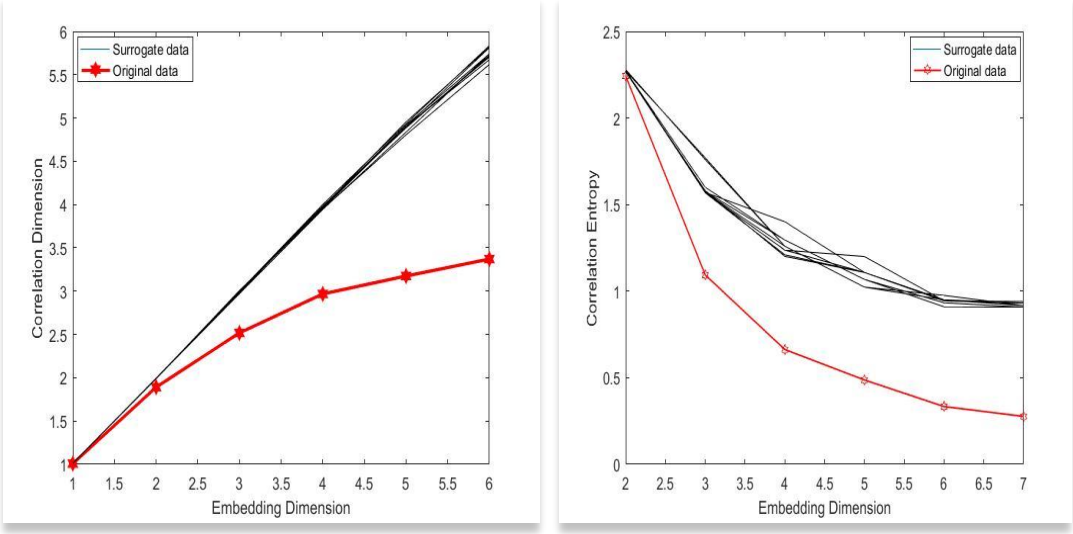


Fig. 6.6 Surrogate analysis of pathological signal

Table 6.2 shows the statistical significance level(S) of healthy and pathological signals obtained from the surrogate analysis. D_{2m} and K_{2m} in the six-dimensional hyperspace are used as nonlinear discriminating measures. It is clear from Table 6.2 that the 'S' value is enough to prove the nonlinear structure in the data.

Table 6.2 Significance level for healthy and pathological voices

Signal	$\langle D_{2m} \rangle$	$\langle D_{2m} \rangle_{surr}$	σ_{surr}	S	$\langle K_{2m} \rangle$	$\langle K_{2m} \rangle_{surr}$	σ_{surr}	S
Healthy	3.51	4.53	0.045	22.6	0.62	0.21	0.019	21.5
Pathological	3.66	4.86	0.047	25.53	0.75	0.30	0.017	26.47

6.3.3 f(α) spectrum

The strange attractors of the phase space of chaotic dynamical systems are characterized by the associated singularity spectrum f(α).The system bears a generalized dimension spectrum (D_q) from which f(α) can be generated. The precise mathematical description of the multifractal measure in the strange attractor is provided by the f(α) spectrum [155] calculated by the relation

$$f(\alpha) = q\alpha - (q-1)D_q \tag{6.1}$$

Where $D_q$is the generalized dimension which can be determined from the generalized correlation sum $C_d(R)$ with q taking values 0,1,2,3,.....

$$D_q = \frac{1}{q-1}\lim_{R\to 0}\frac{\log C_q(R)}{\log R} \tag{6.2}$$

$$C_q(R) = \frac{1}{N_c}\sum_i^{N_\vartheta}\left\{\lim_{N\to\alpha}\frac{1}{N_\vartheta}\sum_{\substack{j=1,\\ j\neq i}}^{N_\vartheta} H\left(R - \left|\overrightarrow{x_i} - \overrightarrow{x_j}\right|\right)\right\}^{q-1} \tag{6.3}$$

Since f(α) has a single maximum and it falls to zero for two value of α,namely α_{min} and $\alpha_{max,}$ it can be expressed as a polynomial in α as

$$f(\alpha) = A(\alpha - \alpha_{min})^{\gamma_1}(\alpha_{max} - \alpha)^{\gamma_2} \tag{6.4}$$

Where A, α_{min}, α_{max} γ_1 and γ_2, are a set of features realizing a particular f(α) graph of the given sample. Since these parameters are mutually connected, four of them can be taken as independent. Hence α_{min}, $\alpha_{max,}$ γ_1 and γ_2 are taken as parameters in this work for discriminating healthy and pathological signals.

6.4 SVM Classifier

SVM is a supervised machine learning algorithm that has been applied to a variety of real-world problems, particularly classification. Pattern recognition research began in 1936 with the work of R. A. Fisher, who proposed the first pattern recognition algorithm. SVM is based on Vapnik's 1974 statistical learning theory and quadratic programming. It's a nonlinear version of the Generalized Portrait algorithm, which was first developed in 1963. Boser et al. first proposed the current form of SVM in 1992. SVM has been modified in various ways, but it remains an active algorithm with relatively simple concepts. SVM was built from the ground up to be a binary nonlinear classifier that could determine whether the input data belonged in class 1 or class 2. In a higher-dimensional feature space, it defines an

optimum hyperplane that best separates different classes with a maximum margin between the boundary points (support vectors).

6.4.1 Binary classifier

The data set for a supervised classifier was divided into two parts: training data and testing data. Let (x_i, y_i) be the training data, with y_i being the target output for x_i and i=1, 2, 3, ..., n (number of observations). The goal is to use an imaginary surface to divide testing data x_i into two classes, y_i= +1 and y_i = -1, with the goal of maximising the separation between boundary points and minimising error. After training, the unlabelled data (testing data set) is used to evaluate the classifier's performance. The decision boundary is the imaginary surface that SVM uses to perform separation. Depending on the nature of the problem, the decision boundary has different sizes and shapes. It will be a line in two dimensions, a plane in three dimensions, and a hyperplane in N dimensions for the separable case. SVM employs a kernel trick to transform n-dimensional sequence of feature vectors into a linearly separable higher-dimensional kernel feature space in non-separable cases [156].

For a d-dimensional space, the decision boundary is a hyperplane specified by

$$\sum_{i=1}^{d} w_i x_i + w_0 = 0 \tag{6.5}$$

Thus, the hyperplane can be described as follows

$$W^T x + w_0 = 0 \tag{6.6}$$

where $W = [w_1, w_2, \ldots, w_d]$ and $x = [x_1, x_2, \ldots\ldots, x_d]$

Thus, the decision function of a hyperplane for an input x_i is represented as

$$g(x_i) = W^T x_i + w_0 \tag{6.7}$$

For any points that lies on the hyperplane, then $g(x_i) = 0$. Let the region above the hyperplane be class 1, then $g(x_i)$ will be positive and the region below the hyperplane be class 2, then $g(x_i)$ will be negative as shown in Fig. 6.8.

$$W^T x_i + w_0 > 0 \text{ for } y_i = +1$$

$$W^T x_i + w_0 < 0 \text{ for } y_i = -1$$

$$y_i(W^T x_i + w_0) = \begin{cases} > 0 \; if \; correct \\ < 0 \; if \; incorrect \end{cases} \tag{6.8}$$

The product of predicted and class label would be greater than zero on correct prediction, otherwise less than zero

Any classifier's goal is to reduce the number of misclassifications in the training set, which is known as empirical risk minimisation (ERM). A generalised model must be chosen from a finite data set in machine learning, which leads to the problem of overfitting, or when the model becomes overly tuned to the characteristics of the training set and thus fails to generalise to new data. The Structural Risk Minimisation (SRM) principle is used to solve this problem, which balances the model's complex nature against its ability to match the training data. [157], [158] The optimal hyperplane maximises the margin while minimising the empirical risk. There will be hyperplanes that can accurately categorise all data points, as shown in Fig. 6.7, resulting in zero empirical risk. H1 is preferred over H2 because it has a higher margin and is therefore less prone to overfitting. To put it another way, a linear SVM can be trained to learn a hyperplane that can tolerate a small number of non-separable data points.

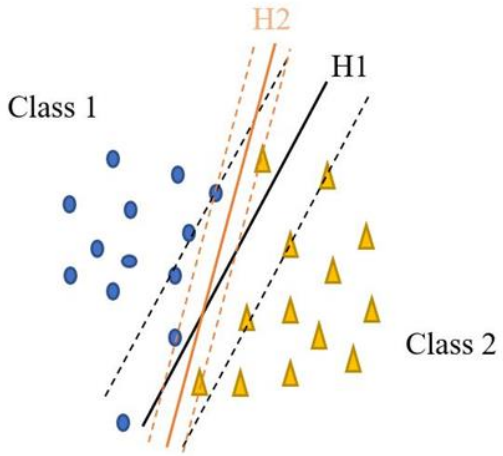


Fig. 6.7 Hyperplanes for Classifying the Non-separable Datapoints

The number of non-separable data items that SVM considers should be kept to a minimum. Many data points may be misclassified if the decision boundaries have a large margin of error. As a result, a trade-off between margin width and misclassification error should exist. A penalization term is added to the optimal condition to avoid this problem, as shown below.

$$M^* = argmin\frac{1}{2}||W||^2 + C\sum_{i=1}^{p}\xi_i$$

$$Subject\ to\ min\ y_i(W^T x_i + w_0) - 1 + \xi_i \geq 0 \qquad (6.9)$$

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

where W is a model parameter vector that defines the decision boundary, C is the penalty parameter that represents misclassification or error term, and ξ_i is the positive slack variables. The misclassification or error term tells the SVM optimization what level of error is acceptable. A smaller C value means a smaller margin, whereas a larger C value means a larger margin. Using an iterative search process, the parameter establishes an understanding between ERM and SRM (Grid search). Linear separating hyper plane for the non-separable data points is shown in Fig. 6.8. The number of support vectors

determines the classifier's complexity after the SVM has been trained with training data.

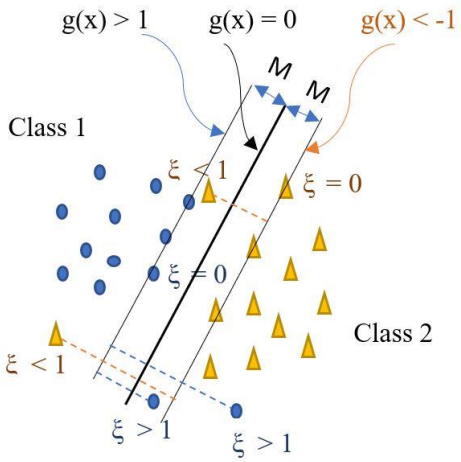


Fig. 6.8 Linear Separating Hyperplane for the Non-separable Datapoints

Because the data points in real-world classification problems are highly overlapping, the above-mentioned classifier does not produce accurate classification. In other words, the decision boundary could be a hypersurface rather than a linear or nonlinear hyperplane. To address this issue, the input data is transferred into a higher-dimensional space, as the data's dimension has no bearing on the classifier's success. Then look for a linear decision boundary to separate the transformed higher-dimensional data. The kernel trick is a revolutionary method for solving the above-mentioned problem. Instead of explicitly applying the transformations f(x) and expressing the data by these modified coordinates in the higher dimensional feature space, the kernel approach conveys the data only through pairwise similarity comparisons between the original data observations x. In simple terms, the kernel function takes lower-dimensional inputs and returns the dot product of converted vectors in higher-dimensional space. When comparing two input

vectors, the dot product is frequently used. . Kernel function $K(x_i, x_j)$ is a real function defined on R such that there exist a function $\phi: R^m \rightarrow R^n$, where $n>m$

$$K(x_i, x_j) = \phi(x_i) \bullet \phi(x_j) \rightarrow x_i \bullet x_j$$

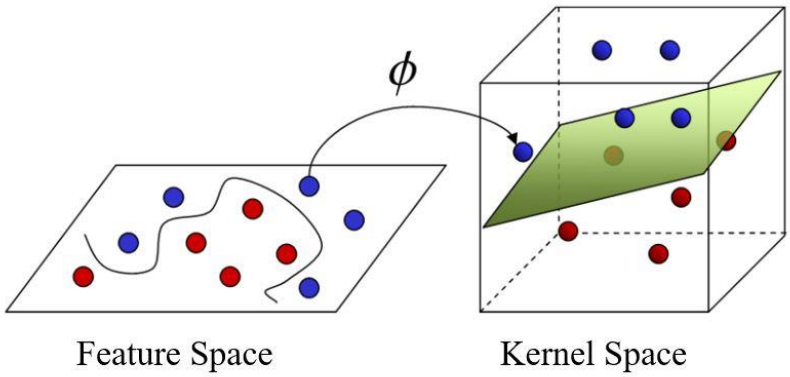


Fig. 6.9 Transformation of Non-Separable Data points in Feature Space to Separable Data points in Kernel Space

In kernel methods, the data set X is represented by a n x n kernel matrix of pairwise similarity comparisons, with the entries defined by the kernel function. Due to the commutative nature of the dot product, only half of the matrix elements should be computed. The kernel-based decision function in the kernel space has the form

$$g(x) = \sum_{i=0}^{N} \alpha_i . y_i . x_i^T x + w_0 = \sum_{i=0}^{N} \alpha_i y_i K(x_i, x) + w_0 \quad (6.10)$$

The kernel trick's ultimate benefit is that the objective function optimising to fit the higher dimensional decision boundary only includes the dot product of the transformed feature vectors, rather than explicitly mapping the data into these spaces. Transformation of non-separable data points in feature space to separable data points in kernel space is seen in Fig. 6.9. Linear, polynomial, Gaussian radial basic, and sigmoid are some of the most commonly used kernel functions.

Linear Kernel: $K_L(x_i, x_j) = x_i \bullet x_j$

Polynomial Kernel: $K_P(x_i, x_j) = (1 + x_i \bullet x_j)^d$

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

Sigmoid Kernel: $K_S(x_i, x_j) = tanh(ax_i \bullet x_j + b)$ (6.11)

The efficiency of the proposed system for classification can be indicated by accuracy and precision. Accuracy is a metric that generally describes how the model performs across all classes as given in Eq. 6.12. Precision attempts to answer how precise the model is, that is how many of them are actual positives out of those predicted positives (Eq. 6.13).

Table 6.3 Confusion Matrix

		Predicted Class	
		Healthy	Pathological
Actual Class	Pathological	True Positive (TP)	False Negative (FN)
	Healthy	False Positive (FP)	True Negative (TN)

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

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

6.4.2 Multi-class Problems

The SVM discussed so far deals with binary class classification, but multi-class classification is required in the coming chapters. A number of binary SVM classifiers are combined to solve the multi-class classification

problem. This can be accomplished using one of two methods: one-vs-all classification or one-vs-one classification.

Each class is compared against the remainder of the classes in a one-vs-all classification. The required number of binary classifiers is equal to the number of class labels in the data set. Data from one class is treated as positive in each binary classifier, while data from all other classes is treated as negative. Each model predicts a score that is similar to a probability score. The argmax (the class index with the highest score) of these scores is then used to predict a class. Fig. 6.10 shows the one-vs-all SVM Classifier.

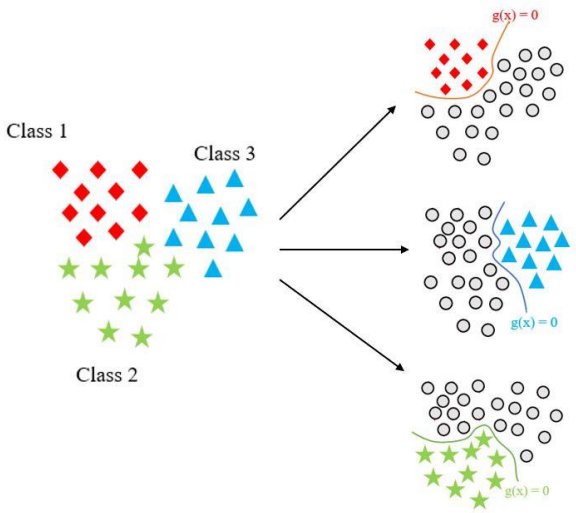


Fig. 6.10 One-vs-All SVM Classifier

Each class is confronted with the other classes independently in a one-vs-one categorization. In this approach, n (n-1)/2 classifier models are required, where n is the number of classes in the problem. To execute this method, the major datasets are separated into one binary classification dataset for each pair of classes. The one-vs.-one SVM Classifier is depicted in Fig. 6.11.

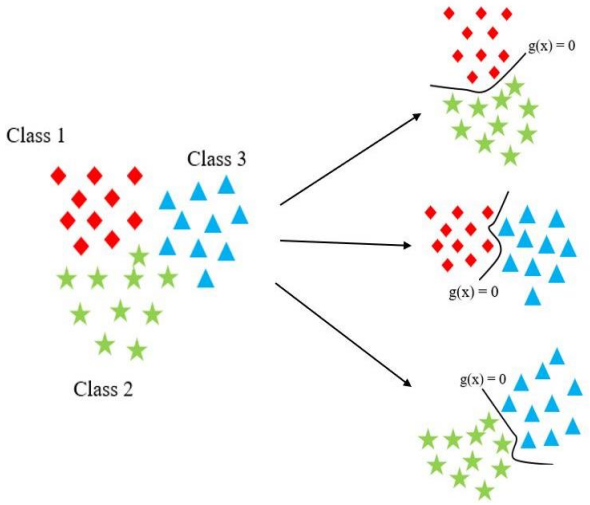


Fig. 6.11 One-vs-One SVM Classifier

6.5 Results and Discussion

The nonlinear feature extraction and the proposed classification system together with the results of SVM classifier are discussed in the following sections**.**

6.5.1 Nonlinear feature extraction

Fifty normal speech signals and fifty voice signals of patients with hyperkinetic dysphonia have been analysed and six nonlinear features, D_{2m}, K_{2m}, γ_1, γ_2, α_{min} and α_{max} were extracted in the experiment. The D_2 and K_2 values calculated at embedding dimensions from 1 to 6 with their error limits are represented in figures 6.12 and 6.14 for healthy voice and in figures 6.13 and 6.15 for pathological voice.

The average value of D_{2m} for healthy voice signal is 3.51 with a standard deviation 0.29 (Fig 6.12) and that of pathological voice is found to be 3.66 with standard deviation 0.30 (Fig 6.13).The average value of K_{2m} is 0.62 with standard deviation 0.15 for normal subjects (Fig 6.14) and 0.75 with standard deviation 0.14 for pathological subjects (Fig 6.15). Due to fluctuations in the value these two variables may not be enough to

discriminate the speech samples. Hence, the parameters of the f(α) spectrum are also combined with D_2 and K_2 to discriminate between two types of signals.

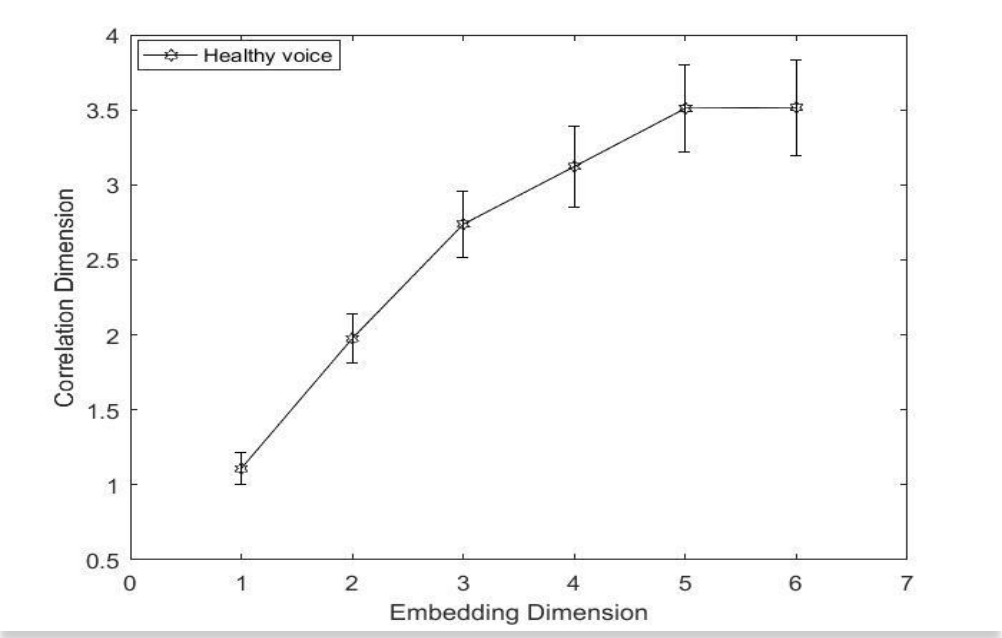


Fig. 6.12 D_2 of healthy voice at various embedding dimensions with error bar

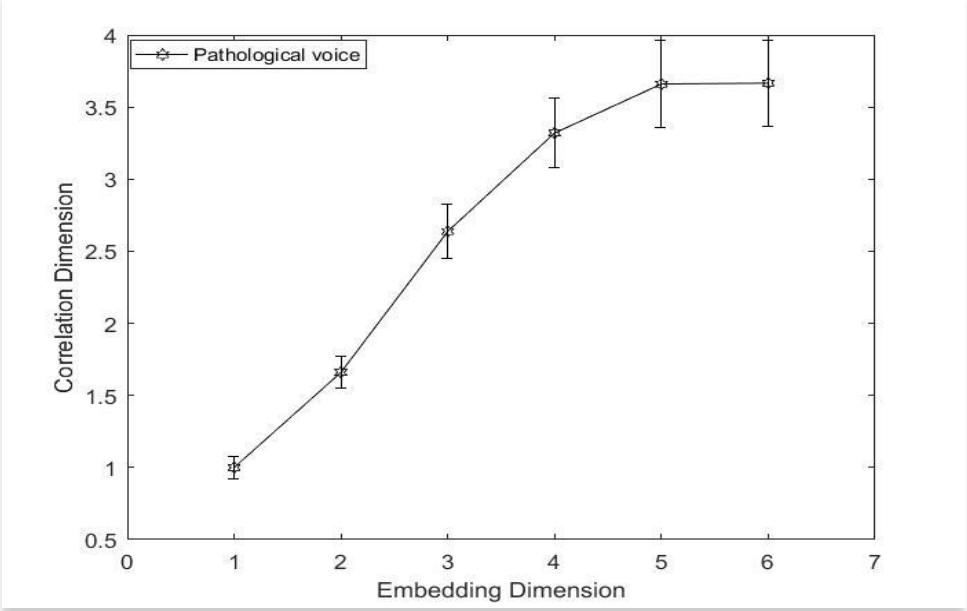


Fig. 6.13 D_2 of pathological voice at various embedding dimensions with error bar

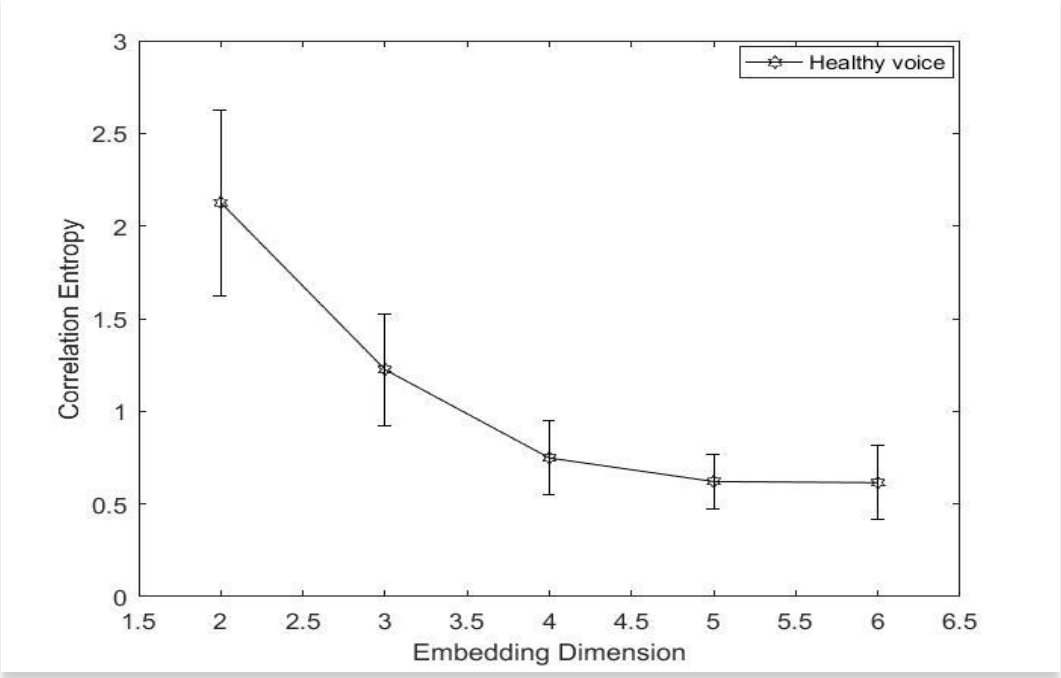


Fig. 6.14 K_2 of healthy voice at various embedding dimensions with error bar

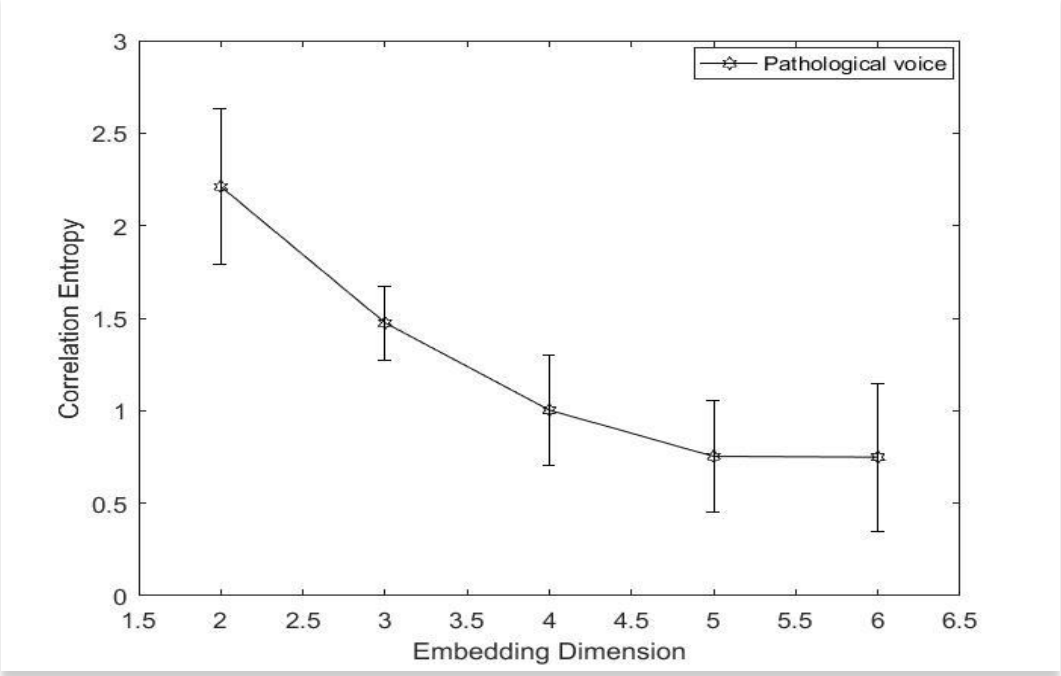


Fig. 6.15 K_2 of pathological voice at various embedding dimensions with error bar

The D_2 values show a decrease with increase in the embedding dimension and K_2 values show a decrease. At a particular dimension (optimal embedding dimension which is already optimised from FNN) these values get saturated and the values corresponding to the said dimension (D_{2m} and K_{2m}) is taken for the classification of samples. D_2 and K_2 values at minimum embedding dimension (six) is tabulated in Table 6.4 together with other discriminating measures, where, <x> denote the mean value and σ is the standard deviation.

The algorithmic approach proposed by Harikrishnan et al [155], which is an improved version of Grassberger et al [129], is used to determine the f(α) spectrum of both healthy and pathological signals. Figures 6.16 and 6.17 show the f(α) spectrum for both types of samples. The different coefficients involved in equation 6.4 determine the shape of the spectrum. The fitting coefficients γ_1, γ_2, α_{min} and α_{max} are derived from the f(α) spectrum and are listed in Table 6.4 along with D_2 and K_2.

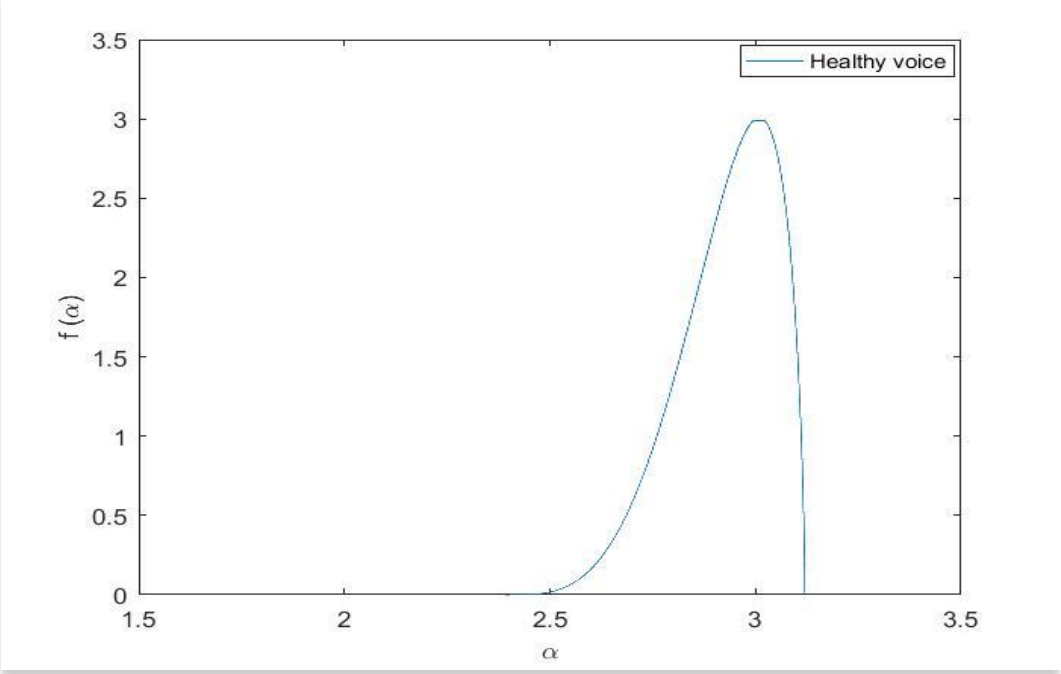


Fig. 6.16 f(α) spectrum of healthy voice

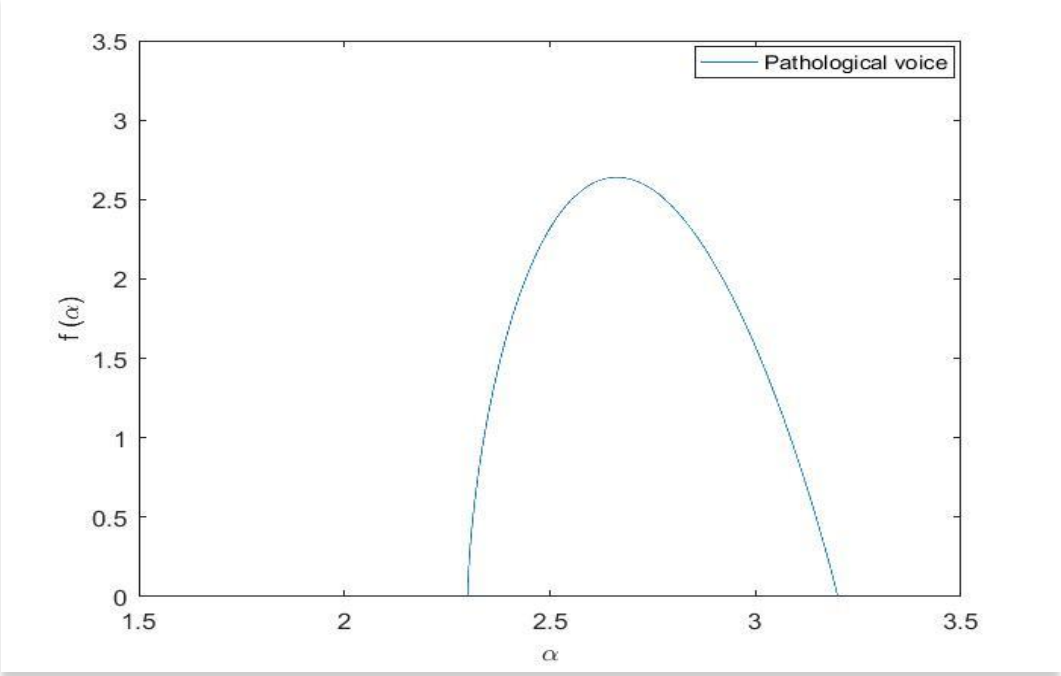


Fig. 6.17 f(α) spectrum of pathological voice

Table 6.4 Nonlinear feature vectors of pathological signal

Feature name	Healthy voice		Pathological voice	
	μ	σ	μ	σ
$\mathbf{D_{2m}}$	3.51± 0.01	0.29	3.66± 0.01	0.30
$\mathbf{K_{2m}}$	0.62± 0.01	0.15	0.75± 0.01	0.14
$\boldsymbol{\alpha_{min}}$	2.54± 0.01	0.24	2.64± 0.01	0.38
$\boldsymbol{\alpha_{max}}$	3.26± 0.01	0.19	3.52± 0.01	0.51
$\boldsymbol{\gamma_1}$	2.78± 0.01	0.68	1.55± 0.01	0.39
$\boldsymbol{\gamma_2}$	0.73± 0.01	0.07	0.79± 0.01	0.07

The proposed classification system for pathology detection from speech signals using the combined nonlinear and multifractal features is given in figure 6.18.

6.5.2 Results from SVM classifier

The six nonlinear parameters(D_{2m}, K_{2m}, $\gamma 1$, $\gamma 2$, α_{min} and α_{max}) extracted from 50 normal and 50 patients with hyperkinetic dysphonia is utilized for SVM classification. The extracted features are fed to the SVM classifier for training and testing purpose. Accuracy and precision of the classification is determined using different types of kernels. Individual classification accuracy of parameters is also tested along with the combined features. The confusion matrix for the combined nonlinear features with different types of kernels is shown in figure 6.19. Results of the proposed classification system are given in table 6.5.It is clear from Table 6.5 that the linear kernel gives better result than other kernels for the given data set. The combined use of the nonlinear measures gives a better accuracy than individual

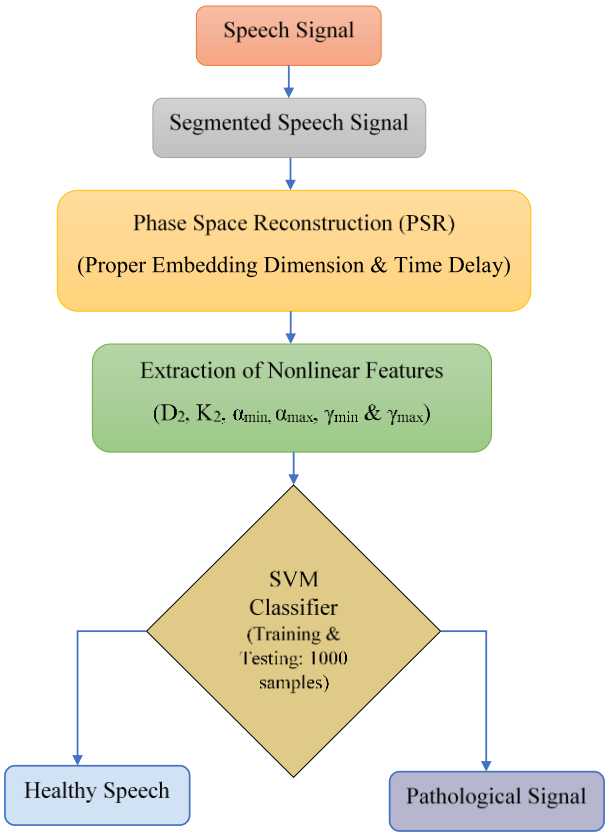


Fig. 6.18 Proposed classification system

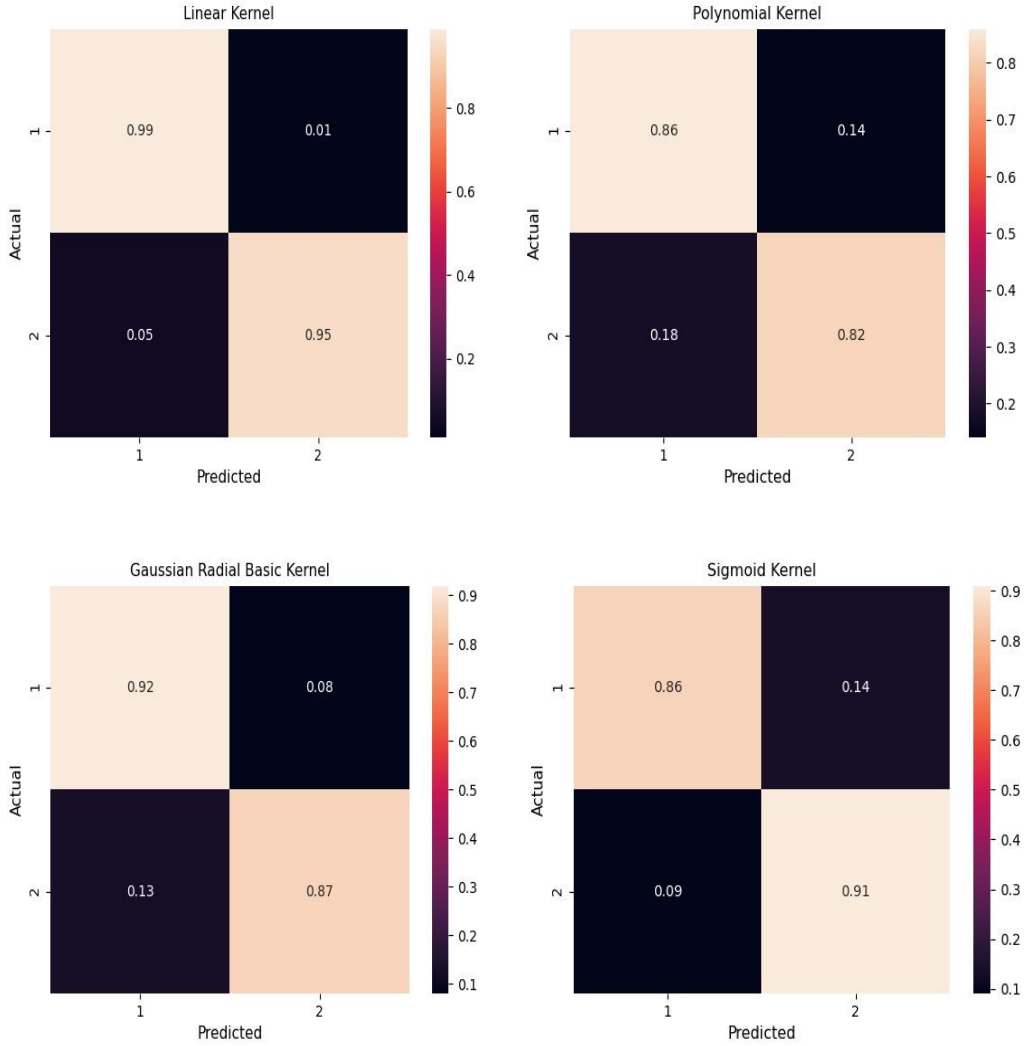


Fig. 6.19 Confusion matrix for different types of kernels

Table 6.5 Accuracy and precision of classification by SVM classifier

Feature	Accuracy %				Precision%			
	K1	K2	K3	K4	K1	K2	K3	K4
D_{2m}	65%	52%	58%	63%	68%	56%	59%	64%
K_{2m}	65%	55%	61%	64%	66%	60%	63%	64%
f(α)	74%	68%	75%	68%	71%	70%	76%	72%
D_{2m}&K_{2m}	82%	75%	81%	76%	80%	73%	81%	75%
D_{2m},K_{2m}&f(α)	97%	84%	90%	89%	99%	86%	92%	86%

K1-Linearkernel, K2-Polynomial kernel, K3-Gaussian kernel, K4-Sigmoid kernel

6.6 Conclusion

The utility of six nonlinear chaotic characteristics in distinguishing pathological from healthy voice signals, including correlation dimension at minimum embedding dimension, correlation entropy at minimum embedding dimension, and four fitting coefficients of the f(α) spectrum of strange attractor, has been studied. The study relied on the VOICE database. FNN and MI have optimised the embedding dimension and time delay of RPS. The data was subjected to a statistical surrogate analysis to ensure that the characteristics used in the analysis were discriminated, and a reasonable significance level indicated the presence of nonlinearity. Based on the measures examined, a classification system is proposed. SVM was used to assess the performance of the proposed classification system in distinguishing between pathological and normal voices. The success rates obtained with combined features are higher than those obtained with individual parameters, and the linear kernel provides the best accuracy and precision. The precision is 99%, and the accuracy is 97%. When compared to recognition algorithms based on linear feature vectors and other nonlinear parameters, this accuracy is promising. This demonstrates that these six parameters can distinguish between healthy and pathological speakers. The use of measured nonlinear features in the speech production system can help with pathology diagnosis.

CHAPTER 7

NOISE IDENTIFICATION IN SPEECH BY MULTIFRACTAL DETRENDED FLUCTUATION ANALYSIS

7.1 Introduction

The speaker's voice is always mixed with background noise when it is received by the listener through any means. Identification of the noise kind and SNR will aid in the removal of noise from voice data for improved perception. Since there is a lack of information on the type and amount of noise, it is difficult to eliminate it. Speech augmentation, speech processing, and crime investigation are all applications of noise identification. Background noise can be used to pinpoint a possible location during communication. Many studies based on linear methods are reported recently for noise identification [159], [160].Feature extraction is a challenging task under environment noise conditions [161], [162].

Nonlinear tools have evolved as an alternative tool in speech applications in recent years. Multifractality in speech time series data has been used in a variety of speech research, including speaker identification, emotion recognition, speech synthesis, and speech processing. Linear Predictive Coefficients (LPC) and Mel Frequency Cepstral Coefficients (MFCC), in conjunction with multifractal variables, provide good speaker recognition accuracy [87]. In noisy situations, a combination of Gammatone Frequency Cepstral Coefficients (GFCC) and MFCC is used to verify speaker’s identify [88]. Both the correlation dimension and correlation entropy of human voice signals exhibited a statistical reduction after surgical removal of vocal polyps, according to Zhang et al. [76]. For the investigation and detection of voice disorder, Huang et al. [52] employed correlation

dimension and correlation entropy at the minimal embedding dimension. It has been demonstrated that combining MFCC with vector quantisation improves back ground noise estimation [91]. Sarkar et al. [92] used multifractal detrended fluctuation analysis to resolve the language dependency in speaker recognition with Bengali. Audio magnetotellaric signal noise identification has recently used multifractal spectrum analysis and matching pursuit [93]. In recent years, multi scale chaotic speaker recognition systems and its accuracy issues have become a hot topic of research [94]. Despite the fact that a significant number of parameters are employed to estimate noise in speech, the multifractality of noisy time series has yet to be exploited.

The effect of noise on speech data may be linguistically dependent, and features developed for one language model may not be equally applicable to other language models. This study uses the Malayalam voice data base to extract multifractal features for noise identification. At various SNRs, pink noise, red noise, and white Gaussian noise are mixed to the speech signal. The effects of these forms of noises on the environment have been extensively researched [163]. It was observed that the nature of the shift in the singularity spectrum indicates the type of noise in the data, and that the spectrum width shows a reasonable decline with SNR. For noise detection, the spectrum width ($\delta\alpha$) and extremum values of holder exponents (α_{min} and α_{max}) can be employed as feature vectors. With, $\delta\alpha$, α_{min} and α_{max} as feature vectors, an SVM classifier[164] is used to predict noise type with 98 percent accuracy.

The chapter is organised as follows: Section 7.2 covers the simulation of noisy signals utilised in the study. Section 7.3 explains the procedure of multifractal detrended fluctuation analysis (MFDFA). Section7.4 discusses results and statistical analysis. Section 7.5 concludes the work.

7.2 Simulated Noisy Signal

During the recording, processing, and transmission of signals, noise is defined as a signal with numerous frequency components of varying strengths that might affect the nature of the original signal. Various noises alter voice signals in the current world, lowering their perceptual quality and intelligibility. Listener fatigue is caused by low perceptual quality, and poor intelligibility results in poor performance in various speech-based applications. The two types of speech signals are vowels and consonants. Consonant sounds have high-frequency properties, whereas vowel sounds have low-frequency features. The majority of the information in a voice signal is provided by consonants. As a result, disturbances that damage speech in various frequency ranges should be investigated. The spectrum features of three noise signals were examined: white Gaussian noise, pink noise, and red noise. These noises were classified as "coloured noise," which describes their frequency response using the concept of colour.

7.2.1 Signal to Noise Ratio

Any real-world speech-based application must deal with different signal-to-noise ratio (SNR) in diverse noisy environments (SNR). The signal-to-noise ratio is a metric that measures how strong the signal is in comparison to the noise. The SNR is frequently given in decibels and is defined as the ratio of signal power to noise power (dB). On a linear scale, SNR is defined in terms of power as

$$SNR = \frac{P_s}{P_n} \tag{7.1}$$

Where 'P_s' is the clear signal's power and 'P_n' is the noisy signal's power. Eq.7.2 gives the SNR in decibel(dB) scale. Table 7.1 shows the SNR values of the voice signal included in this database, as well as their linear scale representation.

$$SNR\ (dB) = 10log_{10}\ (SNR) \tag{7.2}$$

Table 7.1 SNR in linear scale and dB

SNR in linear scale	SNR in dB
100	20
10	10
1	0
0.1	-10
0.01	-20

7.2.2 Different types of Coloured Noises

(A) White Gaussian Noise

The power distribution of white Gaussian noise, also known as white noise, is homogeneous across all frequencies from zero to half the sampling frequency. At a sampling rate of 44,100 Hz, white noise, for example, has the same power between 100 and 500 Hz as between 20,000 and 20,500 Hz. White noise is defined as a sequence of statistically uncorrelated random numbers generated from a Gaussian distribution, usually with zero mean and unity variance. A human ear hears pure white noise as a hissing sound when the TV or radio is tuned to an unoccupied frequency.

The features of white Gaussian noise are depicted in Figure 7.1. The time-domain representation of white noise with zero mean and unit variance is shown in Fig. 7.1 (a). Fig. 7.1(b) shows the frequency response of white noise in a semi-log graph. One axis of a semi-log graph is on a logarithmic scale, while the other is on a liner scale; the spectral magnitude is on the logarithmic scale. Power Spectral Density (PSD), is a commonly used measure for noise analysis that characterises the average behaviour of fluctuating values. Even if white noise's spectrum isn't perfectly flat, the "average power" is the same at all frequencies. The histogram of white noise

corresponds to the Gaussian random variable's theoretical probability distribution function, as seen in Fig. 7.1 (c). The autocorrelation function calculates the time-varying values associated with 't' and 't+τ' and is an impulse at zero lag is shown in Fig. 7.1 (d), ensuring that white noise's random fluctuation is highly uncorrelated. The influence of white Gaussian noise in the time and frequency domain is seen in Fig. 7.2.

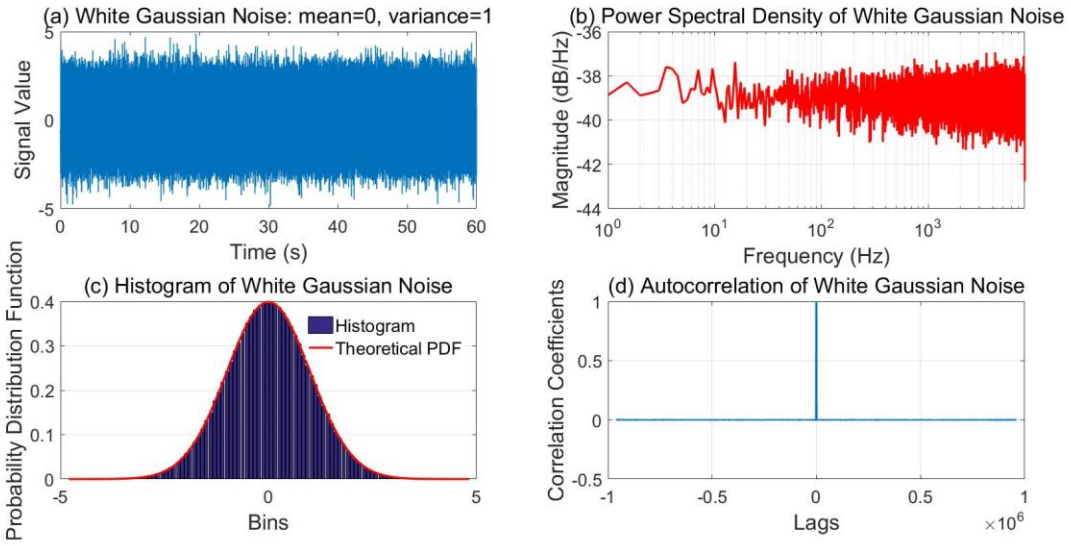


Fig. 7.1 White Gaussian Noise's Characteristics

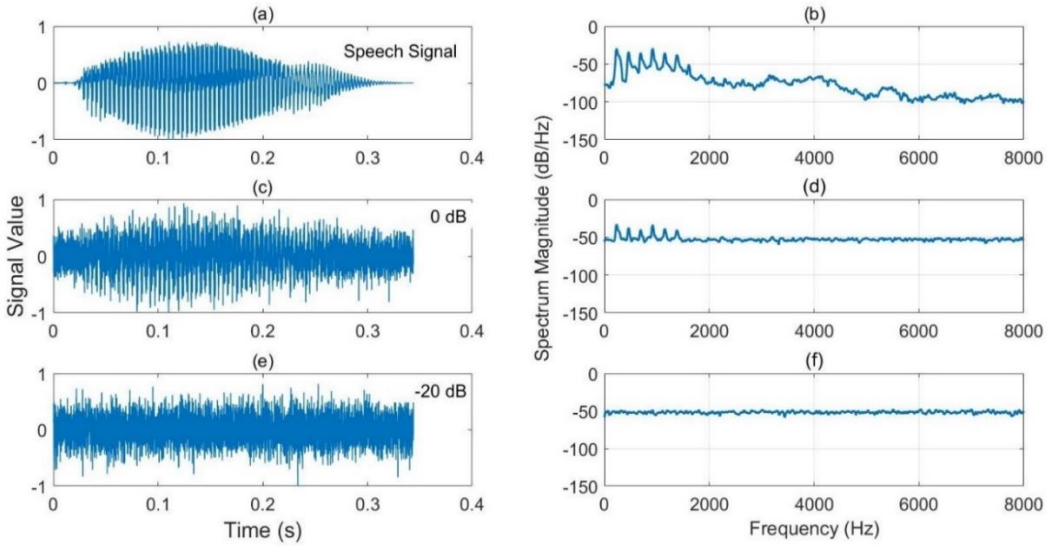


Fig.7.2 Impact of white Gaussian noise in time and frequency domain

(B) Pink Noise

Pink noise, often known as "1/f noise," is a signal with an inversely proportional relationship between power spectral density and its frequency. In the logarithmic scale, pink noise has the same power distribution throughout each octave; hence, the power between 200 Hz and 400 Hz is the same as that of the power between 2,000 Hz and 4,000 Hz. The power within every constant bandwidth of pink noise falls at higher frequencies at a rate of roughly -10dB per decade, since power is proportional to amplitude squared. As a result, pink noise is comparable to white noise, which has a 10 dB per decade falloff. Pink noise is produced by a variety of physical phenomena in the natural world, one of which being flicker noise in electronics.

The time-domain representation of pink noise is shown in Figure 7.3 (a). Figure 7.3(b) shows the frequency response of pink noise in a semi-log graph. Pink noise's spectral property is demonstrated by setting the rolling rate to -10dB/decade. The histogram of pink noise is identical to the theoretical probability distribution function of a Gaussian random variable (see Fig 7.3(c)). Pink noise is white noise that has been filtered at a low frequency. The random samples have a less sharp transition between them now that the high-frequency components have been removed, as shown in Fig. 7.3 (d) as compared to 7.1. (d). The influence of pink noise in the time and frequency domain is seen in Fig 7.4.

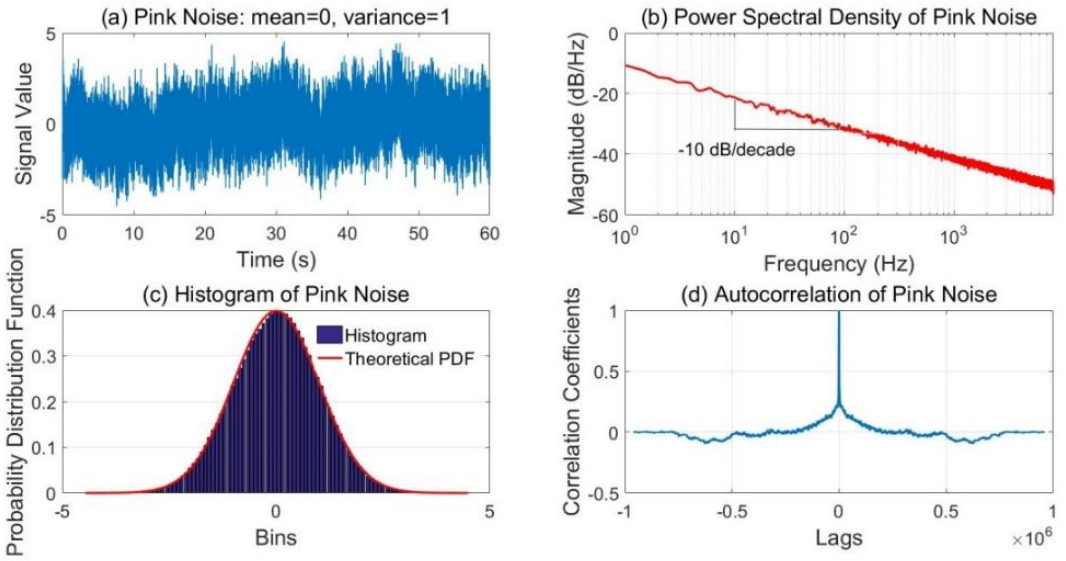


Fig. 7.3 Pink Noise's Characteristics

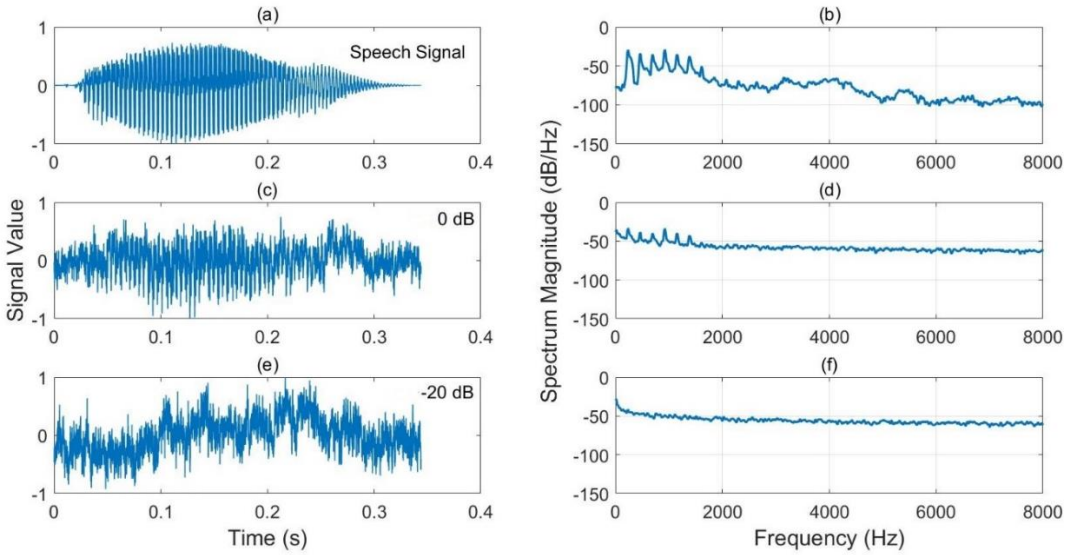


Fig.7.4 Impact of pink noise in time and frequency domain

(C) Red Noise

Pink noise has a power spectrum corresponding to $1/f^2$ and is known as red, brown, or Brownian noise. Lower frequencies in brown noise therefore have more energy than higher frequencies. Brown noise is frequently referred to as "red noise" because of its low frequency, which is similar to that of red

light. Red noise decays at a rate of 20dB per decade at higher frequencies. As a result, red noise is the same as white noise, with 20dB per decade decay.

The time-domain representation of red noise is shown in Fig 7.5 (a), which drifts up and down but has a clear correlation between subsequent values. Figure 7.5(b) shows the frequency response of red noise in a semi-log graph. The rolling rate is set at -20dB/decade to show the spectral property of red noise. The histogram of red noise deviates from the Gaussian random variable's theoretical probability distribution function, as seen in Fig 7.5 (c). Because of the high correlation between the samples in Red noise, the autocorrelation function in Fig 7.5(d) has a repeated pattern. Finally, Fig 7.6 depicts the temporal and frequency domain effects of red noise.

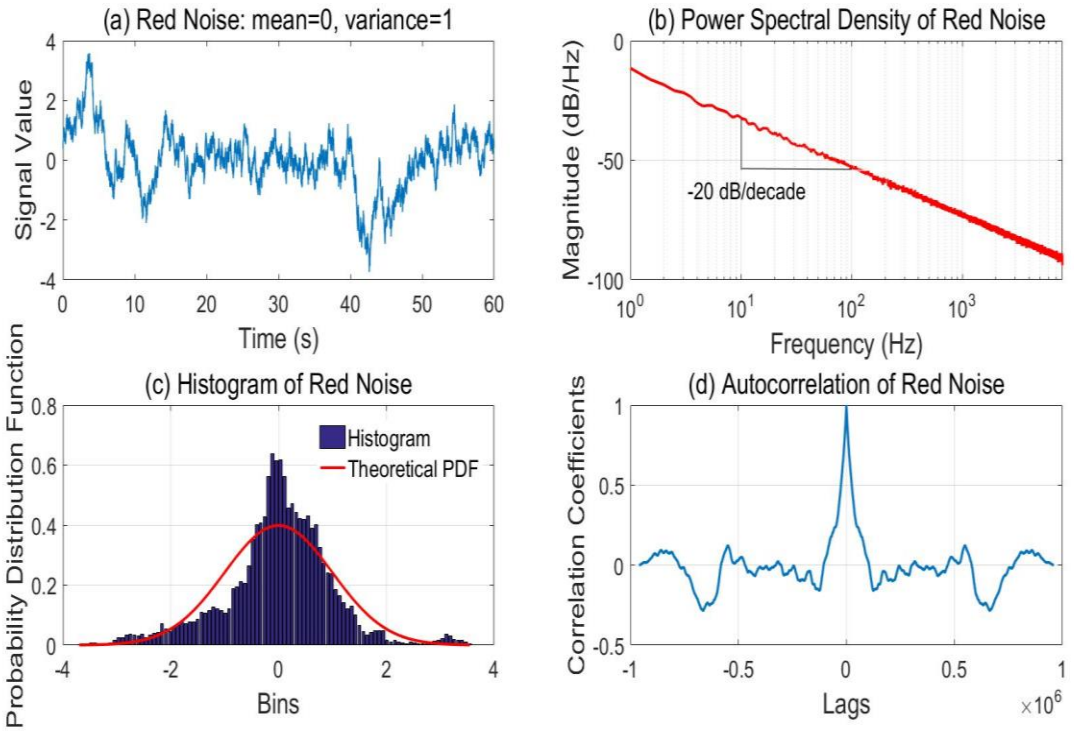


Fig. 7.5 Red Noise's Characteristics

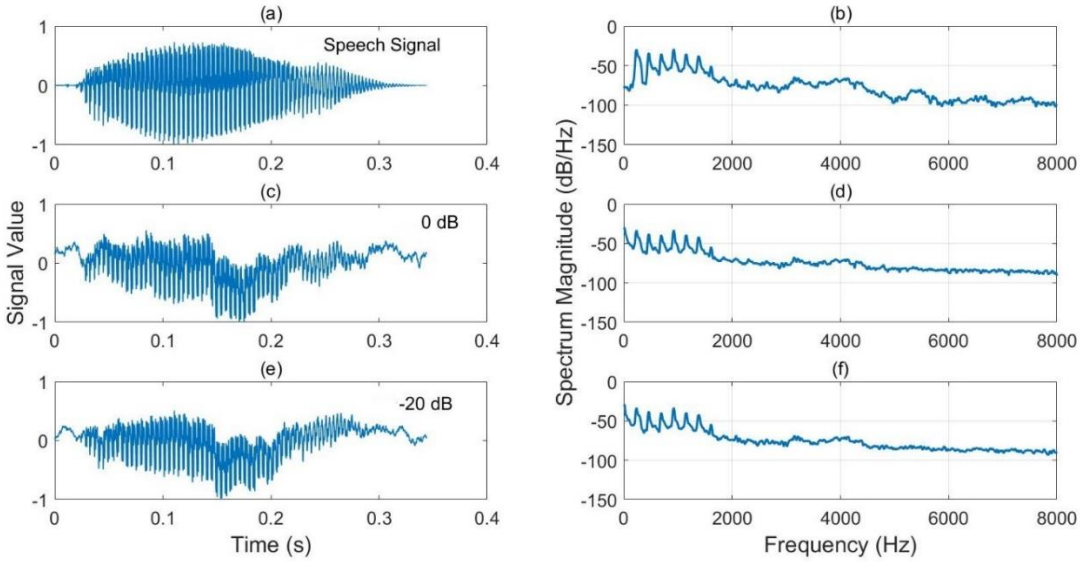


Fig.7.6 Impact of red noise in time and frequency domain

The corrupted speech signal must be appropriately labelled before analysing the influence of the noise signal in speech. Each phoneme and allophone had previously been represented by 12 and 13 alphanumeric characters, respectively. Four more letters are utilised to add noise information. The first character stands for the first letter of the noise type utilised in this database, i.e. 'W' for White noise, 'P' for Pink noise, and 'R' for Red noise. The noise level is represented by the next three characters, which are +20, +10, +00dB, -10dB, and -20dB on a dB scale. A corrupted phoneme file is labelled with 16 alphanumeric characters, and its matching corrupted contextual variation is labelled with 17 alphanumeric characters. This labelling method aids in the selection of the appropriate database subset and the retrieval of all relevant information about the speech signal.

7.2.3 Generation of coloured noisy signal

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

Step1: Create randomised signals from a finite-length Gaussian distribution.

Step2: Take Fast Fourier Transform (FFT) of the randomised signal

Step3: Remove the symmetric component of the spectrum to modify the left half of the spectrum.

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

-Don't make any changes to the spectrum.

PSD equals K/f for pink noise, where K is a constant and f is the frequency.

-To manipulate the spectrum, divide the amplitude of the spectrum by the square root of the frequency indexes. Because power is related to amplitude squared, power per Hz will drop at higher frequencies by around 10 dB every decade.

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

-Divide the spectrum amplitude with frequency indices to manipulate the spectrum. At higher frequencies, the power per Hz will drop by around 20 dB each decade.

Step5: Reconstruct the entire spectrum.

Step6: Use the Inverse Fast Fourier Transform (IFFT) to transfer noisy signals from the frequency domain to the time domain.

Step7: Ascertain that the mean is zero and the standard deviation is one.

At various SNR values, the produced coloured noise is additively merged with the speech signal. Convert the given SNR in dB into the linear scale and plug it into the equation 7.3 to generate the noisy signal with the desired SNR. Then, with a suitable SNR, mix the speech signal with the noisy signal.

$$\text{Noise of desired } SNR = \sqrt{\frac{P_{signal}}{SNR_{linear}}} * noise[m] \quad (7.3)$$

7.3 Multifractal Detrended Fluctuation Analysis(MFDFA)

Many signals don't have a straight forward monofractal scaling behaviour that can be explained by a single scaling exponent. There may be crossover (time-) scales dividing regimes with different scaling exponents. The scaling behaviour is more difficult in other circumstances, and different scaling exponents are necessary for different portions of the series. Different scaling behaviour can be seen for multiple interwoven fractal subsets of the time series in even more complicated scenarios. In this scenario, a multifractal analysis is required to fully describe the scaling behaviour over the same range of time scales, which necessitates the use of a large number of scaling exponents. Higher order correlations can be discovered via a multifractal analysis of time series.

The standard partition function multifractal formalism, which was designed for the multifractal characterisation of normalised, stationary measures, is the most basic sort of multifractal analysis. Unfortunately, for non-stationary time series that are affected by trends or cannot be normalised, this basic approach does not produce accurate results. As a result, in the early 1990s, the wavelet transform modulus maxima (WTMM) method was created, which is based on wavelet analysis and involves tracing the maxima lines in the continuous wavelet transform over all scales. The multifractal DFA (MF-DFA) algorithm is a significant alternative, as it does not require the modulus maxima procedure and hence requires less programming effort than the ordinary DFA.

MFDFA, a procedure to analyse a biomedical time series, was introduced by Kantelhardt et al. "It is a generalisation of the detrended fluctuation analysis and can be obtained in five steps" [165].

1) The profile of the time series x_k of length N and average value <x> is determined by the relation

$$Y(j) = \sum_{k=1}^{j} |x_k - < x >|, \quad j=1, 2, ..., N \tag{7.4}$$

2) The profile is divided in to Np segments from both sides to obtain 2Np segments each of length p (=N/Np).

3) The variance of the time series is determined after least square fit of the time series with appropriate polynomial function.

$$F^2(p, \epsilon) = \frac{1}{p} \sum_{j=1}^{s} \{Y[(\varepsilon - 1)p + j] - Y_\epsilon(j)\}^2 \tag{7.5}$$

For ε=1,2,.....N_p and

$$F^2(p, \epsilon) = \frac{1}{p} \sum_{j=1}^{s} \{Y[N - (\varepsilon - N_p)p + j] - y_\epsilon(j)\}^2 \tag{7.6}$$

For ε=N_p+1,N_p+2,.....,2N_p

Where $y_\varepsilon(j)$ is the polynomial used for least square curve fitting. The order of polynomial will be different for different time series and the best fit should be used in order to eliminate local trends in time series. The ε should selected such that ε > m+2,where m is the order of polynomial function.

4) The q [th] order fluctuation function is evaluated from the equation

$$F_q(p) = \frac{1}{2N_p} \left\{ \sum_{\varepsilon=1}^{2N_p} [F^2(p, \varepsilon)]^{q/2} \right\}^{1/q} \tag{7.7}$$

5) From the log $F_q(p)$ versus p plots the behaviour of the fluctuation function can be predicted. The $F_q(p)$ can be expressed as a power law

$$F_q(p) = p^{h(q)} \tag{7.8}$$

Where h(q) is the generalised Hurst exponent. The generalized mass exponent and singularity spectrum can be calculated from the relations

$$\tau(q) = qh(q) - 1 \tag{7.9}$$

$$f(\alpha) = q\alpha - \tau(q) \tag{7.10}$$

$$f(\alpha) = q[\alpha - h(q)] + 1 \tag{7.11}$$

α is the singularity strength or Holder exponent. The width of singularity spectrum (δα) can be calculated by taking the maximum and minimum α values in the f(α) spectrum.

7.4 Experiments and Results

The MFDFA analysis of clean and simulated noisy speech signals together with the results of SVM classifier is discussed in following sections.

7.4.1 Analysis of Clean Speech Signal

The analysis is based on a Malayalam speech vowel data set uttered by 100 speakers of the same age group (20-25) as discussed in chapter 3. Three types of noisy signals have been simulated using the algorithm given in section 7.2.3 [166] .The generalised Hurst exponents are calculated by varying the "q" value between -10 and 10, and the best least squared polynomial was found to be three. Figures 7.7 to 7.11 show the outcome of multifractal analysis of five Malayalam vowels sampled at 16 kHz.

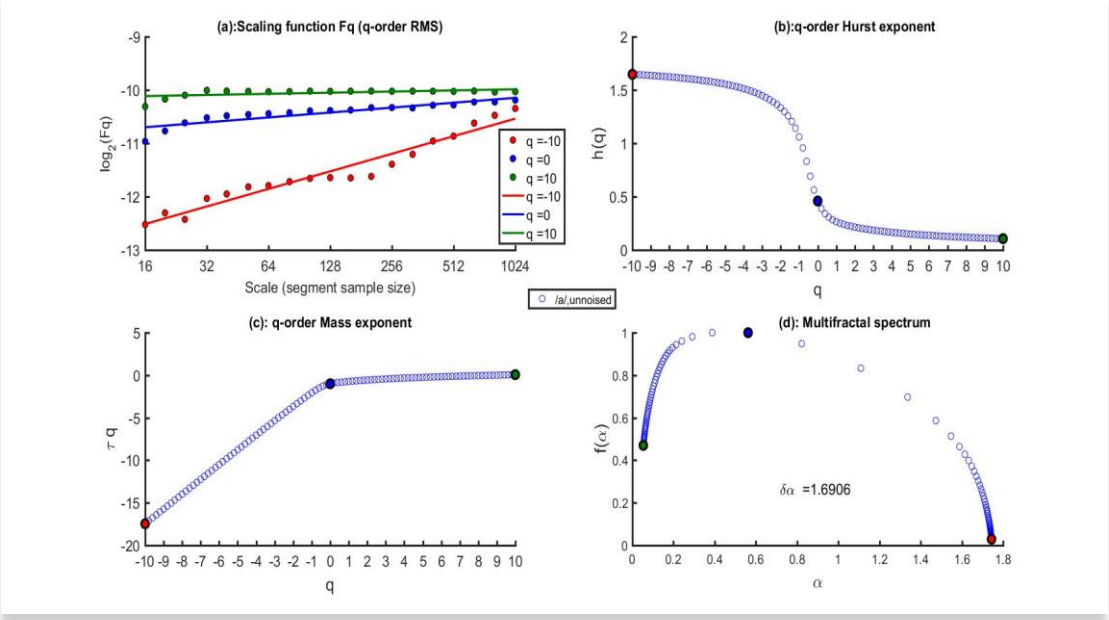


Fig. 7.7 Multifractal analysis of clean Malayalam vowel ഓ /a/

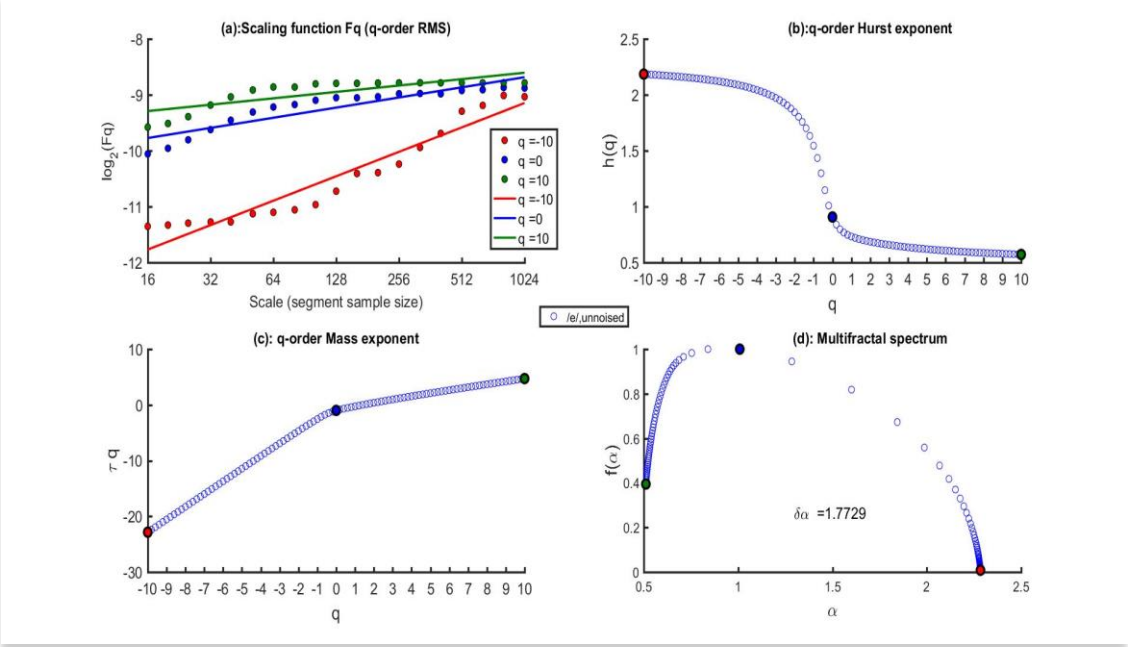


Fig. 7.8 Multifractal analysis of clean Malayalam vowel ഐ /e/

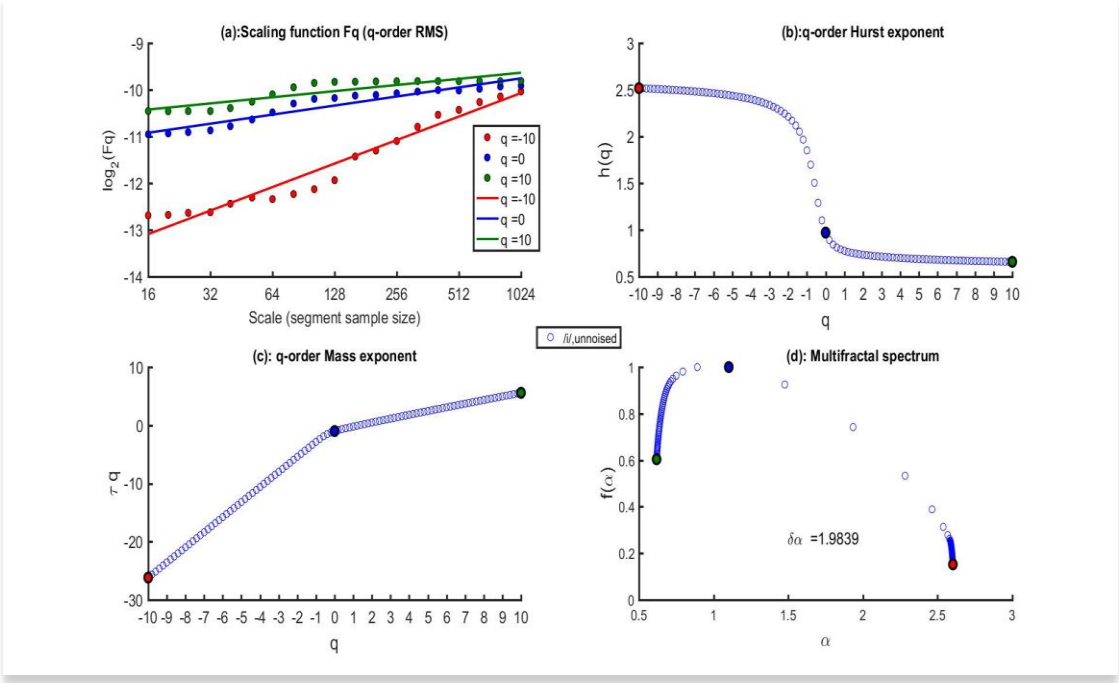


Fig. 7.9 Multifractal analysis of clean Malayalam vowel ഇ /i/

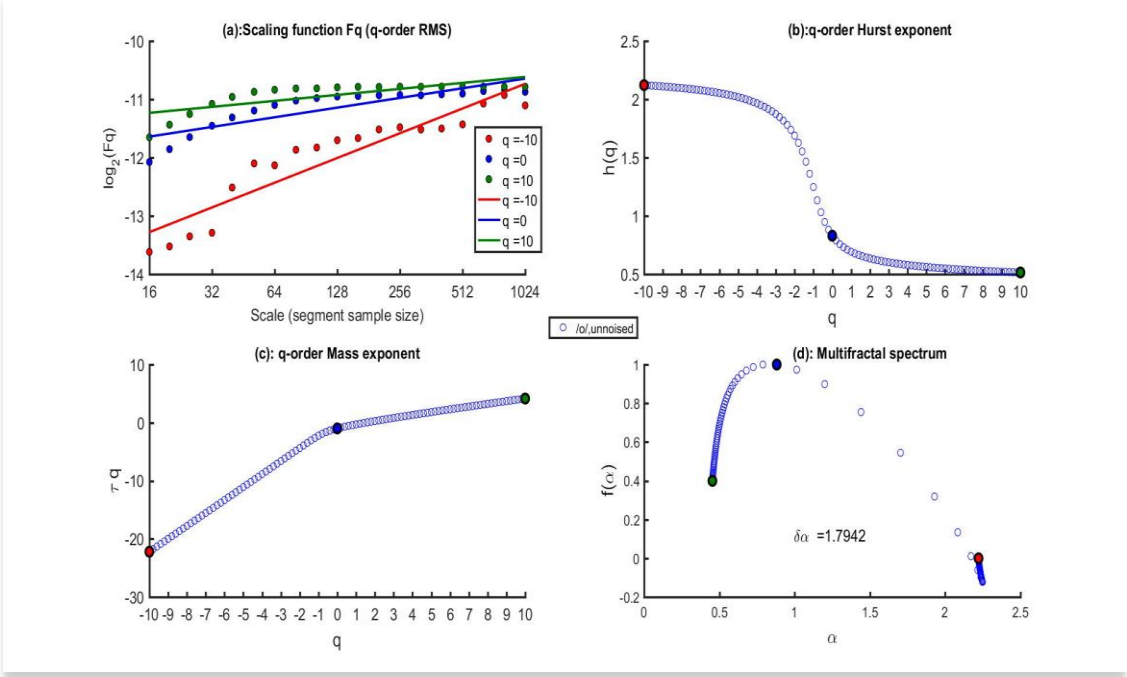


Fig. 7.10 Multifractal analysis of clean Malayalam vowel ഒ /o/

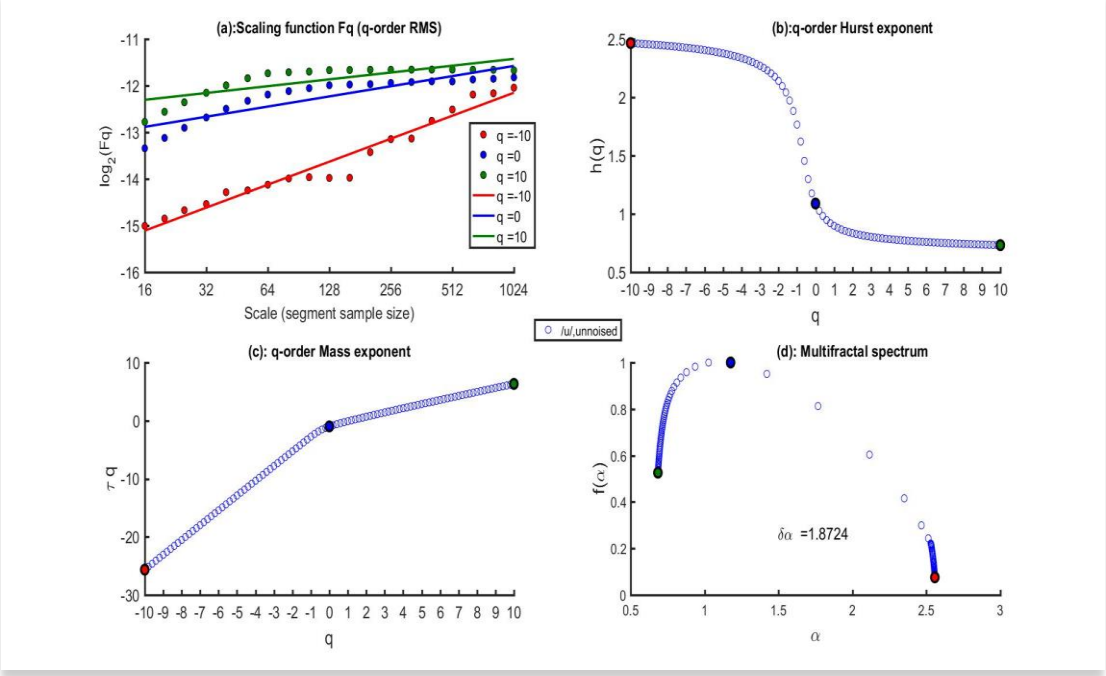


Fig. 7.11 Multifractal analysis of clean Malayalam vowel ഉ /u/

For all of the samples tested, the log of F (q) varies linearly with segment sample size for optimum and null values of q, as shown in figures 7.7 (a) to 7.11 (a). The q-order Hurst exponent falls with the q value for all samples, as shown in figures 7.7 (b) to 7.11 (b). The mass exponent 'τ(q)' has a curved q dependency, as shown in figures 7.7 (c) to 7.11 (c). It designates a multifractal spectrum with power law exponents. The singularity spectrum width changes from speaker to speaker or even with different moods of the same speaker, and the variation of the generalised Hurst exponent indicates a unique structure for all vowels. The lowest value for the specified speaker is 1.7729, and the maximum is 1.9839, as shown in figures 7.7 (d) to 7.11 (d). When the full database was analysed, it was found that practically every sample's spectrum width falls between 1 and 3, and the spectrum's structure and behaviour appear to be the same.

7.4.2 Analysis of Noisy Speech Signal

Three forms of noise are added to the speech data samples to explore the noise impacts on the data: pink noise, white Gaussian noise, and red noise. By combining the three types of noise, samples with signal to noise ratios of 0 dB, 3 dB, 7 dB, 10 dB, 14 dB, 17 dB, and 20 dB are created. The effect of noise on all vowels was found to be identical; hence the effect can be used to detect the presence of noise in any speech data. The multifractal spectrum for SNR 0dB in the vowel /a/ of a particular speaker for pink noise is shown in Fig 7.12.Fig 7.13 shows the effect of 0dB red noise on the vowel അ /a/.The variation in multifractal spectrum for 0dB white Gaussian noise on the vowel അ /a/ is seen in Fig 7.14.

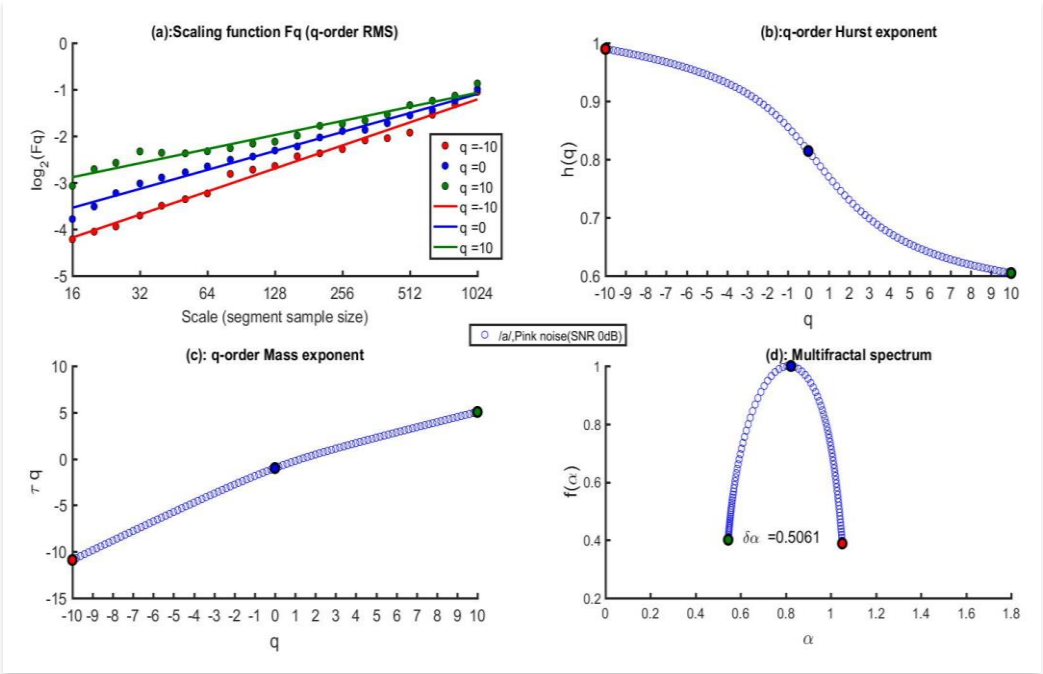


Fig. 7.12 Multifractal analysis of Malayalam vowel അ /a/ with 0 dB Pink noise

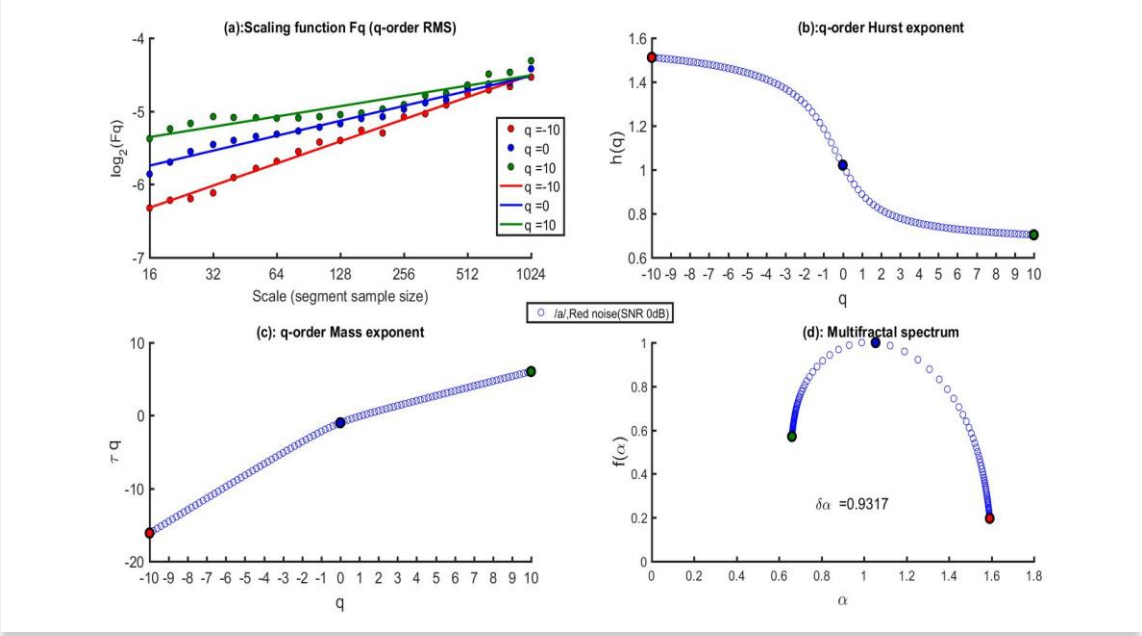


Fig. 7.13 Multifractal analysis of Malayalam vowel അ /a/ with 0 dB Red noise

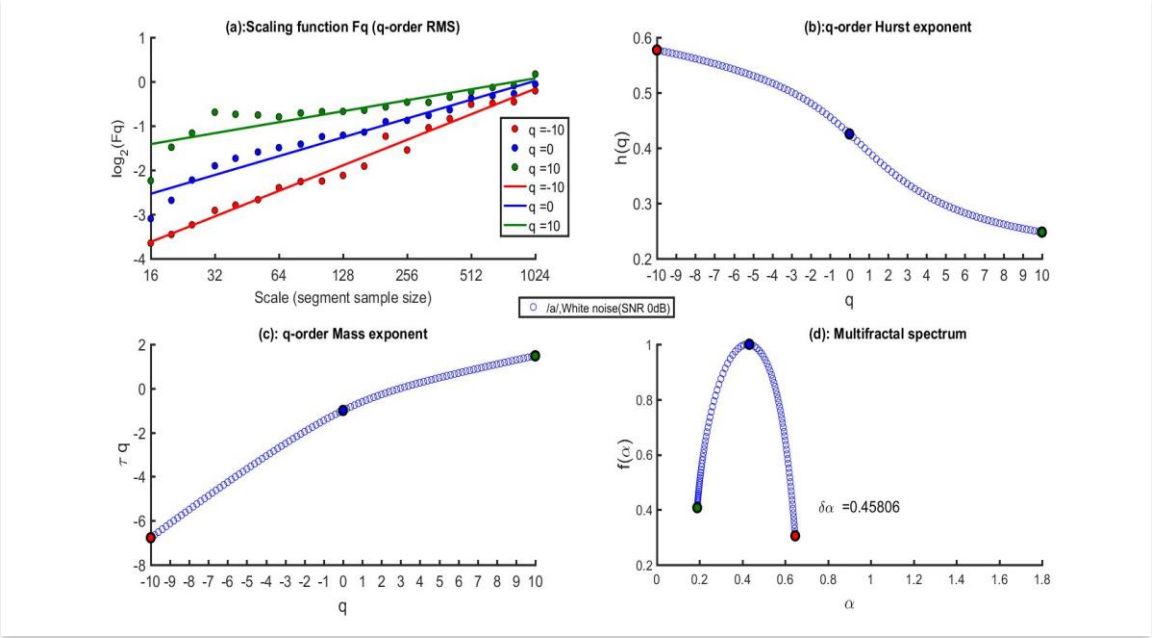


Fig. 7.14 Multifractal analysis of Malayalam vowel അ /a/ with 0 dB White noise

The log plots show a linear relationship even with the addition of different kinds of noise, and the corresponding generalised Hurst exponent also shows the same behaviour. The mass exponent has a curvy dependence on q, and the q^{th} order fluctuation shows a power law relationship with the generalised Hurst exponent. As can be seen in figures 7.12,7.13 and 7.14, $\delta\alpha$ decreases due to the presence of noise in the data. The decrease varies depending on the type of noise. It is the highest for white noise and lowest for red noise.

As a result of the noise effect, the singularity spectrum shifts, changing the values of the Holder exponent's minimum and maximum. It was observed that there is no reasonable change in the height of the spectrum ($f(\alpha)$) due to additive noise. Hence, it cannot be taken as a candidate for noise identification. For all circumstances, the minimum changes to the right and maximum shifts to the left (Figures 7.12–7.14). The degree of shifting differs despite the fact that both signals have the same SNR. The fluctuation of the $f(\alpha)$ spectrum with additive noise ranging from 0 to 20 dB for various types of noise is plotted. It was discovered that when the amount of noise increases, the width of the singularity spectrum narrows (i.e., with the decrease of SNR). For Malayalam vowel /a/, Fig 7.15 displays the changes in the $f(\alpha)$ as a result of the addition of pink noise compared to the clean signal. Figure 7.16 depicts the shift in spectrum width for red noise spanning from 0 dB to 20 dB, whereas Fig. 7.17 depicts the same for white Gaussian noise.

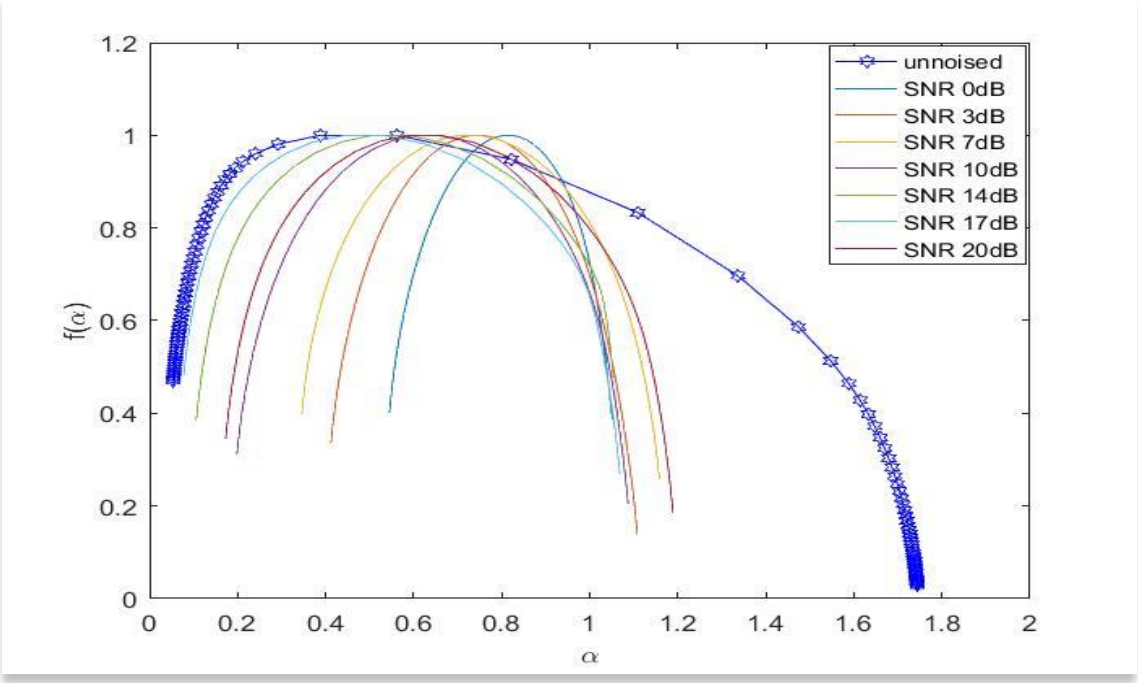


Fig. 7.15 The change in f(α) spectrum with addition of Pink noise

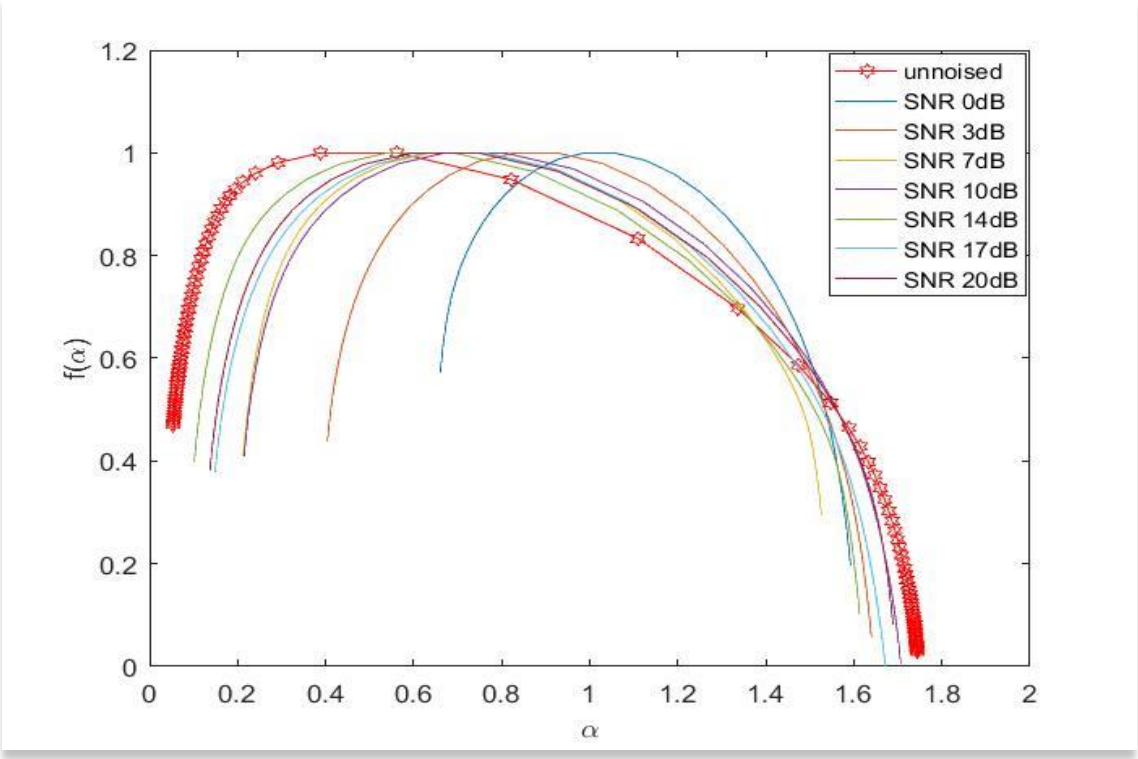
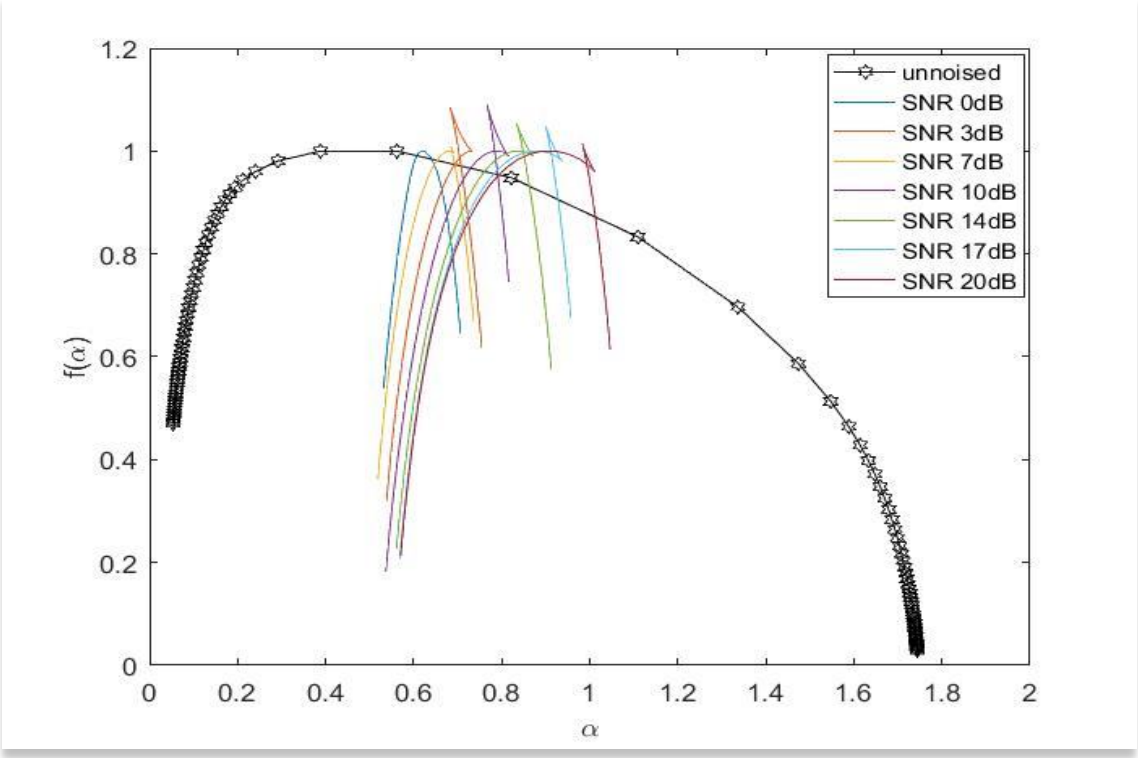


Fig. 7.16 The change in f(α) spectrum with addition of Red noise



7.17 The change in f(α) spectrum with addition of White Gaussian noise

With SNR, there is a systematic reduction in δα, but the reduction is not the same for all types of noise. For a given SNR, the decline is greatest for white noise and lowest for red noise. The asymptotic value of tends to equal the value for unnoised data.

Even if the value of varies from sample to sample, the fall in samples follows the same pattern. The reduction in the δα shows a particular pattern. Fig 7.18 shows the values of δα for various sorts of noises for two different speakers and the same vowel utterance.

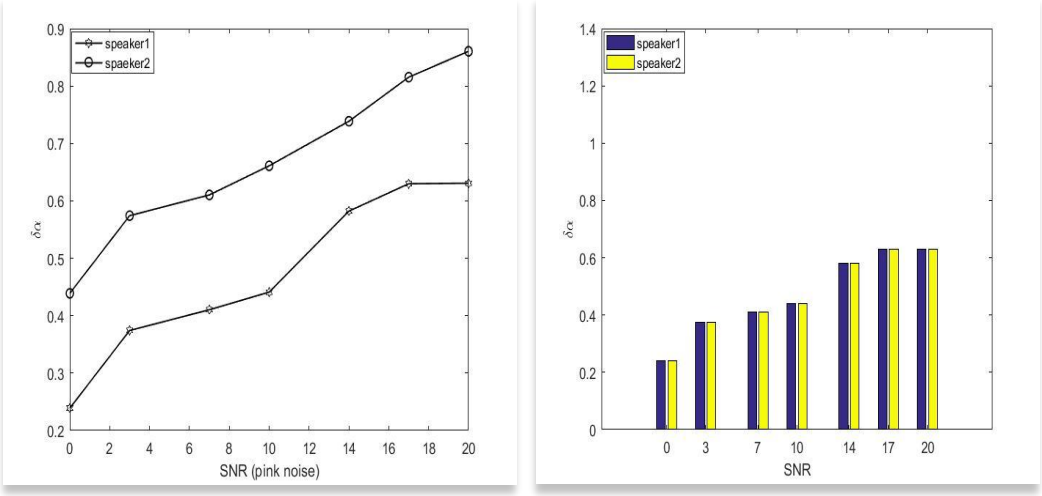


Fig.7.18 Singularity spectrum width of two different speakers for Pink noise

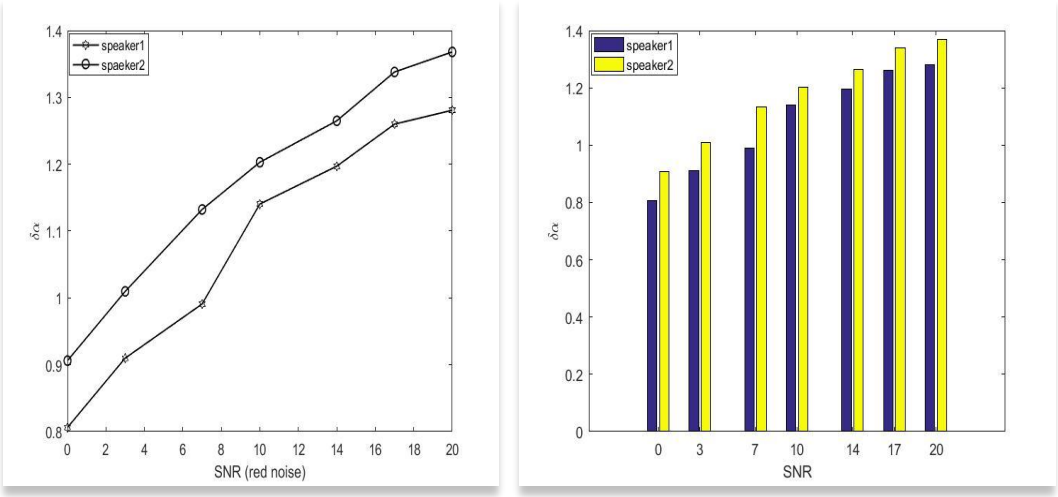


Fig.7.19 Singularity spectrum width of two different speakers for Red noise

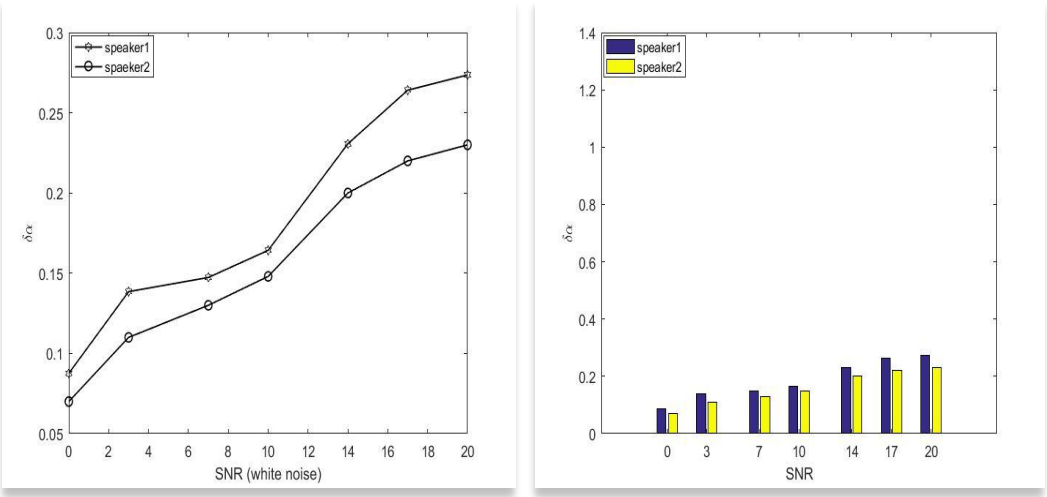


Fig.7.20 Singularity spectrum width of two different speakers for Red noise

7.4.3 Proposed noise identification system

As the Holder exponent (α) varies considerably with noise, it can be used for noise identification from a speech signal. The five Malayalam vowel sounds from 100 speakers (50 male and 50 female) together with noise simulated signals of them are used for analysis. Fig 7.21 shows the average percentage reduction in δα for samples from 100 speakers for the three types of noises with SNR of 0 dB, 3 dB, 7 dB, 10 dB, 14 dB, 17 dB, and 20 dB.

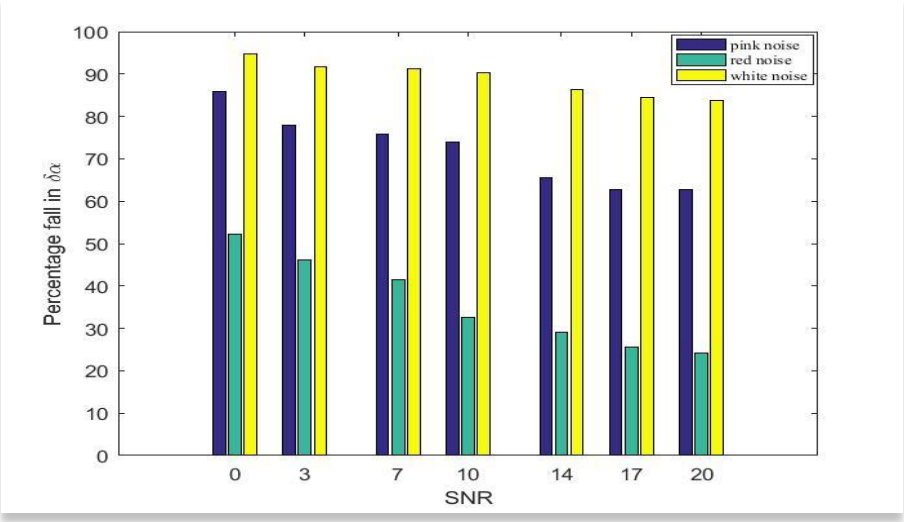


Fig.7.21 Average percentage reduction in singularity spectrum width with SNR

Table 7.2 summarises the average percentage rise in α_{min} percentage fall in α_{max} and percentage fall in δα with error limit due to the addition of 0dB noises in the samples. The proposed noise identification system for identifying the type of noise signal (having a particular noise level) using the multifractal features is shown in Fig 7.22.

Table 7.2 Multifractal features used for noise identification

	Pink Noise (0 dB)	Red Noise (0 dB)	White Noise (0 dB)
$\delta\alpha$	85±20	50±15	95±18
α_{min}	400±38	550±50	100±20
α_{max}	8.57±2.3	3.71±1.1	62.85±8.8

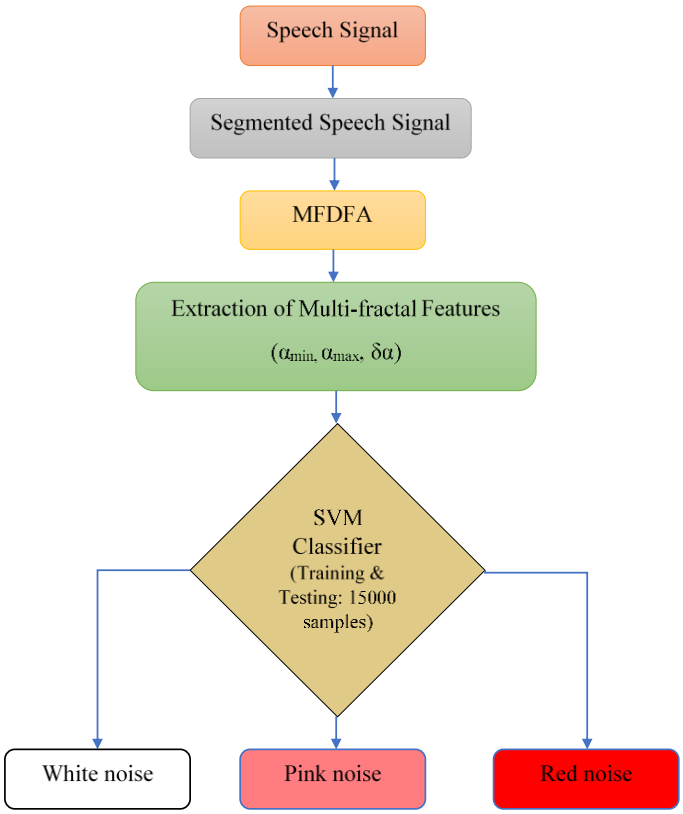


Fig.7.22 Proposed noise type identification system

7.4.4 Results of SVM classifier

To identify the type of noise in the speech signals, an SVM classifier (section 6.4) is employed. α_{min},α_{max} and $\delta\alpha$are taken as feature vectors, and the accuracy and precision of the identification is measured using four kernels. Fig 7.23 shows the confusion matrix corresponding to four kernels and it is

clear from the confusion matrix that the accuracy and precision are maximum when the Gaussian radial basic kernel is used. In the confusion matrix, the index 1 corresponds to pink noise, 2 corresponds to red noise, and 3 corresponds to white noise. The accuracy and precision for the Gaussian kernel are listed in Table 7.3.

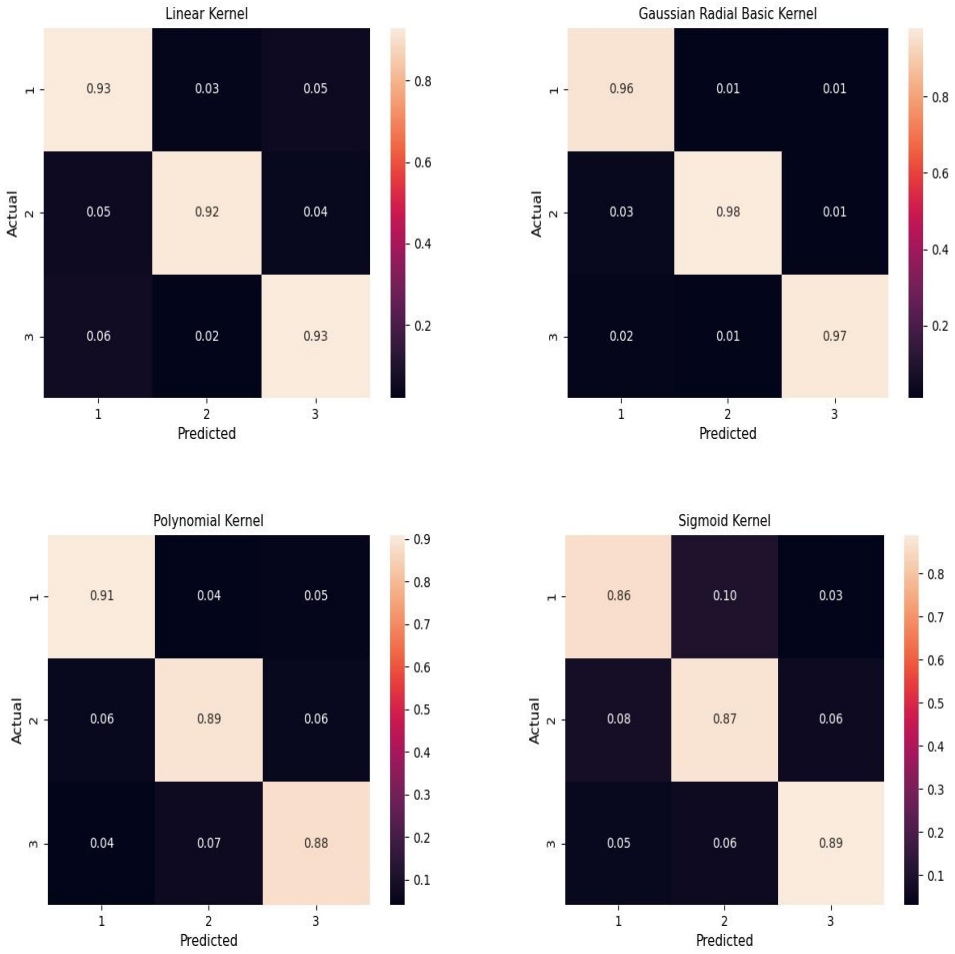


Fig. 7.23 Confusion matrix for noise identification (different kernels)

Table 7.3 Accuracy and precision of identification from Confusion Matrix (Gaussian Radial Basic Kernel)

Noise	Pink	Red	White	Accuracy(%)
Pink	96	1	1	96.96
Red	3	98	1	96.07
White	2	1	97	97.00
Precision(%)	95.04	98.00	97.97	

The accuracy and precision for distinguishing different types of noise are variable, as shown in Table 7.3. Pink noise has an accuracy of 96.96 percent, red noise has an accuracy of 96.07 percent, and white Gaussian noise has an accuracy of 97 percent. Pink noise is identified with 95.04 percent precision, red noise with 98 percent precision, and white noise with 97.97 percent precision.

7.5 Conclusion

The singularity spectrum width and extremal Holder exponents are reduced in the multifractal detrended fluctuation analysis of the voice samples due to additive noise. These characteristics can be used as feature vectors to distinguish between different types of noise. From a noise-added sample of Malayalam short vowels pronounced by 100 Malayalam native speakers, the feature vectors are utilised to distinguish pink, red, and white noises. The SNR has an effect on the reduction, and the SNR rate can be computed by multiplying the percentage reduction in the parameters. The noise categories are recognised using feature vectors and an SVM classifier, and the accuracy attained indicates that multifractal features are an efficient tool for recognising different types of noise.

CHAPTER 8

SPEECH EMOTION RECOGNITION USING NONLINEAR AND MULTIFRACTAL FEATURES

8.1 Introduction

Emotion is crucial to one's physical and psychological health. Negative emotions like sadness and rage are linked to underlying mental health problems that, if left untreated, can lead to dire consequences like self-harm or suicide. Detrimental emotions that aren't caused by mental health issues can be unpleasant in the moment and have a negative impact on one's everyday life and interactions with others. One's voice can pick up on emotion. As more acoustic sensors become connected to the Internet, the discipline of affective computing is seeing a boom in interest in speech-based emotion detection.

Existing speech processing systems do not effectively process emotional speech. As a result, there is a need to increase speech processing systems' capacities for dealing with emotional speech, because the inclusion of emotions in speech makes interaction more natural. Emotional speech analysis, recognition [77], [167]–[169], and conversion [170], [171] have been a hot topic for decades. One of the most difficult aspects of developing human-machine interfaces is detecting and exploiting emotional information from speech.

Because speech characteristics vary during emotional speech production, traits that represent these variations can be used to detect emotions. The most common characteristics for the classification of emotions is based on fundamental frequency, energy shape, silent duration, formant, Mel-band energies, cepstral coefficients of linear prediction, Mel Frequency

cepstral coefficients and voice quality [96], [108]–[112]. Ramamohan and Dandapat [113] conducted an experiment that demonstrated that sinusoidal characteristics can be used to classify emotions. The acoustic features used in SER are divided into two categories: prosodic features and spectral features. The speaker's emotional cues are provided through prosodic characteristics, which are commonly employed in SER [114]. These characteristics are often estimated as pitch and energy tracking contour statistics [97], [115]. Spectral characteristics, which are often taken from the speech spectrum, have gotten a lot of attention in recent years. These qualities might help with recognition by providing additional information for prosodic features [102]. Both prosodic and spectral aspects of the human speech production system are typically estimated using the traditional linear source-filter models [116].

It is evident that significant nonlinear 3D fluid dynamics events occur during speech generation that are not captured by a linear model [172] . Nonlinear dynamics has provided various methods, such as phase space reconstruction (PSR) and fractal dimensions, to bridge the gap between linear deterministic models and highly unpredictable processes. Mutual information, correlation dimension, correlation entropy, Shannon entropy, and Hurst exponents have all been used to detect human emotion in speech [143], [173], [174]. PSR has been used for a variety of speech processing applications in the recent decade, including speech recognition [75], [175], speech augmentation [176], [177], and detecting sleepiness from speech .

Despite the fact that a great number of studies have been reported on the use of nonlinear features to improve speech emotion recognition, there is still room for improvement in accuracy. The use of nonlinear features extracted from the optimised rebuilt hyperspace for emotion recognition is the chapter's major contribution. In addition to spectral and prosodic features, the correlation dimension at minimum embedding dimension (D_{2m}), correlation

entropy at minimum embedding dimension (K_{2m}), largest Lyapunov exponent at minimum embedding dimension (LLE), height of singularity spectrum ($f(\alpha)$), and values of Holder exponents are used as features.

The sections are arranged as follows. Section 8.2 discusses the peculiarities of the emotional speech database used for the investigation. In section 8.3, the spectral, prosodic, nonlinear, and multifractal features used in the study are described. The feature extraction experiments, the proposed classification system, and the results of the SVM classifier are explained in section 8.4. Section 8.5 concludes the work.

8.2 Emotional Speech Database

The emotional speech database is a key component of the speech emotion recognition system. It is the most important criteria, and its quality is crucial for correctly recognising emotions from speech samples. The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [178] is a validated database of emotional speech and song that includes 24 professional actors (12 female, 12 male) vocalising two lexically-matched phrases in a neutral North American accent. It is one of the best databases freely available. The database was released in 2018 by the Department of Psychology at Ryerson University in Toronto, Canada, in association with the University of Wisconsin-River Falls' Department of Computer Science and Information Systems. The RAVDESS Dataset is a collection of audio and video clips with 24 actors uttering the same two sentences while expressing eight distinct emotions. All conditions are accessible in three modality formats: audio-only (16bit, 48 kHz, .wav), audio-video (720p H.264, AAC 48 kHz, .mp4), and video-only (720p H.264, AAC 48 kHz, .mp4), and video-only (720p H.264, AAC 48 kHz, .mp4) (no sound). In this study, we used male voice samples from the RAVDESS database. In the database, the two

sentences "Kids are talking by the door" and "Dogs are sitting by the door" are repeated twice with two intensities.

Fig 8.1 shows the different emotions: angry, happy, eutral, sad, fearful, calm, disgust and surprise, available in the RAVDESS database. Since the emotional backgrounds of male and female voices are quite different, they should be analysed separately. In this work, only male sounds from the database are used. The samples used for the study are listed in Table 8.1.

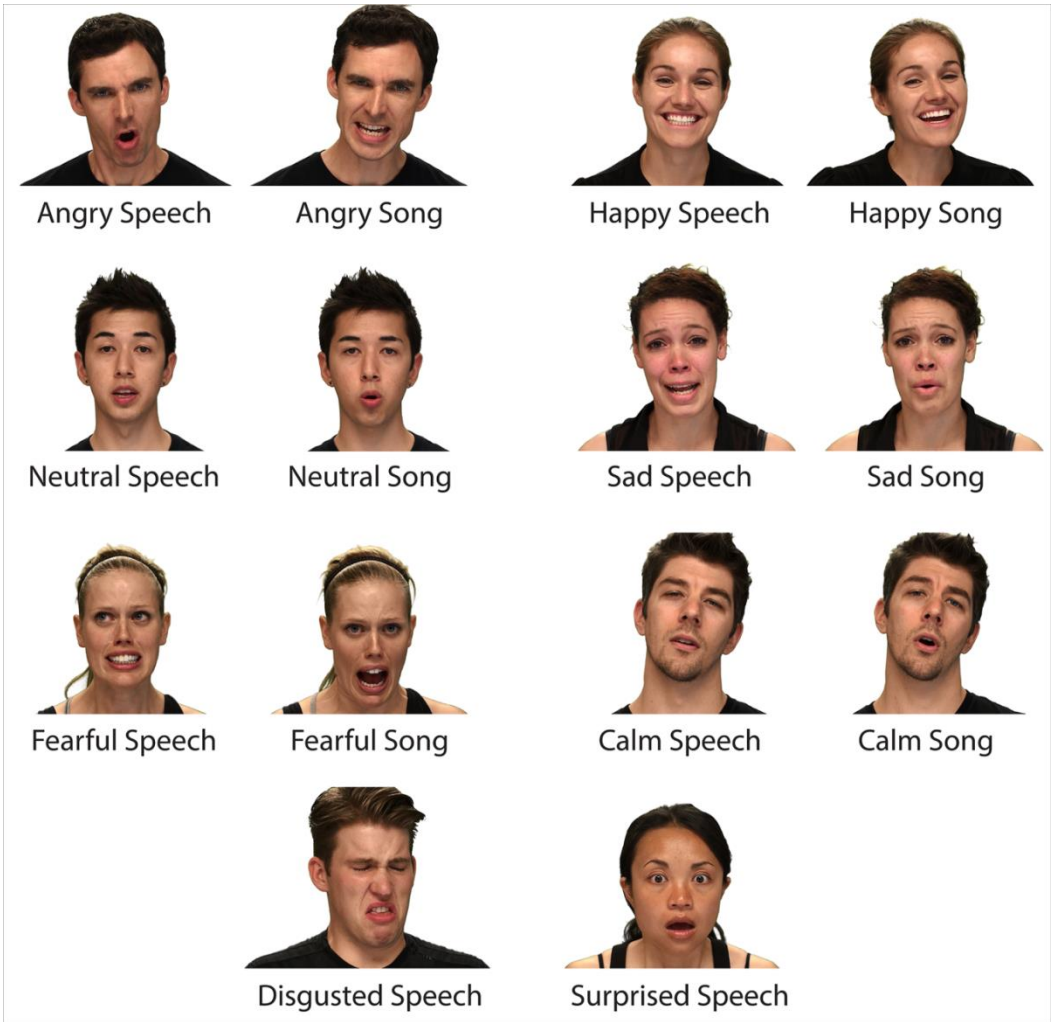


Fig. 8.1 Emotions in RAVDESS database [178]

Table 8.1 Emotional speech samples used for analysis

Emotion(Male)	Number of speakers	Number of statements	Number of repetitions	Number of intense levels	Total samples
Neutral	12	2	2	2	92
Calm	12	2	2	2	92
Happy	12	2	2	2	92
Sad	12	2	2	2	92
Angry	12	2	2	2	92
Fearful	12	2	2	2	92
Disgust	12	2	2	2	92
Surprised	12	2	2	2	92

8.3 Parameterization

The speech signal has a huge number of parameters that indicate emotional traits, and the various parameters cause emotional shifts. As a result, the most important step in speech emotion recognition is to figure out how to extract the feature parameters that can largely reflect speech emotion. A vital step in achieving excellent recognition performance is selecting relevant features. In this work nonlinear features are combined with prosodic and spectral features for emotion recognition.

8.3.1 Prosodic and spectral features

The most widely utilised features in speech emotion recognition are prosodic features. Two of the most widely used features are pitch and formants. The autocorrelation function (ACR) is used in this study to reliably extract distinct properties. The fundamental frequency (F0) is extracted from the voice signal using a repeating pattern in the ACR. It's done by calculating the difference between two successive ACR local maxima. The characterisation of vocal tract properties, particularly formant frequencies, is aided by frequency domain modelling of speech data (F1 and F2). In the frequency domain, formant frequencies are represented as spectral peaks. The

signal should be pre-processed before extracting features. Figure 8.2 summarises the steps involved in extracting acoustic speech features during the pre-processing stage.

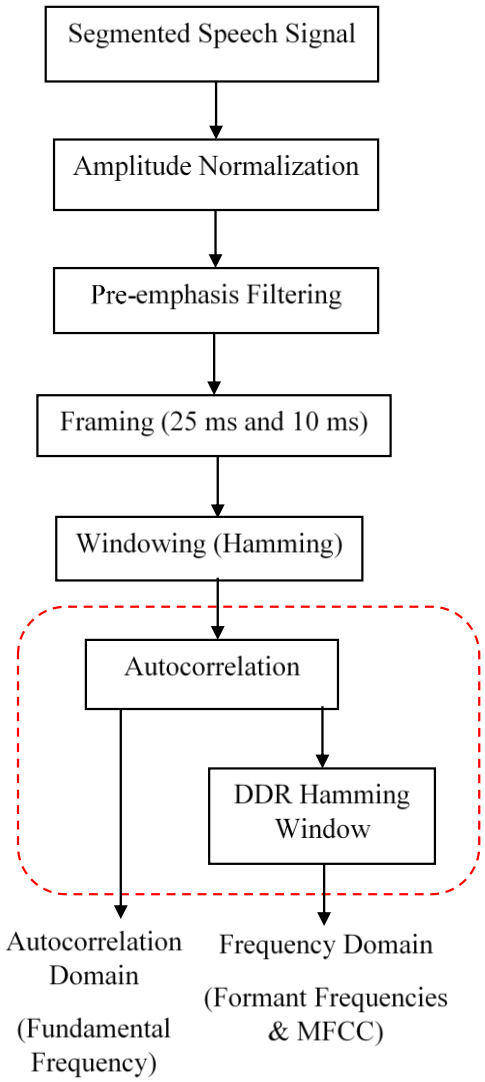


Fig. 8.2 Steps in Pre-processing

The fundamental frequency (F0) is the lowest frequency in a voiced speech signal, as determined by the frequency of the source sounds' quasi-periodic nature (vocal cords). The nature of the source sounds is complex and harmonious. It is made up of a variety of frequencies that are almost integral multiples of the fundamental frequency. The relative strength of the

fundamental frequency and its harmonics characterises the spectrum representation of source sound. F0 is calculated using time-domain algorithms using the voice signal's or its modified version's recurring patterns. The autocorrelation domain, in which the stronger correlation peaks convey the source sound information, is the most explored domain for F0 estimation. To reduce errors in the estimated F0 as in the YIN technique, various modifications to autocorrelation-based methods were implemented. The harmonic structure in the frequency domain contains a lot of information regarding pitch. This kind includes the subharmonics to harmonics ratio (SHRP) [179], summation of residual harmonics [180], Sawtooth Waveform Inspired Pitch Estimator (SWIPE) [181], and others [178], [182]–[184]. In time-frequency domain pitch extraction techniques, the speech signal is separated into many frequency bands, and each sub-band signal is exposed to time-domain operations. A popular time-frequency domain method is the auditory-model correlogram based methodology [185]. To decompose the signal, an auditory filter bank is utilised, followed by autocorrelation computation on each sub-band signal. Some methods [186], [187] use data-driven methodologies to learn how noise affects the amplitude and placement of peaks in the speech spectrum. The methods in [188]–[191]use statistical methodologies to improve F0 estimation. Figure 8.3 depicts the block diagram of the suggested F0 estimation algorithm.

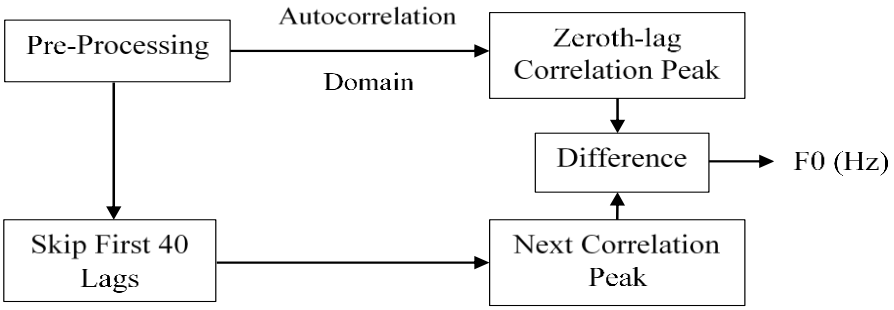


Fig. 8.3 Block Diagram of F0 Estimation Algorithm

The speech signal is made up of a number of frequency components that characterise the differences between source sound (vocal cords) and

system sound (vocal tract) in the speech production mechanism. The content of the spoken word is mostly determined by frequency components. Formant frequencies are related to the vocal tract anatomy because the fundamental frequency of a speech sound is directly tied to the properties of the vocal cords. The vocal tract is an acoustic region that modifies the frequency range of sound as it flows through it from the vocal folds. The shape and size of the vocal tract's frequency response are impacted by the position of active articulators, particularly the tongue. Resonant frequencies, on the other hand, are a related concept that regularly arises in the literature. As a result, some authors considered the two concepts as interchangeable, while others handled them separately. The vocal tract's acoustic property is resonant frequencies, while the speech signal radiated from the lips is formant frequencies. The vocal tract characteristics are retrieved from the speech signal and then filtered in this study. As a result, formant frequencies refer to the vocal tract's equivalent frequency response. It is a spectral peak in the spectrum or dark horizontal bands on the spectrogram that represents a concentration of acoustic energy. Figure 8.4 shows a block diagram of formant frequency estimation.

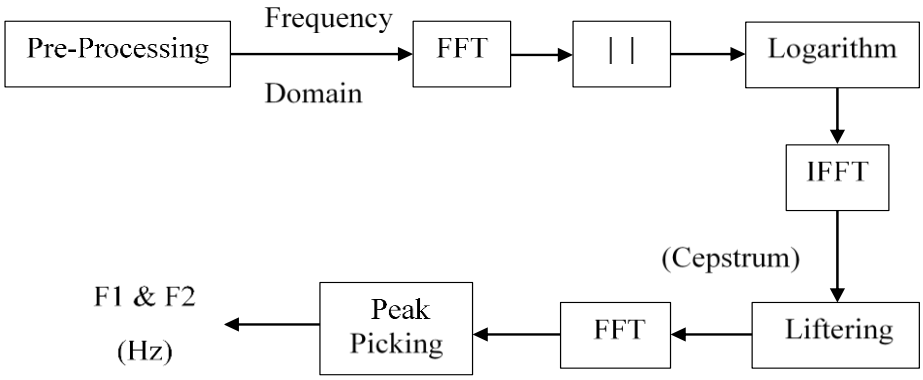


Fig.8.4 Block diagram of F1 and F2 estimation

As effective spectral features for SER, Mel frequency cepstral coefficients (MFCC) are reported. In voice recognition systems, MFCCs are the most extensively used acoustic characteristic [192]. Knowing 'how

humans hear is' more important than 'speaking' in voice recognition and feature extraction. Auditory perception in humans is non-uniform, linear at low frequencies (1000 Hz) but nonlinear above 1000 Hz. As a result, the nonlinear mapping of the recorded speech signal onto the perceived scale is required. The Mel scale represents the recorded frequency in the same way as human auditory perception does. A group of overlapped triangular bandpass filters that imitate the Mel scale's features are employed in the computation. These filter banks are used in MFCC to obtain the Mel scale power spectrum by applying them to the speech spectrum. The output of the log-filter-bank is then obtained using the log operator. Finally, 12 MFCCs are generated using the discrete cosine transform (DCT). The log energy calculated from each speech segment is the 13th parameter. Temporal derivatives were collected as ΔMFCCs and ΔΔMFCCs to capture the dynamic information of the speech lost during frame-by-frame analysis. As a result, each speech segment's final feature vector contains 13 MFCCs, 13 ΔMFCCs, and 13 ΔΔMFCCs. Figure 8.5 depicts the ACR-MFCC Feature Extraction Algorithm. The extraction of entire spectrum and prosodic features employed in this study is shown in Figure 8.6. All of the features are subjected to the statistical functions mean and standard deviation. For emotion recognition, a total of 6 prosodic features (F0, F1, F2) and 78 spectral features (MFCC) are used.

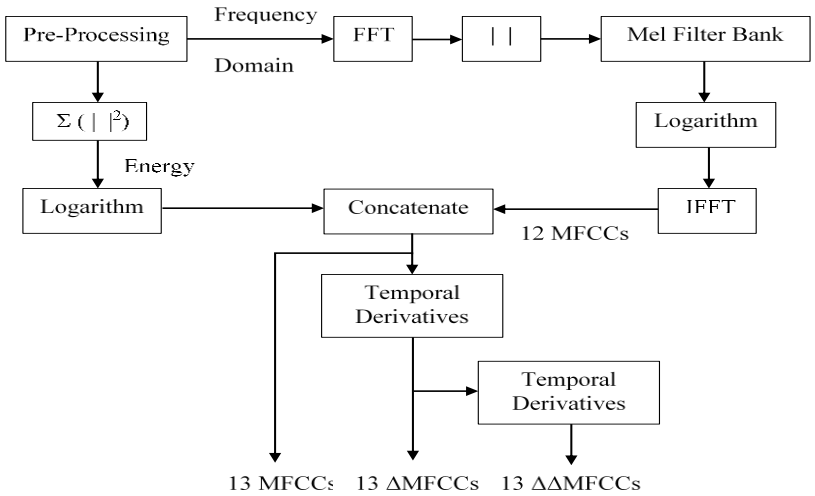


Fig. 8.5 Block Diagram of ACR-MFCC Feature Extraction Algorithm

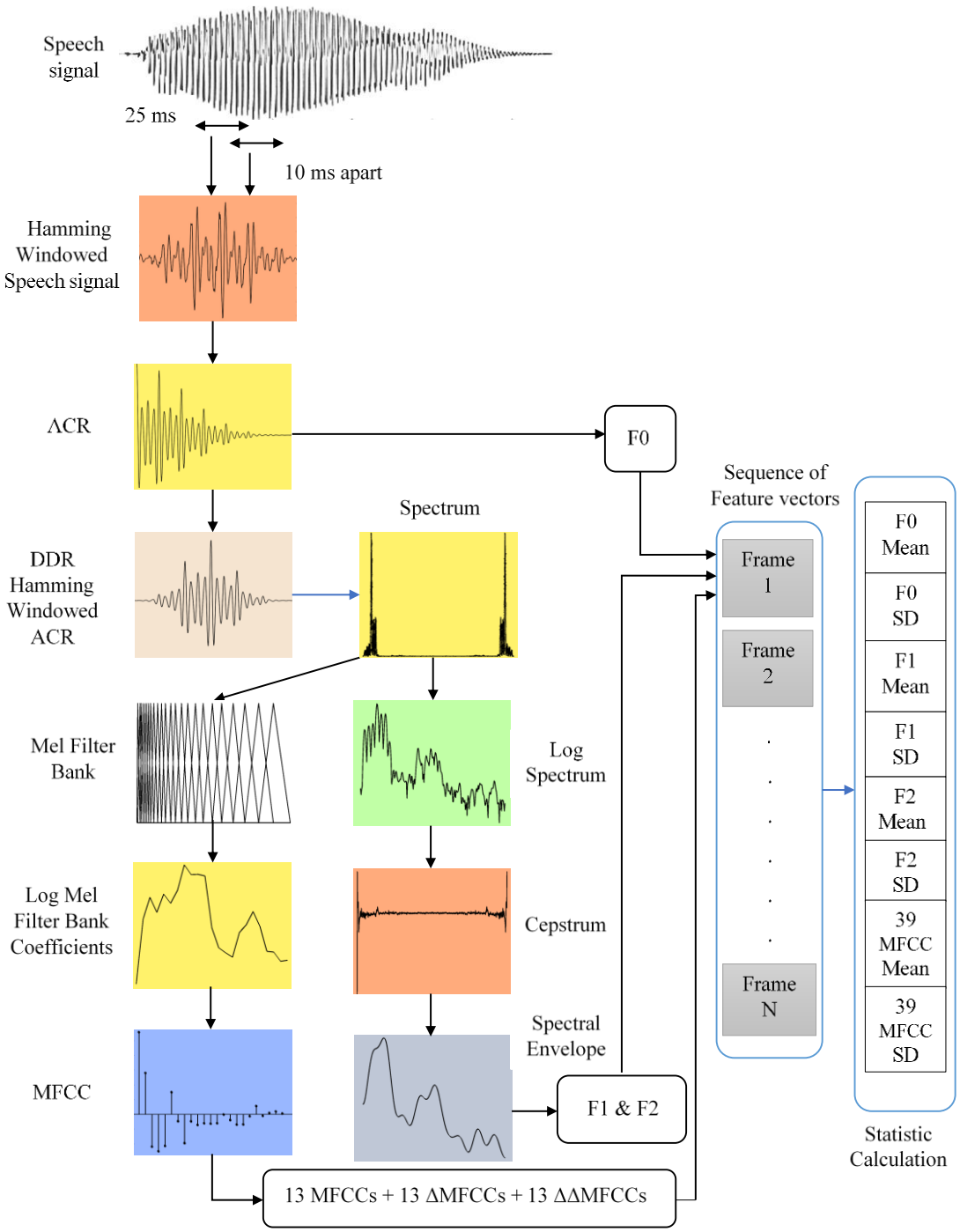


Fig. 8.6 Unified Frame work of Prosodic and Spectral Feature Extraction

8.3.2 Nonlinear and Multifractal Features

Emotion recognition features include the correlation dimension at minimum embedding dimension (D_{2m}), correlation entropy at minimum embedding dimension (K_{2m}), largest Lyapunov exponent at minimum embedding dimension (LLE), 21 Generalised Hurst exponents(h(q)), singularity spectrum width (δα) and singularity spectrum height (f(α)). All of the observed parameters are subjected to the statistical functions mean and standard deviation, and a total of 52 features are analysed. Sections 5.2.3 and 5.2.4 discuss D_{2m} and K_{2m}, respectively. Section 7.3 contains a detailed discussion of h(q), and f(α).

The sensitivity of a chaotic system to initial conditions can be measured using the Lyapunov exponent [75], [124]. Lyapunov exponents with positive values suggest a chaotic attractor. If the initial corresponds to a particular function f(x) are x_0 and $x_0+\delta_0$ (where δ_0 is extremely small), the separation after n iterations is given by

$$|\delta_n| = |\delta_0| e^{n\lambda} \tag{8.1}$$

$$\lambda = \frac{1}{n} ln \left|\frac{\delta_n}{\delta_0}\right| \tag{8.2}$$

$$\lambda = \frac{1}{n} ln \left|\frac{f^n(x_0+\delta_0) - f^n(x_0)}{\delta_0}\right| \tag{8.3}$$

When δ0 tends to zero ,the above equation becomes

$$\lambda = \frac{1}{n} ln \left|(f^n)'(x_0)\right| \tag{8.4}$$

Using chain rule

$$\lambda = \frac{1}{n} ln \left|\prod_{i=0}^{n-1} f'(x_i)\right| \tag{8.5}$$

$$\lambda = \frac{1}{n} \sum_{i=0}^{n-1} ln \left|f'(x_i)\right| \tag{8.6}$$

In the limit n→ α ,the Lyapunov exponent can be expressed as

$$\lambda = \lim_{n\to\infty} \left\{ \frac{1}{n} \sum_{i=0}^{n-1} ln\, |f'(x_i)| \right\} \tag{8.7}$$

The block diagram of nonlinear and multiracial feature extraction is shown in Fig 8.7.

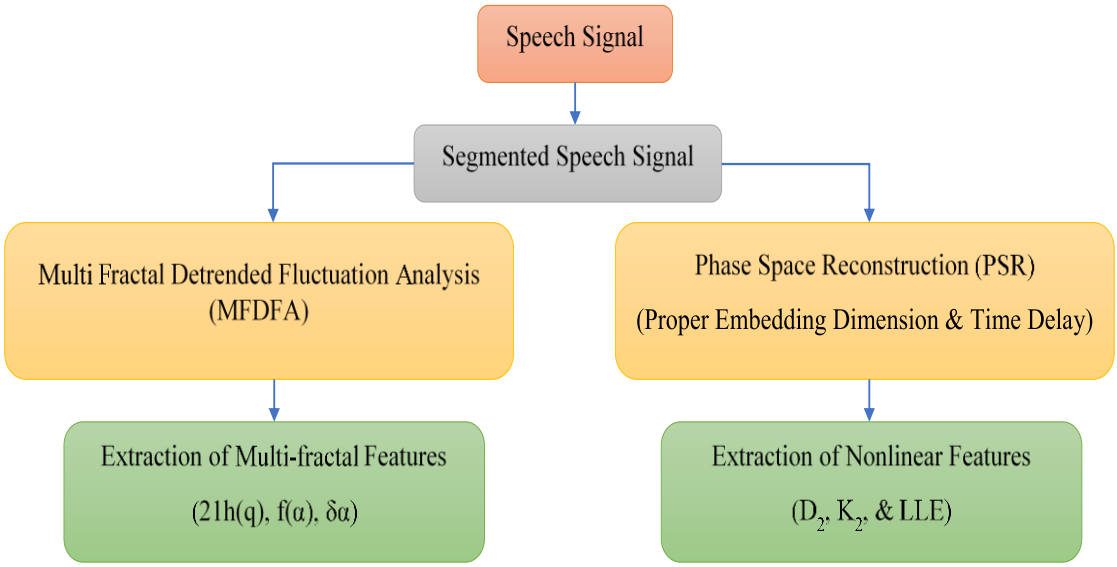


Fig. 8.7 Nonlinear and Multifractal Feature Extraction

8.4 Experiments and Results

The experiments regarding optimisation of embedding dimension, surrogate data analysis and multifractal feature extraction are discussed in the following sections. The proposed classification system and the results of SVM classifier are discussed in the succeeding sections.

8.4.1 Optimising Embedding Dimension

The time delay and embedding dimension for all the emotional signals mentioned in Table8.1 are determined by Mutual information and FNN respectively. The time delay value obtained from the mutual information method is used in FNN method. Fig 8.8 shows the variation of mutual

information with time delay for eight studied emotions, from which the first minimum gives the optimum time delay. The variation of FNN with embedding dimension is given in Fig 8.9

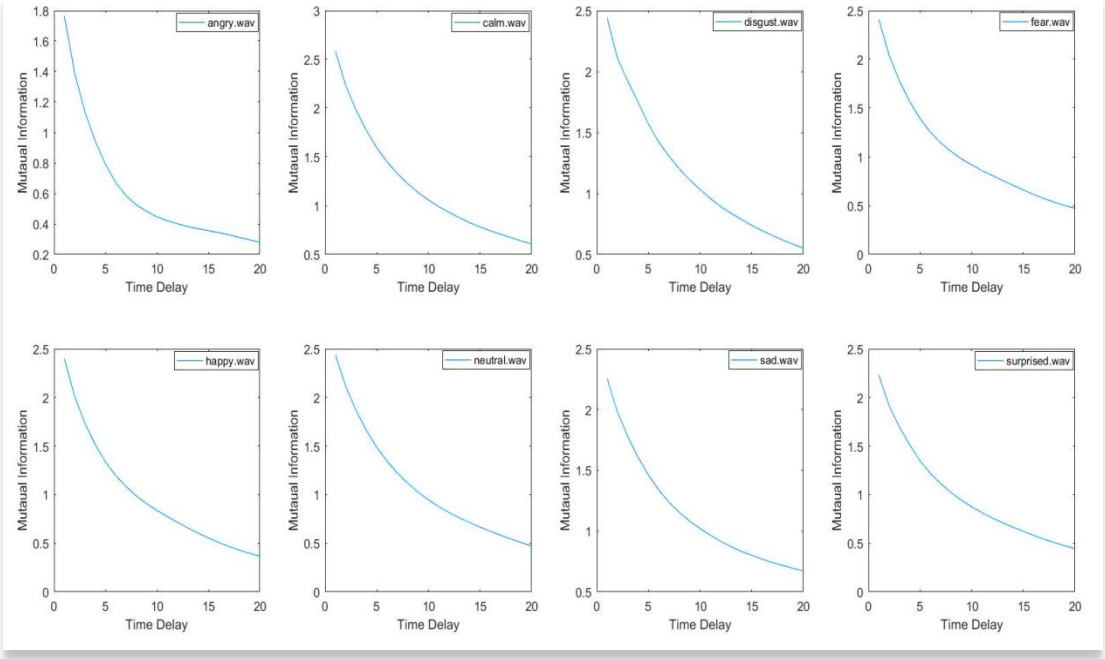


Fig. 8.8 Variation of mutual information with time delay(emotional speech signal)

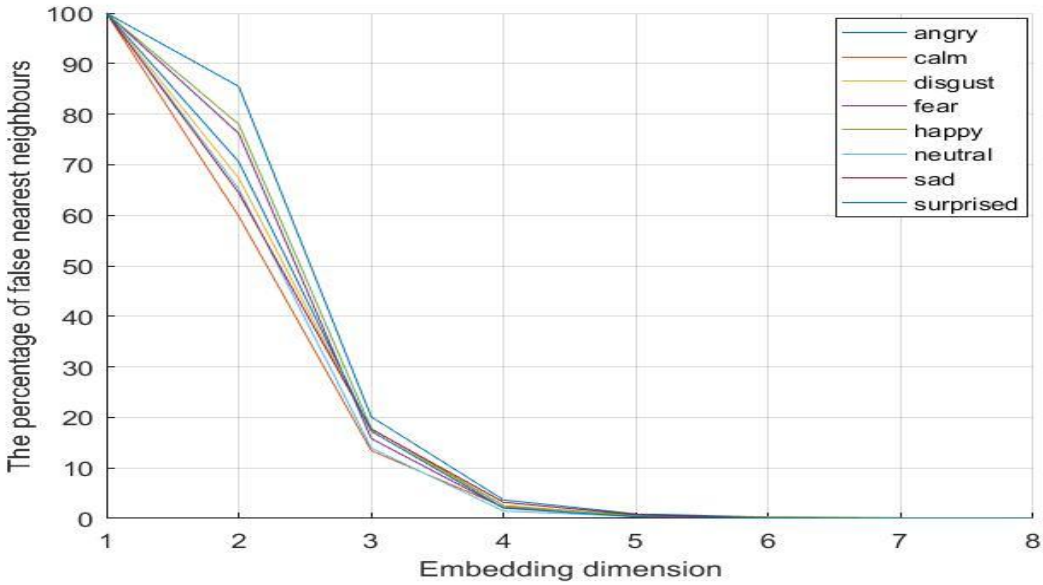


Fig. 8.9 Variation of FNN with dimension (emotional speech signal)

8.4.2 Results of Surrogate Analysis with D_{2m}, K_{2m} and LLE

Surrogate analysis has been permormed for the eight emotional speech signals by constructing 100 IAAFT surrogates.The algorithm described in section 5.3 is used for generating surrogates.Correlation dimension,correlation entropy and Largest lyapunov exponent,all estimated from the six dimensional hyper space,are used as the discreminating measures.Fig 8.10 shows the variation of D_2 with embedding dimension for the eight emotional signals and its 5 surrogates. Fig 8.11 shows the variation of K_2 with embedding dimension for the eight emotional signals and its 5 surrogates.

The statistical significance level (S) for each feature, corresponding to each emotion, with D_{2m}, K_{2m} and LLE as measure are calculated. The calculated significance level from a specific emotional sample (repeated 20 times) and its 100 surrogates,with D_{2m} as measure, is given in Table 8.2. Tables 8.3 and 8.4 demonstrate the same for K_{2m} and LLE. It is clear from the tables that the significance level is high enough to ensure the nonlinearity.

Table 8.2 Significance level with D_{2m} as nonlinear measure

Emotion	$<D_{2m}>$	$<D_{2m}>_{surr}$	σ_{surr}	Significance level (S)
Angry	2.31	4.32	0.071	28.3
Calm	2.56	4.64	0.082	25.4
Disgust	2.62	4.11	0.086	17.3
Fear	2.59	4.95	0.068	34.7
Happy	2.19	4.63	0.076	32.10
Neutral	2.33	4.76	0.066	36.8
Sad	2.81	5.01	0.080	27.5
Surprised	2.95	5.23	0.073	31.23

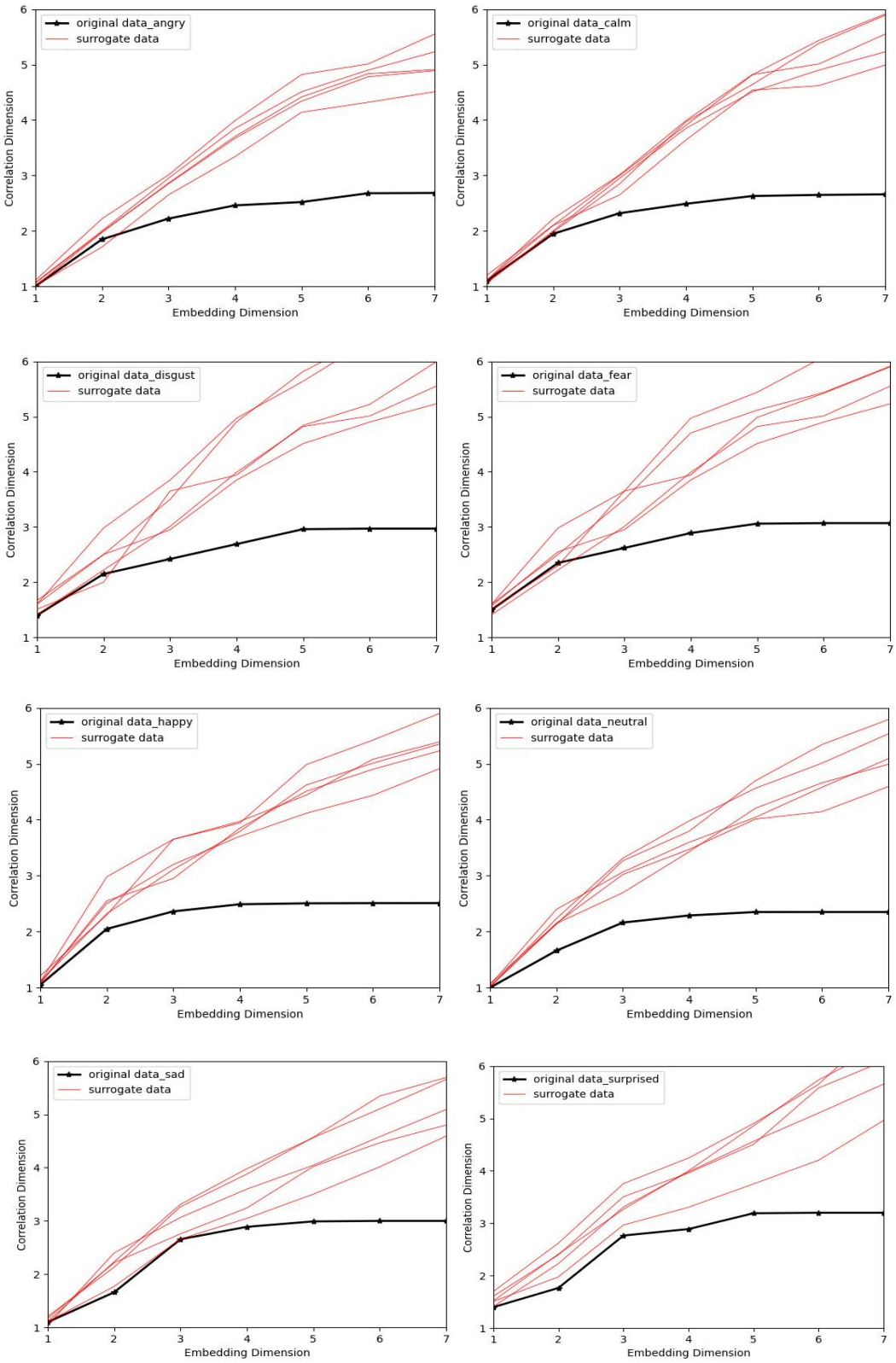


Fig. 8.10 Surrogate analysis with D_2 for emotional speech signals

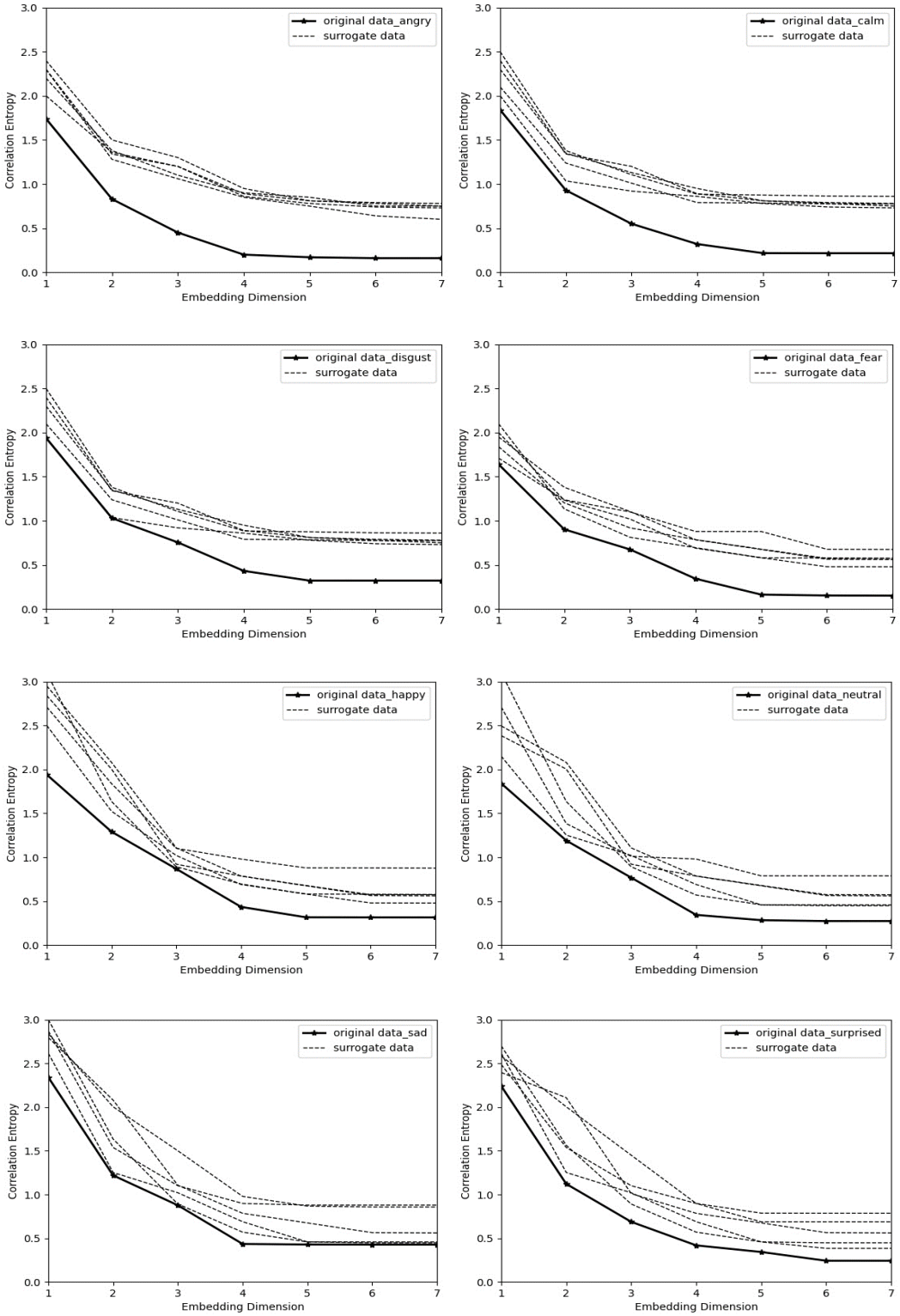


Fig. 8.11 Surrogate analysis with K_2 for emotional speech signals

Table 8.3 Significance level with K_{2m} as nonlinear measure

Emotion	$<K_{2m}>$	$<K_{2m}>_{surr}$	σ_{surr}	Significance level (S)
Angry	0.23	0.67	0.015	29.33
Calm	0.21	0.71	0.017	29.41
Disgust	0.19	0.62	0.019	22.63
Fear	0.30	0.75	0.018	28.12
Happy	0.26	0.73	0.016	29.37
Neutral	0.28	0.80	0.018	28.88
Sad	0.24	0.77	0.015	35.33
Surprised	0.29	0.76	0.016	29.37

Table 8.4 Significance level with LLE as nonlinear measure

Emotion	<LLE>	$<LLE>_{surr}$	σ_{surr}	Significance level (S)
Angry	0.38	0.71	0.014	23.57
Calm	0.41	0.73	0.016	20.00
Disgust	0.46	0.80	0.013	26.15
Fear	0.37	0.77	0.017	23.53
Happy	0.39	0.79	0.021	19.04
Neutral	0.41	0.68	0.011	24.54
Sad	0.39	0.70	0.015	20.66
Surprised	0.36	0.72	0.017	21.17

8.4.3 Multifractal Feature Extraction

The multifractal spectrum of each emotion data sample is constructed using the algorithm described in section 7.3. The ‘q’ values are varied from -10 to 10, and the corresponding 21 generalised Hurst exponents have been estimated for all samples listed in Table 3.1 (chapter 3). The height and width

of the spectrum are evaluated from the spectrum graph. Figures 8.12 to 8.19, respectively, show the multifractal analysis of eight emotions: angry, calm, disgust, fear, happy, neutral, sad, and surprised, used in this work.

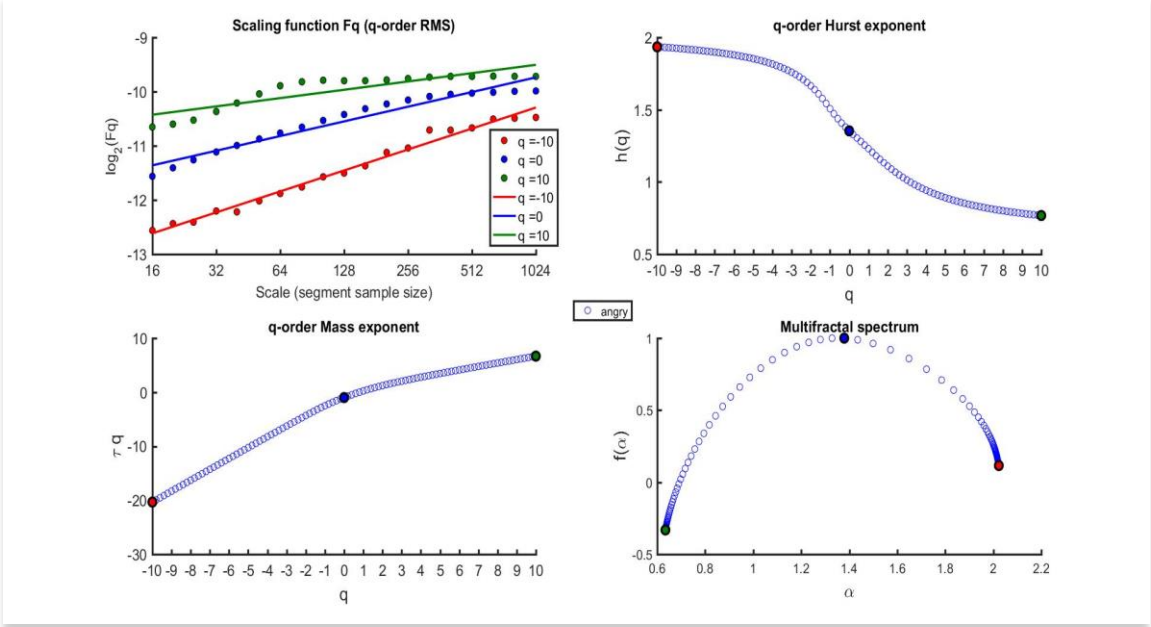


Fig. 8.12 Multifractal analysis of speech emotion (angry)

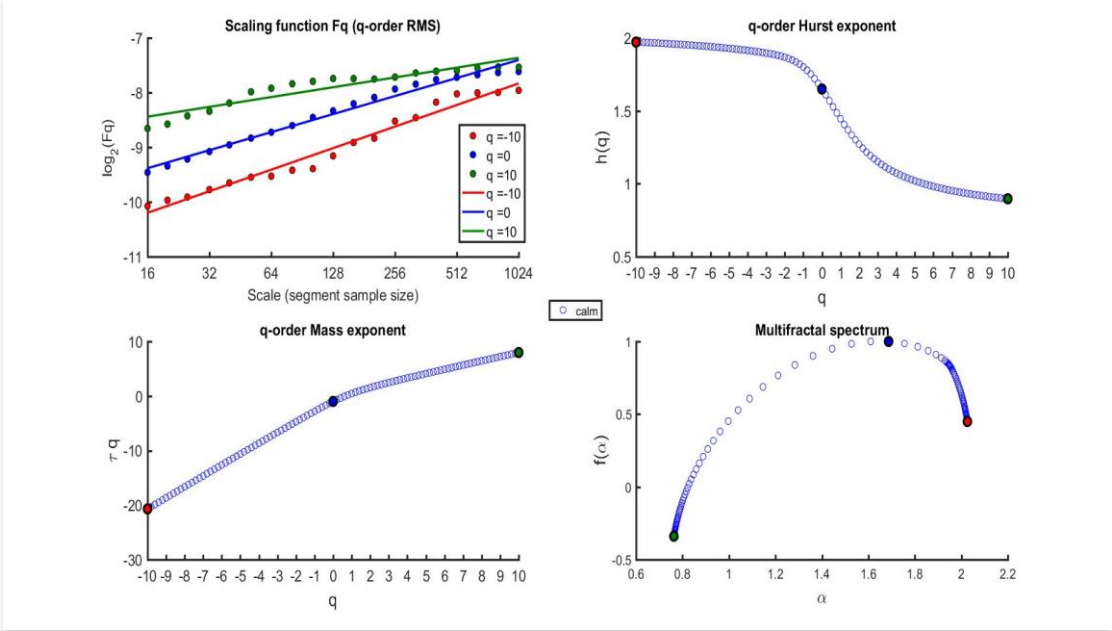


Fig. 8.13 Multifractal analysis of speech emotion (calm)

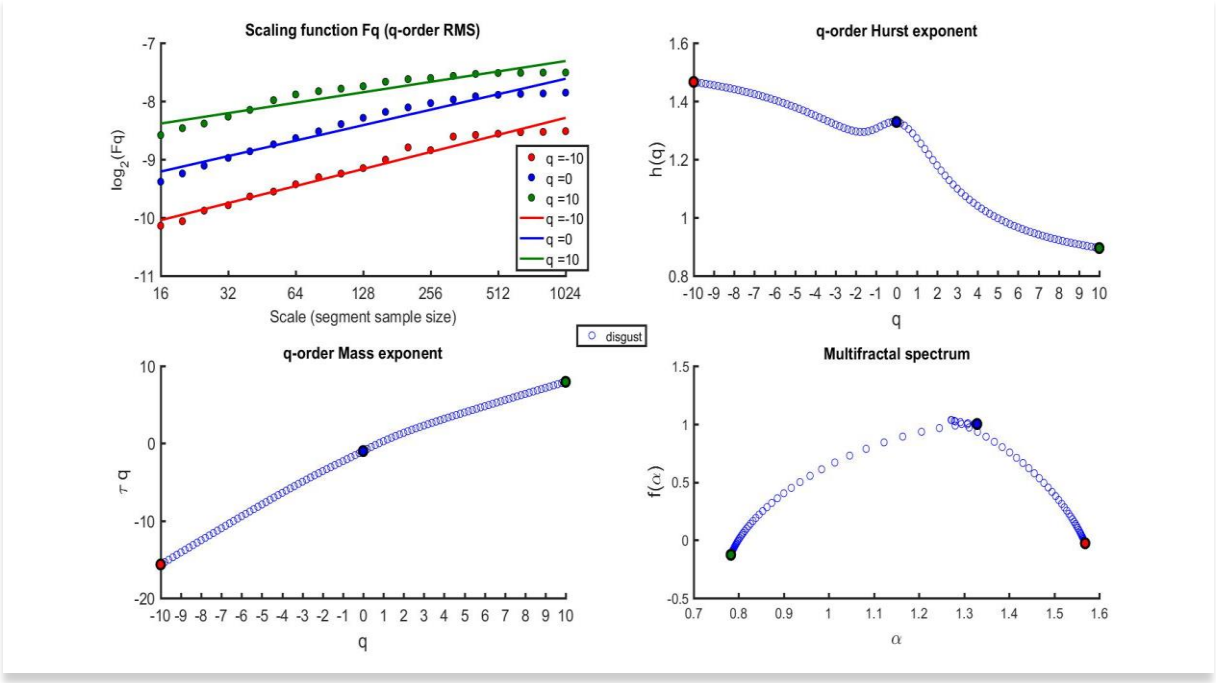


Fig. 8.14 Multifractal analysis of speech emotion (disgust)

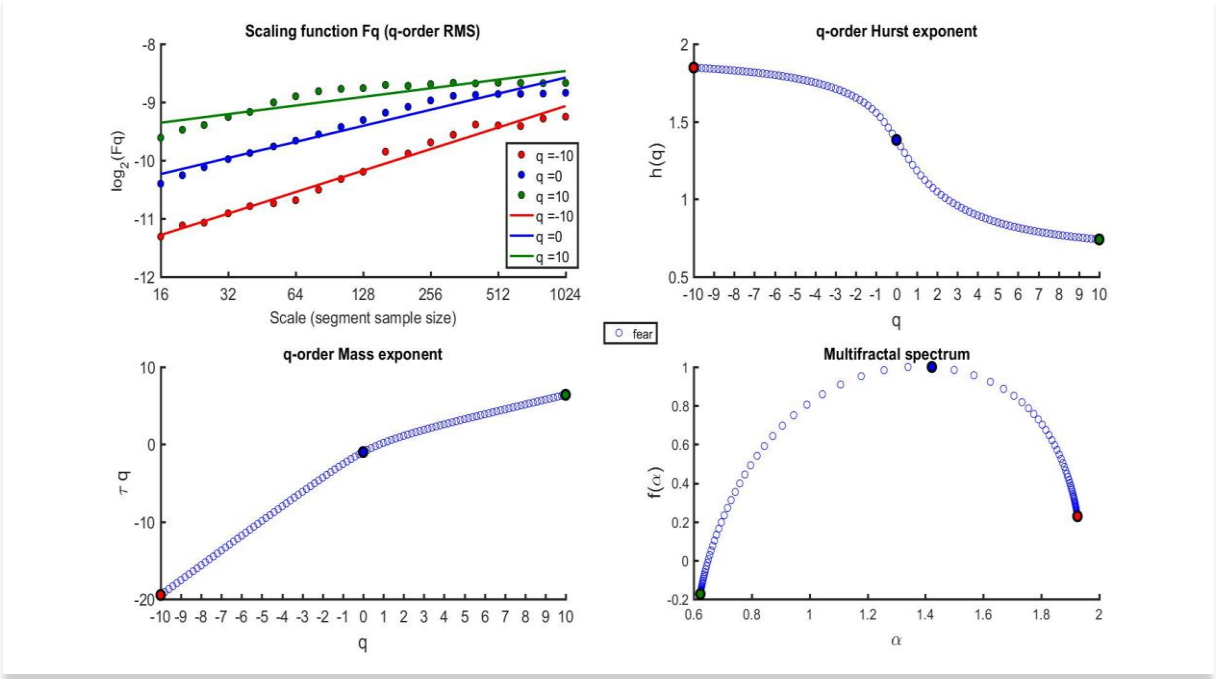


Fig. 8.15 Multifractal analysis of speech emotion (fear)

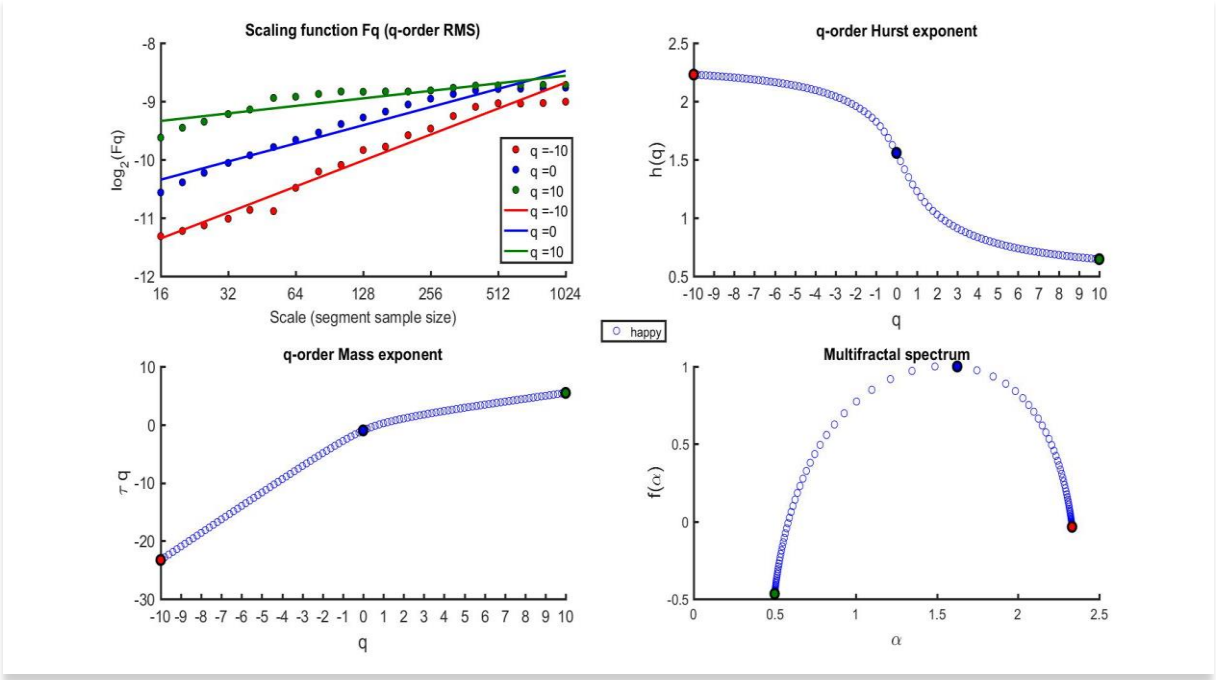


Fig. 8.16 Multifractal analysis of speech emotion (happy)

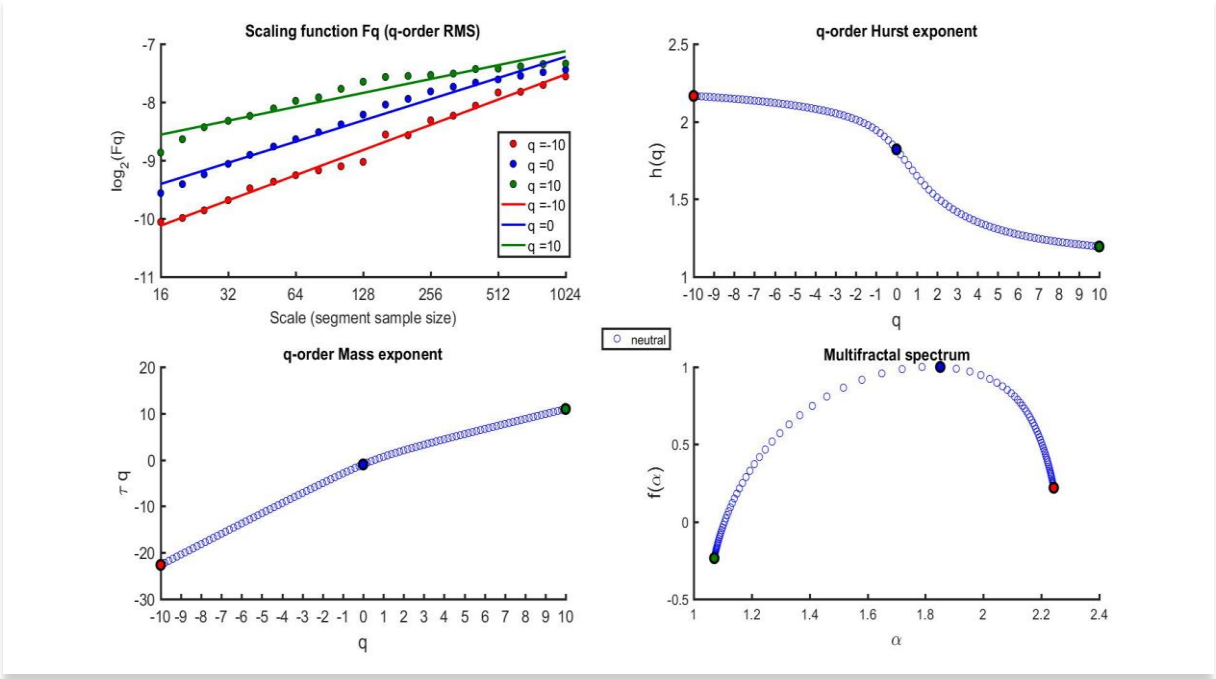


Fig. 8.17 Multifractal analysis of speech emotion (neutral)

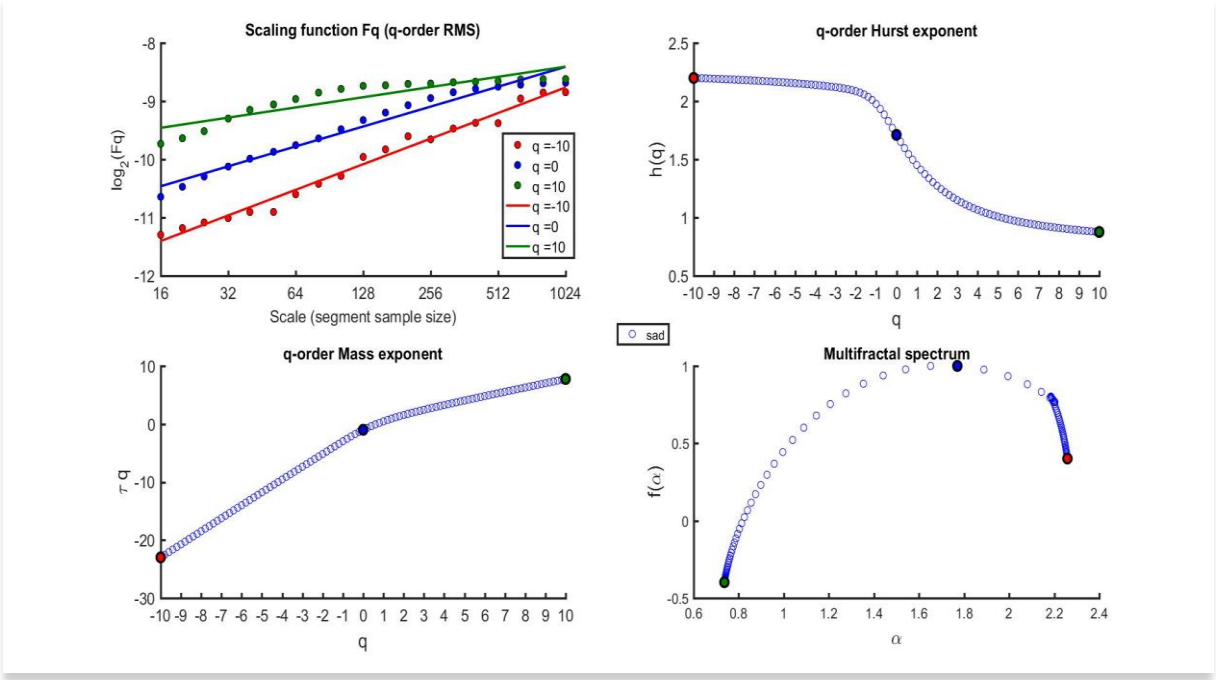


Fig. 8.18 Multifractal analysis of speech emotion (sad)

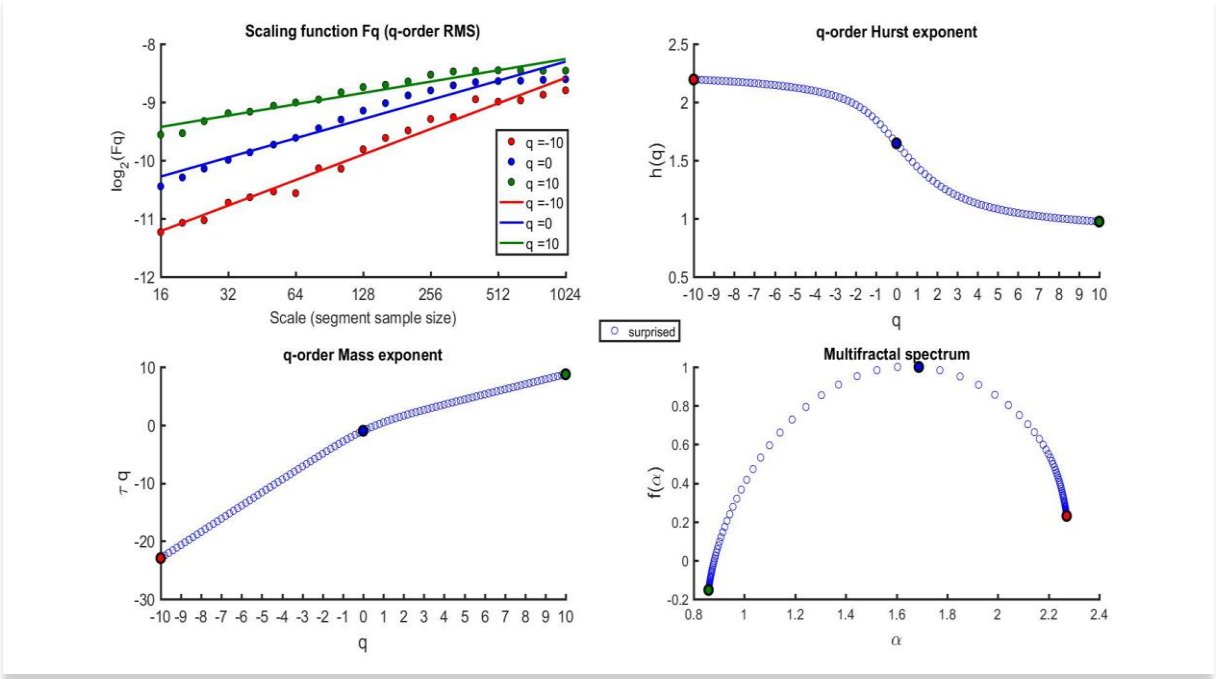


Fig. 8.19 Multifractal analysis of speech emotion (surprised)

8.4.4 Evaluation of Proposed Classification System using SVM

Prosodic, spectral, nonlinear, and multifractal characteristics are combined to create an emotion identification system. Six spectral features (mean and standard deviation of F0, F1, and F2), 78 spectral features (mean and standard deviation of 13 MFCCs, 13 MFCCs, and 13 MFCCs), six nonlinear features in the six-dimensional hyperspace (mean and standard deviation of D_2, K_2, and LLE), and 46 multifractal features are the features used for classification (mean and standard deviation of 21 generalised Hurst exponents, singularity spectrum width and singularity spectrum height). Figure 8.20 depicts the proposed classification scheme.

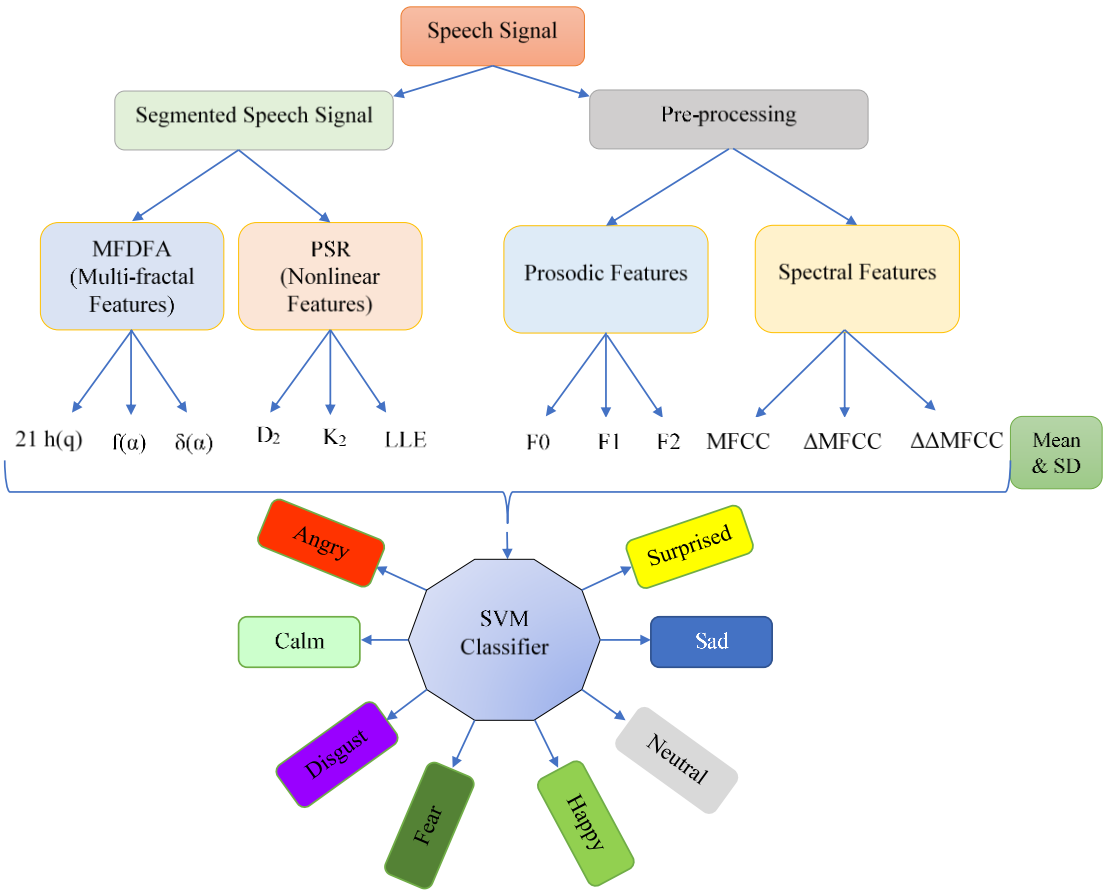


Fig. 8.20 Proposed Emotion Classification System

For classification, an SVM classifier is used, which was previously explained in section 6.4.All the features extracted from the male samples of the RAVDESS database (Table 8.1) are fed to the SVM classifier. The Gaussian radial basic kernel was proven to be the most effective for emotion categorization. The performance of spectral and prosodic characteristics was initially evaluated, and the classification confusion matrix is computed. The analysis was then conducted with nonlinear features mixed with spectral and prosodic features. Table 8.5 shows the accuracy and precision of emotion recognition together with confusion matrix, when prosodic and spectral information are combined. Table 8.6 shows the same for combined prosodic, spectral, and nonlinear properties. When all of the feature vectors are integrated for classification, the accuracy and precision are shown in Table 8.7.

From Table 8.6, it is observed that the recognition efficiency increases while combining the nonlinear features with prosodic and spectral features. The increase in accuracy in eight studied emotions, angry, calm, disgust, fear, happy, neutral, sad, and surprised, is 19.93%, 15.85%, 11.64%, 1.65%, 3.07%, 10.75%, 3.82%, and 2.4%, respectively. The precision increases by 5.41%, 7.44%, 5.55%, 9.65%, 2.86%, 12.37%, 11.35%, and 4.99%, respectively. There is a further improvement in accuracy and precision due to the clubbing of multifractal features. The accuracy of the eight mentioned emotions improves by 7.82%, 9.10%, 17.2%, 22.7%, 10.68%, 14.13%, and 4.59%, respectively. The increments in precision are 2.11%, 17.44%, 16.58%, 20.29%, 11.38%, 4.71%, 12.95%, and 12.09%, respectively. Fig. 8.21 summarise the results of classification.

Table 8.5 Accuracy and Precision for using Combined Prosodic and Spectral Features

Emotion	**Angry**	**Calm**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprised**	**Accuracy (%)**
Angry	0.76	0.05	0.06	0.12	0.21	0.00	0.08	0.10	55.07
Calm	0.00	0.68	0.05	0.05	0.00	0.18	0.07	0.03	64.15
Disgust	0.02	0.10	0.64	0.08	0.05	0.02	0.06	0.04	63.36
Fear	0.05	0.02	0.02	0.58	0.00	0.00	0.09	0.02	74.35
Happy	0.12	0.04	0.07	0.08	0.62	0.00	0.00	0.04	63.9
Neutral	0.00	0.08	0.00	0.00	0.02	0.81	0.00	0.00	80.19
Sad	0.00	0.02	0.04	0.05	0.03	0.11	0.70	0.03	71.42
Surprised	0.00	0.04	0.06	0.08	0.00	0.00	0.01	0.72	79.12
Precision (%)	80.0	66.02	68.08	55.77	66.66	72.32	69.30	72.72	

Table 8.6 Accuracy and Precision for using Combined Prosodic, Spectral and Nonlinear Features

Emotion	**Angry**	**Calm**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprised**	**Accuracy (%)**
Angry	0.75	0.04	0.03	0.05	0.08	0.00	0.05	0.00	75.00
Calm	0.00	0.72	0.04	0.03	0.04	0.00	0.02	0.06	80.00
Disgust	0.03	0.04	0.81	0.05	0.06	0.07	0.00	0.02	75.00
Fear	0.00	0.06	0.05	0.70	0.07	0.00	0.03	0.02	76.08
Happy	0.06	0.00	0.04	0.08	0.73	0.03	0.05	0.10	66.97
Neutral	0.02	0.01	0.00	0.09	0.03	0.83	0.03	0.02	82.17
Sad	0.04	0.05	0.08	0.07	0.00	0.03	0.76	0.00	75.24
Surprised	0.00	0.06	0.05	0.00	0.04	0.02	0.00	0.75	81.52
Precision (%)	83.36	73.42	73.63	65.42	69.52	84.61	80.85	77.31	

Table 8.7 Accuracy and Precision for using Combined Prosodic, Spectral, Nonlinear and Multifractal Features

Emotion	**Angry**	**Calm**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprised**	**Accuracy (%)**
Angry	0.82	0.01	0.04	0.02	0.07	0.00	0.00	0.02	82.82
Calm	0.00	0.90	0.00	0.02	0.03	0.04	0.00	0.02	89.10
Disgust	0.02	0.03	0.83	0.01	0.05	0.00	0.01	0.04	83.83
Fear	0.00	0.00	0.01	0.96	0.03	0.00	0.02	0.01	93.20
Happy	0.04	0.00	0.01	0.00	0.89	0.01	0.03	0.02	89.89
Neutral	0.03	0.04	0.00	0.01	0.00	0.91	0.02	0.00	92.85
Sad	0.00	0.00	0.03	0.04	0.02	0.01	0.85	0.00	89.47
Surprised	0.05	0.01	0.00	0.06	0.01	0.00	0.02	0.93	86.11
Precision (%)	85.41	90.90	90.21	85.71	80.90	89.40	93.80	89.40	

Fig. 8.21 Accuracy of Different Feature Vectors

8.5 Conclusion

In this study, nonlinear features and multifractal features are combined with spectral and prosodic features to recognise speaker emotion. The male sounds from the RAVDESS speech emotion database are used in this work. Nonlinear features: correlation dimension (D_{2m}), correlation entropy (K_{2m}), and largest Lyapunov exponent (LLE), all at minimum embedding dimension, are used as nonlinear features, and singularity spectrum parameters (height and width of singularity spectrum, and q-order Hurst exponents) are taken as multifractal features. Surrogate analysis is performed with D_{2m} and K_{2m} as nonlinear measures in both normal and emotional signals, and the high level of significance indicates the nonlinear structure in the signal. The suggested system's classification accuracy is assessed using a Support Vector Machine

(SVM) Classifier. As per the result, the addition of nonlinear features with spectral and prosodic features improves recognition accuracy and minimises classification ambiguity. The integration of multifractal features further improves the accuracy and precision. The nonlinearity and multifractality of the signal reflect the speaker's emotional content, making these features a supporting tool for recognising the speaker's emotional state.

CHAPTER 9

CONCLUSIONS AND FUTURE DIRECTIONS

9.1 Conclusions

Nonlinearity and multifractality of the Reconstructed Phase Space (RPS) are considered in this thesis in order to develop classification algorithms for diseased, noisy, and emotional speech signals. There are various shortcomings in previously reported speech-based applications for diseased, noisy, and emotional speech analysis, which are discussed in this thesis. As a result, better systems, based on nonlinear and multiracial features, are proposed for the detection of pathology, noise, and emotional expression.

The availability of a phonetically balanced audio-visual speech database in the language in which the application is to be used is the foundation of any speech-based application. A new Malayalam audio speech database has been developed and presented. This collection contains 50 isolated Malayalam phonemes and 207 related words that comprise all allophonic variations. The background noise is removed using the spectral subtraction method. The database is created in open and closed modes. The database was segmented and labelled. The aforementioned database can be utilised to enhance research in a variety of speech-based signal studies. The database is used in this thesis to investigate the nonlinearity of the reconstructed hyperspace of the speech production system.

The developed database is utilised to optimise time delay and embedding dimension of the chaotic attractor of the RPS. The hypothetical abstract space that represents the system under investigation will help analyse its dynamics. The subject of optimizing the embedding dimension is studied

using Lorenz and Rossler systems as models. The time delay of embedding was determined using the mutual information method, and it varied between samples. The embedding dimension for Lorenz and Rossler systems is three, confirming the applicability of FNN and PCA. The embedding dimension of the speech production system is unaffected by age, gender, or sample frequency in Malayalam phoneme time series. The tested samples' mode is six, with a mean close to it. In this case, the standard deviation is so small that the mode value can be used. It is possible to analyse and model the speech production system using a six-dimensional hyperspace reconstruction.

For ensuring the underlying nonlinear structure in the signal a surrogate analysis was performed at the optimised delay and embedding dimension. While comparing the statistical significance level of Malayalam phoneme time series with standard Lorenz and Rossler systems, the significance level of different phonemes was found to be different but comparable. The significance level for D_{2m} and K_{2m} analysis shows that the values are closer to those of standard systems for vowels അ /a/, ഇ /i/, ഏ /e/ and all the analysed syllables. The levels for vowels ഒ /o/and ഉ /u/are comparatively smaller and it is better to avoid these sounds in the nonlinearity studies of the system. It can be concluded that inherent nonlinearity exists in speech production, and the system is time-variant. The amount of nonlinearity is different while uttering different phonemes and the system's nonlinearity is greater while uttering syllables. Thus, D_{2m} and K_{2m} can be used as better tools for the study of nonlinear dynamical structures in the speech production system, emotion recognition, and pathological analysis in place of saturated values of D_2 and K_2.

After ensuring the nonlinearity in the speech production system, the nonlinear features are utilized in distinguishing pathological voice signals from healthy voice signals. The features used are correlation dimension at

minimum embedding dimension, correlation entropy at minimum embedding dimension, and four fitting coefficients of the f(α) spectrum of strange attractor. The study relied on the VOICE database. FNN and MI have optimized the embedding dimension and time delay of RPS. The data was subjected to a statistical surrogate analysis to ensure that the characteristics used in the analysis were discriminated, and a reasonable significance level indicated the presence of nonlinearity. Based on the measures examined, a classification system is proposed. SVM was used to assess the performance of the proposed classification system in distinguishing between pathological and normal voices. The precision is 99%, and the accuracy is 97%. When compared to recognition algorithms based on linear feature vectors and other nonlinear parameters, this accuracy is promising.

The multifractal features derived from the multifractal detrended fluctuation analysis (MFDFA) is used for noise identification. The singularity spectrum width and extremal Holder exponents are seemed to be reduced in the MFDFA of the voice samples due to additive noise. The Malayalam speech database developed together with its noise simulated signals are utilised to distinguish pink, red, and white noises. The SNR has an effect on the reduction, and the SNR rate can be computed by multiplying the percentage reduction in the parameters. The noise categories are recognised using feature vectors and an SVM classifier, and the accuracy attained indicates that multifractal features are an efficient tool for recognising different types of noise.

Finally, nonlinear features and multifractal features are combined with spectral and prosodic features to recognise speaker emotion. The male sounds from the RAVDESS speech emotion database are used. Fundamental frequency (F0), formant frequencies (F1 and F2) and Mel frequency cepstral coefficients (MFCCs) are the most popular prosodic and spectral features.

Instead of using these features directly, its noise-tolerant version was proposed by using the autocorrelation function(F0 estimation from ACR, F1 and F2 estimation from ACR Cepstrum and ACR MFCC). Correlation dimension (D_{2m}), correlation entropy (K_{2m}), and largest Lyapunov exponent (LLE), all at minimum embedding dimension, are used as nonlinear features, and singularity spectrum parameters (height and width of singularity spectrum, and q-order Hurst exponents) are taken as multifractal features. Surrogate analysis is performed with D_{2m} and K_{2m} as nonlinear measures in both normal and emotional signals, and the high level of significance indicates the nonlinear structure in the signal. The suggested system's classification accuracy is assessed using a Support Vector Machine (SVM) Classifier. As per the result, the addition of nonlinear features with spectral and prosodic features improves recognition accuracy and minimises classification ambiguity. The integration of multifractal features further improves the accuracy and precision. The nonlinearity and multifractality of the signal reflect in the speaker's emotional content, which makes these features a supporting tool for recognising the speaker's emotional state.

By studying and analyzing the nonlinear and multifractal features the pathological, noisy and emotional content in speech signal can be captured and it will throw light on the inherent dynamics of the speech production system. These features not only provide high accuracies for pathology detection and excellent noise identification performance but also help in identifying emotional cues in speech.

9.2 Future Research Directions

Despite the positive conclusions of this research, certain areas remain unexplored or require further research. Listed below are a few.

- The proposed database should include continuous speech while capturing all Malayalam dialects. Extend the existing database and make it searchable online.
- Nonlinear features and multifractal features can be used to detect the pathological level of patients, and the stage of the disease can be identified from these parameters. Further work should be done in this direction. Also, the study can be extended to all available databases for different types of pathologies.
- When the SNR levels vary considerably, the identification task becomes difficult. Studies should be conducted to identify the noise type irrespective of SNR.
- The female part of the RAVDESS database is not explored in this study.By using it, the work towards a gender-independent speech emotion system can be constructed.
- The study of emotions in languages like Malayalam is very poor. Hence, a speech emotion database can be constructed to study the linguistic dependence of the approach.

REFERENCES

[1] R. K. Moore, "Spoken language processing: Piecing together the puzzle," *Speech Commun.*, vol. 49, no. 5, pp. 418–435, 2007, doi: 10.1016/j.specom.2007.01.011.

[2] V. Delić *et al.*, "Speech technology progress based on new machine learning paradigm," *Comput. Intell. Neurosci.*, vol. 2019, 2019, doi: 10.1155/2019/4368036.

[3] E. Bradley and H. Kantz, "Nonlinear time-series analysis revisited," 1981.

[4] D. Maria, "Nonlinearity Framework in Speech," pp. 11–26, 2012, doi: 10.1007/978-1-4614-1505-3.

[5] A. Mencattini *et al.*, "Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure," *Knowledge-Based Syst.*, vol. 63, pp. 68–81, 2014, doi: 10.1016/j.knosys.2014.03.019.

[6] P. Birkholz, L. Martin, Y. Xu, S. Scherbaum, and C. Neuschaefer-Rube, "Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis," *Comput. Speech Lang.*, vol. 41, pp. 116–127, 2017, doi: 10.1016/j.csl.2016.06.004.

[7] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, "Automatic Voice Pathology Detection With Running Speech by Using Estimation of Auditory Spectrum and Cepstral Coefficients Based on the All-Pole Model," *J. Voice*, vol. 30, no. 6, pp. 757.e7-757.e19, 2016, doi: 10.1016/j.jvoice.2015.08.010.

[8] J. C. Wang, C. Y. Wang, Y. H. Chin, Y. T. Liu, E. T. Chen, and P. C. Chang, "Spectral-temporal receptive fields and MFCC balanced feature extraction for robust speaker recognition," *Multimed. Tools Appl.*, vol. 76, no. 3, pp. 4055–4068, 2017, doi: 10.1007/s11042-016-3335-0.

[9] M. Sarria-Paja and T. H. Falk, "Fusion of auditory inspired amplitude modulation spectrum and cepstral features for whispered and normal speech speaker verification," *Comput. Speech Lang.*, vol. 45, pp. 437–456, 2017, doi: 10.1016/j.csl.2017.04.004.

[10] J. R. Movellan, "Visual Speech Recognition with Stochastic Networks," *Adv. Neural Inf. Process. Syst. 7*, pp. 851–858, 1995.

[11] S. P.-L. Vandendorpe, "The M2VTS Multimodal Face Database," *First Int. Conf. Audio-and Video-based Biometric Pers. Authentication*, pp. 403–409, 1997.

[12] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB : The Extended M2VTS Database University of Surrey 1 Introduction 2 Database Speci cation 3 The Database Acquisition System," *Proc. Second Int. Conf. audio video-based biometric Pers. authentication*, no. April 2016, pp. 1–6, 1999.

[13] D. H. Brooks, E. L. Miller, C. A. Dimarzio, M. Kilmer, and R. J. Gaudette, "Audiovisual Speech Processing-Lip Reading and Lip Synchronization-IEEE-2001," no. November, pp. 57–75, 2001.

[14] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, 2002, doi: 10.1109/34.982900.

[15] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Cuave: A

new audio-visual database for multimodal human-computer interface research," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2, 2002.

[16] C. Sanderson and K. K. Paliwal, "The vidtimit database," *Idiap Commun.*, vol. 06, pp. 2–6, 2002.

[17] A. G. Chițu and L. J. M. Rothkrantz, "Building a data corpus for audio-visual speech recognition," *EUROMEDIA 2007 - 13th Annu. Sci. Conf. Web Technol. New Media Commun. Telemat. Theory Methods, Tools Appl. D-TV*, vol. 1, no. Movellan 1995, pp. 88–92, 2007.

[18] "The BANCA-Database and Evaluation Protocol-2003."

[19] T. J. Hazen, K. Saenko, C. H. La, and J. R. Glass, "A segment-based Audio-Visual speech recognizer: Data collection, development, and initial experiments," *ICMI'04 - Sixth Int. Conf. Multimodal Interfaces*, pp. 235–242, 2004.

[20] R. Göcke and J. B. Millar, "A Detailed Description of the AVOZES Data Corpus," *Technology*, pp. 486–491, 2004.

[21] L. Liang, Y. Luo, F. Huang, and A. V. Nefian, "A multi-stream audio-video large-vocabulary Mandarin Chinese speech database," *2004 IEEE Int. Conf. Multimed. Expo*, vol. 3, pp. 1787–1790, 2004.

[22] N. A. Fox, B. A. O'Mullane, and R. B. Reilly, "VALID: A new practical audio-visual database, and comparative results," *Lect. Notes Comput. Sci.*, vol. 3546, pp. 777–786, 2005.

[23] P. Císař, J. Zelinka, M. Železný, A. A. Karpov, and A. L. Ronzhin, "Audio-Visual Speech Recognition for Slavonic Languages (Czech and Russian) Department of Cybernetics , University of West Bohemia in

Pilsen (UWB), Czech Republic Speech Informatics Group , St . Petersburg Institute for Informatics and Automation of th," *Specom*, no. June, pp. 493–498, 2006.

[24] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, 2006, doi: 10.1121/1.2229005.

[25] J. Trojanová, M. Hrúz, P. Campr, and M. Železný, "Design and recording of Czech audio-visual database with Impaired conditions for continuous speech recognition," *Proc. 6th Int. Conf. Lang. Resour. Eval. Lr. 2008*, pp. 1239–1243, 2008.

[26] A. Vorwerk, X. Wang, D. Kolossa, S. Zeiler, and R. Orglmeister, "WAPUSK20 - A database for robust audiovisual speech recognition," *Proc. 7th Int. Conf. Lang. Resour. Eval. Lr. 2010*, pp. 3016–3019, 2010.

[27] A. Bastanfard, M. Fazel, and A. A. Kelishami, "The Persian Linguistic Based Audio-Visual Data Corpus , AVA II , Considering Coarticulation," pp. 284–294.

[28] Y. Benezeth and G. Bachman, "BL-Database: A French audiovisual database for speech driven lip animation systems," no. August, 2011.

[29] Y. W. Wong *et al.*, "A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities," *Pattern Recognit. Lett.*, vol. 32, no. 13, pp. 1503–1510, 2011, doi: 10.1016/j.patrec.2011.06.011.

[30] C. McCool *et al.*, "Bi-modal person recognition on a mobile phone: Using mobile phone data," *Proc. 2012 IEEE Int. Conf. Multimed. Expo Work. ICMEW 2012*, no. July 2012, pp. 635–640, 2012, doi:

10.1109/ICMEW.2012.116.

[31] S. Antar, A. Sagheer, S. Aly, and M. F. Tolba, "AVAS: Speech database for multimodal recognition applications," *13th Int. Conf. Hybrid Intell. Syst. HIS 2013*, pp. 123–128, 2014, doi: 10.1109/HIS.2013.6920467.

[32] A. Biswas and M. Chandra, "Audio Visual Isolated Oriya Digit Recognition Using HMM and DWT," *Conf. Adv. Commun. Control Syst. 2013*, no. April, pp. 234–238, 2013.

[33] P. Żelasko, B. Ziółko, T. Jadczyk, and D. Skurzok, "AGH corpus of Polish speech," *Lang. Resour. Eval.*, vol. 50, no. 3, pp. 585–601, 2016, doi: 10.1007/s10579-015-9302-y.

[34] I. Anina, Z. Zhou, G. Zhao, and M. Pietikainen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," *2015 11th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2015*, 2015, doi: 10.1109/FG.2015.7163155.

[35] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Trans. Multimed.*, vol. 17, no. 5, pp. 603–615, 2015, doi: 10.1109/TMM.2015.2407694.

[36] P. Upadhyaya, O. Farooq, M. R. Abidi, and P. Varshney, "Comparative Study of Visual Feature for Bimodal Hindi Speech Recognition," *Arch. Acoust.*, vol. 40, no. 4, pp. 609–619, 2015, doi: 10.1515/aoa-2015-0061.

[37] M. R. A. R. Maulana and M. I. Fanany, "An Audio-Visual Corpus for Multimodal Automatic Speech Recognition-2017," *2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017*, vol. 2018-Janua, pp. 381–385, 2018, doi: 10.1109/ICACSIS.2017.8355062.

[38] M. R. A. R. Maulana and M. I. Fanany, "Indonesian audio-visual speech corpus for multimodal automatic speech recognition," *2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017*, vol. 2018-Janua, pp. 381–385, 2018, doi: 10.1109/ICACSIS.2017.8355062.

[39] A. H. Abdelaziz, "NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2017-Augus, pp. 3752–3756, 2017, doi: 10.21437/Interspeech.2017-860.

[40] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, "A corpus of audio-visual Lombard speech with frontal and profile views," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. EL523–EL529, 2018, doi: 10.1121/1.5042758.

[41] L. R. Kishline, S. W. Colburn, and P. W. Robinson, "A multimedia speech corpus for audio visual research in virtual reality (L)," *J. Acoust. Soc. Am.*, vol. 148, no. 2, pp. 492–495, 2020, doi: 10.1121/10.0001670.

[42] A. Kashevnik *et al.*, "Multimodal Corpus Design for Audio-Visual Speech Recognition in Vehicle Cabin," vol. 9, 2021, doi: 10.1109/ACCESS.2021.3062752.

[43] D. G. Childers, K. S. Bae, and a Datu, "Detection of Laryngeal Function Using Speech and Elec trog lo ttog raphic Data," *IEEE Trans. Biomed. Eng.*, vol. 39, no. 9104288, pp. 19–25, 1992.

[44] D. A. Cairns, J. H. L. Hansen, and J. E. Riski, "A noninvasive technique for detecting hypernasal speech using a nonlinear operator," *IEEE Trans. Biomed. Eng.*, vol. 43, no. 1, pp. 35–45, 1996, doi: 10.1109/10.477699.

[45] A. P. Accardo and E. Mumolo, "An algorithm for the automatic

differentiation between the speech of normals and patients with friedreich's ataxia based on the short-time fractal dimension," *Comput. Biol. Med.*, vol. 28, no. 1, pp. 75–89, 1998, doi: 10.1016/S0010-4825(97)00039-5.

[46] V. Parsa and D. G. Jamieson, "Identification of Pathological," pp. 469–486, 2000.

[47] S. Hadjitodorov and P. Mitev, "A computer system for acoustic analysis of pathological voices and laryngeal diseases screening," *Med. Eng. Phys.*, vol. 24, no. 6, pp. 419–429, 2002, doi: 10.1016/S1350-4533(02)00031-0.

[48] M. D. O. Rosa, J. C. Pereira, and M. Grellet, "Pathology Diagnosis," *Pathology*, vol. 47, no. 1, pp. 96–104, 2000.

[49] C. R. Watts, R. Clark, and S. Early, "Acoustic measures of phonatory improvement secondary to treatment by oral corticosteroids in a professional singer: A case report," *J. Voice*, vol. 15, no. 1, pp. 115–121, 2001, doi: 10.1016/S0892-1997(01)00011-X.

[50] R. C. Guido *et al.*, "Support vector machines and wavelets for voice disorder sorting," *Proc. Annu. Southeast. Symp. Syst. Theory*, vol. 2006, pp. 434–438, 2006, doi: 10.1109/ssst.2006.1619117.

[51] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, "Discrimination of pathological voices using a time-frequency approach," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 3, pp. 421–430, 2005, doi: 10.1109/TBME.2004.842962.

[52] N. Huang, Y. Zhang, W. Calawerts, and J. J. Jiang, "Optimized Nonlinear Dynamic Analysis of Pathologic Voices With Laryngeal Paralysis Based on the Minimum Embedding Dimension," *J. Voice*,

vol. 31, no. 2, pp. 249.e1-249.e7, 2017, doi: 10.1016/j.jvoice.2016.07.021.

[53] K. Szklanny and P. Wrzeciono, "The Application of a Genetic Algorithm in the Noninvasive Assessment of Vocal Nodules in Children," *IEEE Access*, vol. 7, pp. 44966–44976, 2019, doi: 10.1109/ACCESS.2019.2908313.

[54] P. Gómez-Vilda *et al.*, "Evaluation of Voice Pathology Based on the Estimation of Vocal Fold Biomechanical Parameters," *J. Voice*, vol. 21, no. 4, pp. 450–476, 2007, doi: 10.1016/j.jvoice.2006.01.008.

[55] P. Gómez-Vilda *et al.*, "Characterizing Neurological Disease from Voice Quality Biomechanical Analysis," *Cognit. Comput.*, vol. 5, no. 4, pp. 399–425, 2013, doi: 10.1007/s12559-013-9207-2.

[56] Y. Zhang and J. J. Jiang, "Acoustic Analyses of Sustained and Running Voices From Patients With Laryngeal Pathologies," *J. Voice*, vol. 22, no. 1, pp. 1–9, 2008, doi: 10.1016/j.jvoice.2006.08.003.

[57] A. I. R. Fontes, P. T. V. Souza, A. D. D. Neto, A. D. M. Martins, and L. F. Q. Silveira, "Classification system of pathological voices using correntropy," *Math. Probl. Eng.*, vol. 2014, 2014, doi: 10.1155/2014/924786.

[58] L. Gavidia-Ceballos and J. H. L. Hansen, "Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection," *IEEE Trans. Biomed. Eng.*, vol. 43, no. 4, pp. 373–383, 1996, doi: 10.1109/10.486257.

[59] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "An improved method for voice pathology detection by means of a HMM-based feature space

transformation," *Pattern Recognit.*, vol. 43, no. 9, pp. 3100–3112, 2010, doi: 10.1016/j.patcog.2010.03.019.

[60] G. Muhammad *et al.*, "Automatic voice pathology detection and classification using vocal tract area irregularity," *Biocybern. Biomed. Eng.*, vol. 36, no. 2, pp. 309–317, 2016, doi: 10.1016/j.bbe.2016.01.004.

[61] Z. Ali, G. Muhammad, and M. F. Alhamid, "An Automatic Health Monitoring System for Patients Suffering from Voice Complications in Smart Cities," *IEEE Access*, vol. 5, no. c, pp. 3900–3908, 2017, doi: 10.1109/ACCESS.2017.2680467.

[62] R. Behroozmand and F. Almasganj, "Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis," *Comput. Biol. Med.*, vol. 37, no. 4, pp. 474–485, 2007, doi: 10.1016/j.compbiomed.2006.08.016.

[63] M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 7, pp. 1938–1948, 2011, doi: 10.1109/TASL.2010.2104141.

[64] N. Erfanian Saeedi, F. Almasganj, and F. Torabinejad, "Support vector wavelet adaptation for pathological voice assessment," *Comput. Biol. Med.*, vol. 41, no. 9, pp. 822–828, 2011, doi: 10.1016/j.compbiomed.2011.06.019.

[65] M. K. Arjmandi and M. Pooyan, "An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine,"

Biomed. Signal Process. Control, vol. 7, no. 1, pp. 3–19, 2012, doi: 10.1016/j.bspc.2011.03.010.

[66] V. Uloza *et al.*, "Categorizing normal and pathological voices: Automated and perceptual categorization," *J. Voice*, vol. 25, no. 6, pp. 700–708, 2011, doi: 10.1016/j.jvoice.2010.04.009.

[67] G. Muhammad and M. Melhem, "Pathological voice detection and binary classification using MPEG-7 audio features," *Biomed. Signal Process. Control*, vol. 11, no. 1, pp. 1–9, 2014, doi: 10.1016/j.bspc.2014.02.001.

[68] P. Saidi and F. Almasganj, "Voice Disorder Signal Classification Using M-Band Wavelets and Support Vector Machine," *Circuits, Syst. Signal Process.*, vol. 34, no. 8, pp. 2727–2738, 2015, doi: 10.1007/s00034-014-9927-x.

[69] C. M. Travieso, J. B. Alonso, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, E. Nöth, and A. G. Ravelo-García, "Detection of different voice diseases based on the nonlinear characterization of speech signals," *Expert Syst. Appl.*, vol. 82, pp. 184–195, 2017, doi: 10.1016/j.eswa.2017.04.012.

[70] A. Benba, A. Jilbab, and A. Hammouch, "Discriminating Between Patients With Parkinson's and Neurological Diseases Using Cepstral Analysis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 10, pp. 1100–1108, 2016, doi: 10.1109/TNSRE.2016.2533582.

[71] A. Al-nasheri, G. Muhammad, M. Alsulaiman, and Z. Ali, "Investigation of Voice Pathology Detection and Classification on Different Frequency Regions Using Correlation Functions," *J. Voice*, vol. 31, no. 1, pp. 3–15, 2017, doi: 10.1016/j.jvoice.2016.01.014.

[72] G. Vaziri, F. Almasganj, and R. Behroozmand, "Pathological assessment of patients' speech signals using nonlinear dynamical analysis," *Comput. Biol. Med.*, vol. 40, no. 1, pp. 54–63, 2010, doi: 10.1016/j.compbiomed.2009.10.011.

[73] A. Behrman and R. J. Baken, "Correlation dimension of electroglottographic data from healthy and pathologic subjects," *J. Acoust. Soc. Am.*, vol. 102, no. 4, pp. 2371–2379, 1997, doi: 10.1121/1.419621.

[74] I. Hertrich, W. Lutzenberger, S. Spieker, and H. Ackermann, "Fractal dimension of sustained vowel productions in neurological dysphonias: An acoustic and electroglottographic analysis," *J. Acoust. Soc. Am.*, vol. 102, no. 1, pp. 652–654, 1997, doi: 10.1121/1.419711.

[75] A. Giovanni, M. Ouaknine, and J. M. Triglia, "Determination of largest Lyapunov exponents of vocal signal: Application to unilateral laryngeal paralysis," *J. Voice*, vol. 13, no. 3, pp. 341–354, 1999, doi: 10.1016/S0892-1997(99)80040-X.

[76] Y. Zhang, C. McGilligan, L. Zhou, M. Vig, and J. J. Jiang, "Nonlinear dynamic analysis of voices before and after surgical excision of vocal polyps," *J. Acoust. Soc. Am.*, vol. 115, no. 5, pp. 2270–2277, 2004, doi: 10.1121/1.1699392.

[77] M. J. Gangeh, P. Fewzee, A. Ghodsi, M. S. Kamel, and F. Karray, "Multiview supervised dictionary learning in speech emotion recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 22, no. 6, pp. 1056–1068, 2014, doi: 10.1109/TASLP.2014.2319157.

[78] R. Aggarwal, J. Karan Singh, V. Kumar Gupta, S. Rathore, M. Tiwari, and A. Khare, "Noise Reduction of Speech Signal using Wavelet

Transform with Modified Universal Threshold," *Int. J. Comput. Appl.*, vol. 20, no. 5, pp. 14–19, 2011, doi: 10.5120/2431-3269.

[79] G. Yu, E. Bacry, and S. Mallat, "Audio signal denoising with complex wavelets and adaptive block attenuation," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 3, no. 1, pp. 869–872, 2007, doi: 10.1109/ICASSP.2007.366818.

[80] G. Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1830–1839, 2008, doi: 10.1109/TSP.2007.912893.

[81] K. Siedenburg and M. Dörfler, "Audio denoising by generalized time-frequency thresholding," *Proc. AES Int. Conf.*, pp. 174–183, 2012.

[82] I. I. Conference and S. Processing, "Bayram Istanbul Technical University , Istanbul , Turkey," pp. 2917–2921, 2014.

[83] G. Bhattacharya and P. Depalle, "Sparse denoising of audio by greedy time-frequency shrinkage," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 2898–2902, 2014, doi: 10.1109/ICASSP.2014.6854130.

[84] B. M. Ismail, B. Eswara Reddy, and T. Bhaskara Reddy, "Cuckoo inspired fast search algorithm for fractal image encoding," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 30, no. 4, pp. 462–469, 2018, doi: 10.1016/j.jksuci.2016.11.003.

[85] B. Mohammed Ismail, T. B. Reddy, and B. E. Reddy, "Spiral architecture based hybrid fractal image compression," *2016 Int. Conf. Electr. Electron. Commun. Comput. Optim. Tech. ICEECCOT 2016*, pp. 21–26, 2017, doi: 10.1109/ICEECCOT.2016.7955179.

[86] A. Qaroush, I. Abu Farha, W. Ghanem, M. Washaha, and E. Maali, "An efficient single document Arabic text summarization using a combination of statistical and semantic features," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 33, no. 6, pp. 677–692, 2019, doi: 10.1016/j.jksuci.2019.03.010.

[87] Y. Zhou, J. Wang, and X. Zhang, "Research on Speaker Recognition Based on Multifractal Spectrum Feature," vol. 1, no. c, pp. 4–7, 2010, doi: 10.1109/ICCMS.2010.66.

[88] X. Zhao and D. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2013, pp. 7204–7208, doi: 10.1109/ICASSP. 2013.6639061.

[89] M. Ismail, V. Harsha Vardhan, V. Aditya Mounika, and K. Surya Padmini, "An effective heart disease prediction method using artificial neural network," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 8, pp. 1529–1532, 2019.

[90] B. Mohammed Ismail, M. Alam, M. Tahernezhadi, H. K. Vege, and P. Rajesh, "A Machine Learning Classification Technique for Predicting Prostate Cancer," *IEEE Int. Conf. Electro Inf. Technol.*, vol. 2020-July, no. July 2020, pp. 228–232, 2020, doi: 10.1109/EIT48999. 2020.9208240.

[91] J. Hsieh and C. Hsieh, "Cepstral Coefficients (MFCC) and Vector Quantisation (VQ) for Piglets Sound Detection Written for presentation at the 2018 ASABE Annual International Meeting Sponsored by ASABE," in *ASABE*, 2018, pp. 1–9, doi: 10.13031/aim.201800780.

[92] U. Sarkar *et al.*, “Speaker recognition in bengali language from nonlinear features,” pp. 1–6, 2018.

[93] J. I. N. Li, X. Zhang, J. Tang, J. I. N. Cai, and X. Liu, “SIGNAL-NOISE IDENTIFICATION AND SEPARATION BASED ON MULTIFRACTAL SPECTRUM AND MATCHING PURSUIT,” vol. 27, no. 1, pp. 1–13, 2019, doi: 10.1142/S0218348X19400073.

[94] J. Lin, Y. Yumei, Z. Maosheng, C. Defeng, W. Chao, and W. Tonghan, “A multiscale chaotic feature extraction method for speaker recognition,” *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/8810901.

[95] C. M. Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, 2005, doi: 10.1109/TSA.2004.838534.

[96] T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” *Speech Commun.*, vol. 41, no. 4, pp. 603–623, 2003, doi: 10.1016/S0167-6393(03)00099-2.

[97] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011, doi: 10.1016/j.patcog.2010.09.020.

[98] J. Rong, G. Li, and Y. P. P. Chen, “Acoustic feature selection for automatic emotion recognition from speech,” *Inf. Process. Manag.*, vol. 45, no. 3, pp. 315–328, 2009, doi: 10.1016/j.ipm.2008.09.003.

[99] C. N. Anagnostopoulos, T. Iliou, and I. Giannoukos, “Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011,” *Artif. Intell. Rev.*, vol. 43, no. 2, pp. 155–177, 2015, doi: 10.1007/s10462-012-9368-5.

[100] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 312–315, 2009, doi: 10.21437/interspeech.2009-103.

[101] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," *Signal Processing*, vol. 90, no. 5, pp. 1415–1423, 2010, doi: 10.1016/j.sigpro.2009.09.009.

[102] S. Wu, T. H. Falk, and W. Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Commun.*, vol. 53, no. 5, pp. 768–785, 2011, doi: 10.1016/j.specom.2010.08.013.

[103] C. C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Commun.*, vol. 53, no. 9–10, pp. 1162–1171, 2011, doi: 10.1016/j.specom.2011.06.004.

[104] B. Basharirad and M. Moradhaseli, "Speech emotion recognition methods: A literature review," *AIP Conf. Proc.*, vol. 1891, 2017, doi: 10.1063/1.5005438.

[105] C. H. Wu and W. Bin Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels (Extended abstract)," *2015 Int. Conf. Affect. Comput. Intell. Interact. ACII 2015*, no. 1, pp. 477–483, 2015, doi: 10.1109/ACII.2015.7344613.

[106] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digit. Signal Process. A Rev. J.*, vol. 22, no. 6, pp. 1154–1160, 2012, doi: 10.1016/j.dsp.2012.05.007.

[107] W. Dai, D. Han, Y. Dai, and D. Xu, "Emotion recognition and affective computing on vocal social media," *Inf. Manag.*, vol. 52, no. 7, pp. 777–788, 2015, doi: 10.1016/j.im.2015.02.003.

[108] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using MFCC," *Proc. 2017 Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2017*, vol. 2018-Janua, pp. 2257–2260, 2018, doi: 10.1109/WiSPNET.2017.8300161.

[109] S. Lalitha, D. Geyasruti, R. Narayanan, and M. Shravani, "Emotion Detection Using MFCC and Cepstrum Features," *Procedia Comput. Sci.*, vol. 70, pp. 29–35, 2015, doi: 10.1016/j.procs.2015.10.020.

[110] H. Aouani and Y. Ben Ayed, "Emotion recognition in speech using MFCC with SVM, DSVM and auto-encoder," *2018 4th Int. Conf. Adv. Technol. Signal Image Process. ATSIP 2018*, pp. 1–5, 2018, doi: 10.1109/ATSIP.2018.8364518.

[111] T. M. Rajisha, A. P. Sunija, and K. S. Riyas, "Performance Analysis of Malayalam Language Speech Emotion Recognition System Using ANN/SVM," *Procedia Technol.*, vol. 24, pp. 1097–1104, 2016, doi: 10.1016/j.protcy.2016.05.242.

[112] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, "Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions," pp. 1–12, 2019, [Online]. Available: http://arxiv.org/abs/1906.05681.

[113] S. Ramamohan and S. Dandapat, "Sinusoidal model-based analysis and classification of stressed speech," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 3, pp. 737–746, 2006, doi:

10.1109/TSA.2005.858071.

[114] A. Guran and F. Pfeiffer, “Dynamics with friction,” vol. 7, p. 329, 1996, [Online]. Available: http://saba.kntu.ac.ir/eecd/taghirad/Ebooks/New TOC/Mechanical Engineering/Dynamics with friction.pdf.

[115] B. G. Perumana and T. K. Basu, “Emotion Recognition in Malayalam Speech Using Mel Filter Bank Based Feature Set,” pp. 2–6.

[116] M. Syamala and N. J. Nalini, “A speech-based sentiment analysis using combined deep learning and language model on real-time product review,” *Int. J. Eng. Trends Technol.*, vol. 69, no. 1, pp. 172–178, 2021, doi: 10.14445/22315381/IJETT-V69I1P226.

[117] H. L. Bear, R. W. Harvey, B. J. Theobald, and Y. Lan, “Which phoneme-to-viseme maps best improve visual-only computer lip-reading?,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8888, pp. 230–239, 2014.

[118] H. L. Bear and R. Harvey, “Phoneme-to-viseme mappings: the good, the bad, and the ugly,” *Speech Commun.*, vol. 95, pp. 40–67, 2017, doi: 10.1016/j.specom.2017.07.001.

[119] K. T. Bibish Kumar, R. K. Sunil Kumar, E. P. A. Sandesh, S. Sourabh, and V. L. Lajish, “Viseme set identification from Malayalam phonemes and allophones,” *Int. J. Speech Technol.*, vol. 22, no. 4, pp. 1149–1166, 2019, doi: 10.1007/s10772-019-09655-0.

[120] U. Bhattacharjee, S. Gogoi, and R. Sharma, “A statistical analysis on the impact of noise on MFCC features for speech recognition,” *2016 Int. Conf. Recent Adv. Innov. Eng. ICRAIE 2016*, 2017, doi: 10.1109/ICRAIE.2016.7939548.

[121] A. M. Ugena, “Phase space reconstruction from a biological time series .,” *J. Appl. Sci.*, vol. 10, pp. 1–12, 2019, doi: 10.3390/app10041430.

[122] J. Augusto *et al.*, “Quantifying Features Using False Nearest Neighbors : An Unsupervised Approach,” *IEEE 23rd Int. Conf. Tools with Artif. Intell. ICTAI 2011*, vol. 2011, doi: 10.1109/ICTAI.2011.170.

[123] L. Cao, “Practical method for determining the minimum embedding dimension of a scalar time series,” *Phys. D Nonlinear Phenom.*, vol. 110, no. 1–2, pp. 43–50, 1997, doi: 10.1016/S0167-2789(97)00118-8.

[124] E. Bradley and H. Kantz, “Nonlinear time-series analysis revisited,” *Chaos*, vol. 25, no. 9, 2015, doi: 10.1063/1.4917289.

[125] M. A. Hong-guang and H. A. N. Chong-zhao, “Selection of Embedding Dimension and Delay Time in Phase Space Reconstruction,” no. October 2000, pp. 111–114, 2006, doi: 10.1007/s11460-005-0023-7.

[126] B. L. Khubchandani, “Accurate determination of time delay and embedding dimension for state space reconstruction from a scalar time series.”

[127] N. Davey *et al.*, “ScienceDirect Time Time Series Series Analysis Analysis using using Embedding Embedding Dimension Dimension on on Heart Heart Rate Rate Variability Variability,” *Procedia Comput. Sci.*, vol. 145, pp. 89–96, 2019, doi: 10.1016/j.procs.2018.11.015.

[128] F. K. Mohamed and V. L. Lajish, “Nonlinear Speech Analysis and Modeling for Malayalam Vowel Recognition,” *Procedia - Procedia Comput. Sci.*, vol. 93, no. September, pp. 676–682, 2016, doi: 10.1016/j.procs.2016.07.261.

[129] P. Grassberger and I. Procaccia, “Characterization of strange

attractors," *Phys. Rev. Lett.*, vol. 50, no. 5, pp. 346–349, 1983, doi: 10.1103/PhysRevLett.50.346.

[130] R. Hegger, H. Kantz, and T. Schreiber, "Practical implementation of nonlinear time series methods: The TISEAN package," *Chaos*, vol. 9, no. 2, pp. 413–435, 1999, doi: 10.1063/1.166424.

[131] J. A. A. Filho, A. C. P. L. F. Carvalho, R. F. Mello, S. Alelyani, and H. Liu, "Quantifying features using false nearest neighbors: An unsupervised approach," *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, pp. 994–997, 2011, doi: 10.1109/ICTAI.2011.170.

[132] V. Pitsikalis and P. Maragos, "Analysis and classification of speech signals by generalized fractal dimension features," *Speech Commun.*, vol. 51, no. 12, pp. 1206–1223, 2009, doi: 10.1016/j.specom.2009.06.005.

[133] I. Kokkinos, S. Member, and P. Maragos, "Nonlinear Speech Analysis Using Models for Chaotic Systems," *IEEE Trans. speech audio Process.*, vol. 13, no. 6, pp. 1098–1109, 2005.

[134] J. J. Jiang, Y. Zhang, and C. McGilligan, "Chaos in voice, from modeling to measurement," *J. Voice*, vol. 20, no. 1, pp. 2–17, 2006, doi: 10.1016/j.jvoice.2005.01.001.

[135] A. Stefanovska, "Surrogate data for hypothesis testing of physical systems," *Phys. Rep.*, 2018, doi: 10.1016/j.physrep.2018.06.001.

[136] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. Doyne Farmer, "Testing for nonlinearity in time series: the method of surrogate data," *Phys. D Nonlinear Phenom.*, vol. 58, no. 1–4, pp. 77–94, 1992, doi: 10.1016/0167-2789(92)90102-S.

[137] A. S. Thomas Schreiber, "Improved Surrogate Data for Nonlinearity Tests," *Phys. Rev. Lett.*, vol. 77, no. 4, pp. 635–638, 1996, doi: 10.1152/ajpheart.1988.255.6.h1535.

[138] K. P. Harikrishnan, G. Ambika, and R. Misra, "An algorithmic computation of correlation dimension from time series," *Mod. Phys. Lett. B*, vol. 21, pp. 129–138, 2007.

[139] K. P. Harikrishnan, R. Misra, and G. Ambika, "Commun Nonlinear Sci Numer Simulat Combined use of correlation dimension and entropy as discriminating measures for time series analysis," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 14, no. 9–10, pp. 3608–3614, 2009, doi: 10.1016/j.cnsns.2009.01.021.

[140] R. Kunhimangalam, P. K. Joseph, and O. K. Sujith, "Nonlinear analysis of EEG signals: Surrogate data analysis," *Irbm*, vol. 29, no. 4, pp. 239–244, 2008, doi: 10.1016/j.rbmret.2007.09.006.

[141] I. Tokuda, T. Miyano, K. Aihara, and I. Introduction, "Surrogate analysis for detecting nonlinear dynamics," vol. 110, no. December, 2001, doi: 10.1121/1.1413749.

[142] K. P. Harikrishnan, R. Misra, and G. Ambika, "Efficient use of correlation entropy for analysing time series data," *Pramana-journel Phys.*, vol. 72, no. 2, pp. 325–326, 2009.

[143] P. Henríquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, J. I. Godino-Llorente, and F. Díaz-de-María, "Characterization of Healthy and Pathological Voice Through Measures Based on Nonlinear Dynamics," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 17, no. 6, pp. 1186–1195, 2009, doi: 10.1109/TASL.2009.2016734.

[144] B. Boyanov, S. Hadjitodorov, B. Teston, and D. Doskov, "Robust

hybrid pitch detector for pathological voice analysis," *Larynx'97,france,International speech Commun. Assoc.*, pp. 55–58, 1997, [Online]. Available: http://www.isca-speech.org/archive_open/larynx_97/lar7_055.html.

[145] R. J. T. De Sousa, "a New Accurate Method of Harmonic-To-Noise Ratio Extraction," pp. 351–356, 2011, doi: 10.5220/0001552903510356.

[146] Y. Qi and R. E. Hillman, "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," *J. Acoust. Soc. Am.*, vol. 102, no. 1, pp. 537–543, 1997, doi: 10.1121/1.419726.

[147] L. Gu and J. G. Harris, "Disordered Speech Assessment Using Automatic Methods Based on Quantitative Measures," *EURASIP J. Appl. Signal Process.*, vol. 9, no. 1, pp. 1400–1409, 2005, doi: 10.1155/ASP.2005.1400.

[148] B. Boyanov and S. Hadjitodorov, "Acoustic analysis of pathological voices: A voice analysis system for the screening and laryngeal diseases," *IEEE Eng. Med. Biol. Mag.*, vol. 16, no. 4, pp. 74–82, 1997, doi: 10.1109/51.603651.

[149] J. H. . H. E.J Wallen, "A screening test for speech pathology assessment using objective quality measures," doi: 10.1109/CSLP.96.607478.

[150] S. R. Kadiri and P. Alku, "Analysis and Detection of Pathological Voice Using Glottal Source Features," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 2, pp. 367–379, 2020, doi: 10.1109/JSTSP.2019.2957988.

[151] C. Fang, H. Li, L. Ma, and X. Zhang, "Nonlinear dynamic analysis of pathological voices," *Lect. Notes Comput. Sci. (including Subser. Lect.*

Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7996 LNAI, pp. 401–409, 2013, doi: 10.1007/978-3-642-39482-9_46.

[152] R. Palaniappan, "Reliable System for Respiratory Pathology Classification from Breath Sound Signals," pp. 152–156, 2016.

[153] D. Panek, A. Skalski, and J. Gajda, "Quantification of Linear and Non-linear Acoustic Analysis Applied to Voice Pathology Detection," vol. 4, pp. 355–364, 2014, doi: 10.1007/978-3-319-06596-0_33.

[154] U. Cesari, G. De Pietro, E. Marciano, C. Niri, G. Sannino, and L. Verde, "A new database of healthy and pathological voices," *Comput. Electr. Eng.*, vol. 68, no. December 2017, pp. 310–321, 2018, doi: 10.1016/j.compeleceng.2018.04.008.

[155] K. P. Harikrishnan, R. Misra, G. Ambika, and R. E. Amritkar, "Computing the multifractal spectrum from time series : An algorithmic approach," pp. 1–9, 2009, doi: 10.1063/1.3273187.

[156] K. Aida-Zade, A. Xocayev, and S. Rustamov, "Speech recognition using Support Vector Machines," *Appl. Inf. Commun. Technol. AICT 2016 - Conf. Proc.*, vol. 1, 2017, doi: 10.1109/ICAICT.2016.7991664.

[157] A. C. G. Thome, "SVM Classifiers – Concepts and Applications to Character Recognition," *Adv. Character Recognit.*, pp. 25–49, 2012, [Online]. Available: http://www.intechopen.com/books/advances-in-character-recognition.

[158] B. A. Sonkamble and D. D. Doye, "An overview of speech recognition system based on the support vector machines," *Proc. Int. Conf. Comput. Commun. Eng. 2008, ICCCE08 Glob. Links Hum. Dev.*, pp. 768–771, 2008, doi: 10.1109/ICCCE.2008.4580709.

[159] T. Söderström, "Some Methods for Identification of Linear Systems with Noisy Input Output Data," *IFAC Proc. Vol.*, vol. 12, no. 8, pp. 357–363, 1979, doi: 10.1016/s1474-6670(17)65439-9.

[160] K. V. Fernando and H. Nicholson, "Identification of Linear Systems With Input and Output Noise; the Koopmans-Levin Method.," *IEE Proc. D Control Theory Appl.*, vol. 132, no. 1 pt D, pp. 30–36, 1985, doi: 10.1049/ip-d.1985.0007.

[161] R. Singh, M. L. Seltzer, B. Raj, and R. M. Stern, "Speech in noisy environments: Robust automatic segmentation, feature extraction, and hypothesis combination," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 1, pp. 273–276, 2001, doi: 10.1109/icassp.2001.940820.

[162] U. Shrawankar and V. Thakare, "Feature extraction for a speech recognition system in noisy environment: A study," *2010 2nd Int. Conf. Comput. Eng. Appl. ICCEA 2010*, vol. 1, pp. 358–361, 2010, doi: 10.1109/ICCEA.2010.76.

[163] K. Mustin, C. Dytham, T. G. Benton, and J. M. J. Travis, "Red noise increases extinction risk during rapid climate change," *Divers. Distrib.*, vol. 19, no. 7, pp. 815–824, 2013, doi: 10.1111/ddi.12038.

[164] Y. Zhang, "Support vector machine classification algorithm and its application," *Commun. Comput. Inf. Sci.*, vol. 308 CCIS, no. PART 2, pp. 179–186, 2012, doi: 10.1007/978-3-642-34041-3_27.

[165] J. W. Kantelhardt, S. A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H. E. Stanley, "Multifractal detrended fluctuation analysis of nonstationary time series," *Phys. A Stat. Mech. its Appl.*, vol. 316, no. 1–4, pp. 87–114, 2002, doi: 10.1016/S0378-4371(02)01383-3.

[166] E. A. F. Ihlen, "Introduction to multifractal detrended fluctuation analysis in Matlab," *Front. Physiol.*, vol. 3 JUN, no. June, pp. 1–18, 2012, doi: 10.3389/fphys.2012.00141.

[167] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. September, pp. 223–227, 2014, doi: 10.21437/interspeech.2014-57.

[168] S. Karimi and M. H. Sedaaghi, "How to categorize emotional speech signals with respect to the speaker's degree of emotional intensity," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 24, no. 3, pp. 1306–1324, 2016, doi: 10.3906/elk-1312-196.

[169] A. Shahzadi, A. Ahmadyfard, A. Harimi, and K. Yaghmaie, "Speech emotion recognition using nonlinear dynamics features," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 23, no. June 2019, pp. 2056–2073, 2015, doi: 10.3906/elk-1302-90.

[170] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1145–1153, 2006, doi: 10.1109/TASL.2006.876113.

[171] A. Haque and K. S. Rao, "Modification of energy spectra, epoch parameters and prosody for emotion conversion in speech," *Int. J. Speech Technol.*, vol. 20, no. 1, pp. 15–25, 2017, doi: 10.1007/s10772-016-9386-9.

[172] G. Sugihara *et al.*, "Detecting strange attractors in turbulence," *J. Anim. Ecol.*, vol. 84, no. 6, pp. 388–400, 2012.

[173] A. Shahzadi, A. Ahmadyfard, A. Harimi, and K. Yaghmaie, "Speech emotion recognition using nonlinear dynamics features," *Turkish J.*

Electr. Eng. Comput. Sci., vol. 23, pp. 2056–2073, 2015, doi: 10.3906/elk-1302-90.

[174] Y. Sun, X. Y. Zhang, J. H. Ma, C. X. Song, and H. F. Lv, "Nonlinear Dynamic Feature Extraction Based on Phase Space Reconstruction for the Classification of Speech and Emotion," *Math. Probl. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/9452976.

[175] K. M. Indrebo, R. J. Povinelli, and M. T. Johnson, "Sub-banded reconstructed phase spaces for speech recognition," *Speech Commun.*, vol. 48, no. 7, pp. 760–774, 2006, doi: 10.1016/j.specom.2004.12.002.

[176] J. Sun, N. Zheng, and X. Wang, "Enhancement of Chinese speech based on nonlinear dynamics," *Signal Processing*, vol. 87, no. 10, pp. 2431–2445, 2007, doi: 10.1016/j.sigpro.2007.03.020.

[177] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, 2015, doi: 10.1016/j.specom.2015.03.004.

[178] S. Livingstone and F. Russo, *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*, vol. 13. 2018.

[179] T. Ikuma, A. J. McWhorter, L. Adkins, and M. Kunduk, "Development of parameters towards voice bifurcations," *Appl. Sci.*, vol. 11, no. 12, pp. 1–14, 2021, doi: 10.3390/app11125469.

[180] M. Chougala and K. Shridhar, "Novel Formant Estimation Techniques for Speech Processing," vol. 3, no. 6, pp. 229–234, 2015.

[181] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3,

pp. 1638–1652, 2008, doi: 10.1121/1.2951592.

[182] H. Huang and J. Pan, "Speech pitch determination based on Hilbert-Huang transform," *Signal Processing*, vol. 86, no. 4, pp. 792–803, 2006, doi: 10.1016/j.sigpro.2005.06.011.

[183] S. Gonzalez, S. Member, and M. Brookes, "PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise," vol. 22, no. 2, pp. 518–530, 2014.

[184] A. De Cheveigne, "YIN, a fundamental frequency estimator for speech and music," vol. 111, no. April, 2002, doi: 10.1121/1.1458024.

[185] L. N. Tan and A. Alwan, "Multi-band summary correlogram-based pitch detection for noisy speech," *Speech Commun.*, vol. 55, no. 7–8, pp. 841–856, 2013, doi: 10.1016/j.specom.2013.03.001.

[186] W. Chu and A. Alwan, "SAFE: A statistical approach to F0 estimation under clean and noisy conditions," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 3, pp. 933–944, 2012, doi: 10.1109/TASL.2011.2168518.

[187] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirovic, "SPICE: Self-Supervised Pitch Estimation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1118–1128, 2020, doi: 10.1109/TASLP.2020.2982285.

[188] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum A-Posteriori Probability Pitch Tracking in Noisy Environments Using Harmonic Model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76–87, 2004, doi: 10.1109/TSA.2003.819950.

[189] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A maximum

likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," *Eurasip J. Audio, Speech, Music Process.*, vol. 2007, 2007, doi: 10.1155/2007/84186.

[190] K. Han and D. L. Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 2158–2168, 2014, doi: 10.1109/TASLP.2014.2363410.

[191] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A Convolutional Representation for Pitch Estimation," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 161–165, 2018, doi: 10.1109/ICASSP.2018.8461329.

[192] S. A. Majeed, H. Husain, S. A. Samad, and T. F. Idbeaa, "Mel frequency cepstral coefficients (Mfcc) feature extraction enhancement in the application of speech recognition: A comparison study," *J. Theor. Appl. Inf. Technol.*, vol. 79, no. 1, pp. 38–56, 2015.

LIST OF PUBLICATIONS

1. **Muraleedharan, K. M**., Bibish Kumar, K. T., Sunil John, Sunil Kumar, R. K. (2021). Reconstruction of Phase Space and Eigenvalue Decomposition from a Biological Time Series: A Malayalam Speech Signal Case Study. *Journal of Interconnection networks, world scientific 21(3):* ISSN: 1793-6173. DOI:10.1142/S02192659214300 39

2. **Muraleedharan, K. M**., Bibish Kumar, K. T., Sunil John, Sunil Kumar, R. K. (2021). Noise Identification in Speech Data by Multi fractal De-trended Fluctuation Analysis. *Special issue on Emerging Techniques, vidyabharati International Interdisciplinary Research journal.* ISSN: 2319-4797

3. **Muraleedharan, K. M**., Bibish Kumar, K. T., Sunil John, Sunil Kumar, R. K. (2019). Analysis of Time Delay and Embedding Dimension of Reconstructed Phase Space of Human Vocal Tract Using Malayalam Vowels. *International Journal of Emerging Technologies and Innovative Research,* 6(3), 549-556. ISSN:2349-5162

4. Bibish Kumar, K. T., Sunil John, **Muraleedharan, K. M.,** & Sunil Kumar, R. K. (2019). Hierarchical Picture of existing Audio-Visual Speech Database. *International Journal of Recent Technology and Engineering, 8*(3), 8372-8379. DOI:10.35940/ijrte.C6483.098319

5. Sunil Kumar, R. K., **Muraleedharan, K. M**., Vivek, P., & Lajish, V. L. (2016). Study of nonlinear properties of vocal tract and its effectiveness in speaker modelling. *Journal of Acoustical Society of India*, *43*(2), 116-124.

6. Bibish Kumar, K. T., Sunil John, **Muraleedharan, K. M.**, & Sunil Kumar, R. K. (2021). Linguistically involved Data-driven Approach for Malayalam Phoneme-to-Viseme Mapping. *In Applied Speech Processing: Algorithms and Case Studies.* Elsevier Press *(Book chapter).*

Conference papers

1 . **Muraleedharan, K. M.,** and Sunil Kumar, R. K.(2016). Characterisation of Vocal tract by Box counting dimension analysis. *International Conference on Nonlinear Physics: Theory and Experiment, organised by Department of Physics and Research Centre, Farook College, Kozhikode, Kerala, during* 13-14 December 2016

2. **Muraleedharan, K. M.,** and Sunil Kumar, R. K. (2018). High frequency spectral coefficients of speech time series and positive Lyapunov exponents of speech time series. *National Conference on Advances in Statistical Methods, organised by Department of Statistical Sciences, Kannur University, Kerala, during 08-10 November 2018.*

3. **Muraleedharan, K. M**., Sunil Kumar, R. K., Lajish, V. L., and Bibish Kumar, K. T. (2018). Optimisation of Time Delay and Embedding Dimension for Phase Space Reconstruction to Analyse the Dynamical Behaviour of Vocal Tract. *13th Western Pacific Acoustics Conference (WESPAC 2018), organised by CSIR-National Physical Laboratory, New Delhi, during 11-15 November 2018*.

4. Bibish Kumar, K. T., **Muraleedharan, K. M**., Sunil John. et al. (2019). Effectiveness of K-S Test to Classify the Time Domain Distribution Pattern of Speech Signal for Forensic Applications. *National Conference on Cyber Security and Digital*

Forensic (CYFO-2019), organised by Department of School of Information Science and Technology, Kannur University, Kerala, during16 -18 January 2019.

Communicated papers

1. **K.M. Muraleedharan[1]**, K.T. Bibish Kumar[2], R.K Sunil Kumar[3], R.K., B. Ismail [4] ·Detecting Nonlinearity in Malayalam Speech: Surrogate Data Analysis with Correlation Dimension and Correlation Entropy as Nonlinear Discriminating Measures**.** *Malaysian journal of computer science.ISSN:0127-9084(Under review)*

2. **Muraleedharan, K. M**., Sunil Kumar, R. K., Bibish Kumar, K. T., Sunil John. Combined Use of Nonlinear Measures for Analysing Pathological Voices. *Journal of Intelligent systems. ISSN:2191-026X(Under review)*