# INVESTIGATIONS ON THE APPLICATIONS OF DYNAMICAL INSTABILITIES AND DETERMINISTIC CHAOS FOR SPEECH SIGNAL PROCESSING

A THESIS SUBMITTED BY

PRAJITH.P

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

UNDER THE

FACULTY OF SCIENCE

UNIVERSITY OF CALICUT
KERALA – 673 635
INDIA

JANUARY – 2008

# DEDICATION

## To my wife, Deepthi

who has greatly encouraged and supported me during my studies

&

## To my children

who always bring joy to my life

# CERTIFICATE

This is to certify that the thesis entitled **“INVESTIGATIONS ON THE APPLICATIONS OF DYNAMICAL INSTABILITIES AND DETERMINISTIC CHAOS FOR SPEECH SIGNAL PROCESSING”** is a report of the original work carried out by Mr. Prajith.P under my supervision and guidance in the Computer Speech and Intelligence Research Centre, Post Graduate Department of Physics, Government College, Madappally, University of Calicut, Kerala and that no part thereof has been presented for the award of any other degree.

Dr. N.K.Narayanan
Professor & Chairman
School of Information Science & Technology
Kannur University
KERALA

Kannur 670 567
January 2, 2008

# DECLARATION

I hereby declare that the work presented in this thesis is based on the original work done by me under the supervision of Dr. N.K. Narayanan in the Computer Speech and Intelligence Research Centre, Post Graduate Department of Physics, Government College Madappally, University of Calicut, Kerala and that no part thereof has been presented for the award of any other degree.

Calicut 673 635
January 2, 2008

Prajith.P

# ACKNOWLEDGEMENTS

This thesis is the outcome of the education, guidance and inspiration I received from my research guide Dr.N.K..Narayanan, Professor & Chairman, School of Information Science & Technology, Kannur University, Kerala. I wish to express my sincere gratitude to him for his invaluable advice, motivation and patience throughout the research work.

I am extremely grateful to Mr.K.Suresh Babu, Head of the Department, and Dr.P.Ramakrishan, Computer Speech & Intelligence Research Centre, Post Graduate Department of Physics, Government College, Madappally, Kerala for their encouragement, helpful suggestions and constructive criticisms on this work.

I thank Prof. A.K.Dutta, Indian Statistical Institute, Culcutta, Prof. Samudravijaya, TIFR, Mumbai, and Prof. K.K.Paliwal, Griffith University, Australia, for extending their valuable suggestions and support during the tenure of this work.

My special thanks are due to my colleagues Dr.R.K.Sunilkumar, Dr.V.L.Lajish and Mr.V.Kabeer for their wholehearted co-operation throughout this endeavor.

I am obliged to University Grants Commission for providing financial assistance for carrying out this study.

# CONTENTS

**Page No.**

# ABSTRACT

## "INVESTIGATIONS ON THE APPLICATIONS OF DYNAMICAL INSTABILITIES AND DETERMINISTIC CHAOS FOR SPEECH SIGNAL PROCESSING"

**By**

**PRAJITH. P**

This study investigates the potential use of Reconstructed Phase Space (RPS) based parameters for Speech Signal Processing by utilizing nonlinear dynamical systems theory. In this approach features are extracted from the time domain. Study of nonlinear dynamical system shows that the RPS is able to capture the nonlinear information of the underlying system, that cannot be captured by frequency domain analysis.

A multimedia based system is converted into a low cost data acquisition system by adding a presampling antialiasing analog filter prior to A/D converter of the sound card. A speech database of short vowels in Malayalam is created. Nonlinear invariant parameters of vowel sounds are calculated. With these parameters one can quantify the chaotic behaviour of the speech signal.

Reconstructed Phase Space is generated for speech sounds by the method of time delay embedding. From the reconstructed space, a unique parameter called Reconstructed Phase Space Distribution Parameter (RPSDP) is extracted. These parameters are found to be similar for same vowel and differ from vowel to vowel. They are further used in the recognition experiments.

A new method for pitch estimation using Reconstructed Phase Space in two dimensions is presented. The proposed new method does not suffer from the limitations of other short term pitch estimation techniques. The problems in choosing the optimal time delay and the minimum embedding dimension for the reconstruction of phase space using the method of delays are addressed in this thesis. A simple procedure that quantifies expansion from the identity line of embedding space is developed for choosing proper time delay. For determining proper embedding dimension, Cao's algorithm is used. With the optimum embedding parameters, Reconstructed Phase Space Distribution Parameter is modified as Modified Reconstructed Phase Space Distribution Parameter (MRPSDP).

Recognitions experiments of Malayalam vowels based on the above discussed parameters are conducted with k-NN classifier and neural network. The highest recognition accuracy is obtained with MRPSDP feature using neural network classifier. The nonlinear RPS derived features are then combined with the traditional MFCC feature set to achieve an improvement in the accuracy of recognition experiments. When this joint feature vector is used as input parameter, there is a significant boost in the recognition accuracy than that is obtained when RPS derived feature or MFCC feature alone is used. The entire system is developed and implemented using MATLAB 7.

# LIST OF FIGURES

## LIST OF TABLES

# Chapter 1
# Introduction

## 1.1 Background

Speech is one of the oldest and most natural means of information exchange between human beings. We, humans speak and listen to each other in human-human interface. For many years, people have tried to develop machines that can understand and produce speech as humans do so naturally. Speech processing is the mathematical analysis and application of electrical signals for information storage and retrieval that results from a speech. The area of speech processing generally can be subdivided into the broad overlapping categories of speech analysis, coding, enhancement, synthesis and recognition. Speech analysis is the study of the speech production mechanism in order to generate a mathematical model of the physical phenomena. The study of coding endeavors to store speech information for subsequent recovery. Enhancement, in contrast, is the process of improving the intelligibility and quality of noise-corrupted speech signals. The generation of speech from coded instructions is known as speech synthesis. Speech recognition is the process of synthesis in reverse; namely, given a speech signal, produce the code that generated it. It deals with analysis of the linguistic content of a speech signal.

In the past few decades emerging from the theoretical development of nonlinear dynamics, the study of chaotic dynamics in deterministic systems has become very popular. Perhaps it is because of the amazing findings that the study of chaotic systems has delivered. Interest in nonlinear dynamics grew rapidly after 1963, when Lorenz published its implications for weather prediction.

It would be natural to think that if a system is deterministic, its behavior should be easily predicted. But there are systems where their behavior turns out to be non-predictable: not because of lack of determinism, but because of the complexity of the underlying dynamics, that require an infinite precision, which is unable to be computed. This can be seen in systems where very similar initial conditions yield very different behaviors. No matter how much precision we have, the most minimal differences will tend in the long time to very different results, or the system is highly sensitive to initial conditions. The extreme ‘sensitivity to initial conditions’ mathematically present in the systems is called dynamical instability, or simply chaos. It occurs in mechanical oscillators such as pendula or vibrating objects, in rotating or heated fluids, in laser cavities, in electronic circuits and in some chemical reactions. The chaotic behaviour of a system, which obeys deterministic laws, is called deterministic chaos.

These two distinct research fields are addressed in this work: nonlinear dynamics and speech processing. Speech generation has classically been modelled as a linear system, which provides a convenient and simple mathematical formulation. However, a number of nonlinear effects are present in the physical process, which limit the effectiveness of the linear model. An improved approach may be to view speech as an output of a nonlinear dynamical system.

Speech production is an extremely complex process. It can be described in basic terms as follows. The lungs generate air pressure and this pressure wave is modulated as it flows through the larynx. Today a number of nonlinear effects in the speech production process are known. Firstly, it has been accepted for some time that the vocal tract and the vocal folds do not function independently of each other, but that there is in fact some form of coupling between them when the glottis is open [Koizumi. T., Taniguchi. S. *et.al.*, 1985]. This can cause significant changes in formant characteristics between open and closed glottis cycles. [Brookes. D. M. and Naylor. P. A., 1988]. Teager and Teager [Teager. H. M. and Teager. S. M, 1989] have claimed (based on physical measurements) that voiced sounds are characterised by highly complex airflows in the vocal tract, rather than well behaved laminar flow. Turbulent flow of this nature is also accepted to occur during unvoiced speech, where the generation of sound is due to a constriction at some point in the vocal tract. In addition, the vocal folds will themselves be responsible for further nonlinear behaviour, since the muscle and cartilage, which comprise the larynx, have nonlinear stretching qualities. Such nonlinearities are routinely included in attempts to model the physical process of vocal fold vibration, which have focussed on two or more mass models [Ishizaka. K. and Flanagan. J. L, 1972], [Koizumi. T., Taniguchi. S. *et.al.*, 1987], [Steinecke. I. and Herzel. H, 1995], in which the movement of the vocal folds is modelled by masses connected by springs, with nonlinear coupling. Observations of the glottal waveform reinforce this evidence, where it has been shown that this waveform can change shape at different amplitudes [Schoentgen. J, 1990]. Such a change would not be possible in a strictly linear system where the waveform shape is unaffected by amplitude changes.

Taking into account the evidences presented here for the nonlinearities in speech, the source-filter model clearly has drawbacks. The assumptions made imply a loss of information, which although allowing a simple, mathematically tractable linear model, the full speech signal dynamics can never be properly captured. If instead of choosing a linear model, the nonlinear option is pursued, then it is not necessary to make many of the assumptions that are required in the linear case. Hence speech can be modelled as the output of a dynamical system, which has nonlinear behaviour.

Speech recognition has a history of more than 50 years. With the emerging of powerful computers and advanced algorithms, speech recognition has undergone a great amount of progress over the last 25 years. The earliest attempts to build systems for Automatic Speech Recognition (ASR) were made in 1950s based on acoustic phonetics. These systems relied on spectral measurements, using spectrum analysis and pattern matching to make recognition decisions, on tasks such as vowel recognition [Forgie. J. W and Forgie. C. D, 1959]. Filter bank analysis was also utilized in some systems to provide spectral information. In the 1960s, several basic ideas in speech recognition emerged. Zero-crossing analysis and speech segmentation were used, and dynamic time aligning and tracking ideas were proposed [Reddy. D. R, 1966]. In the 1970s, speech recognition research achieved major milestones. Tasks such as isolated word recognition became possible using Dynamic Time Warping (DTW). Linear Predictive Coding (LPC) was extended from speech coding to speech recognition systems based on LPC spectral parameters. IBM initiated the effort of large vocabulary speech recognition in the 70s [Rabiner. L.R and Juang. B. H, 1993], which turned out to be highly successful and had a great impact in speech recognition research. Also, AT&T Bell Labs began making truly speaker-independent speech recognition systems by studying clustering algorithms for creating speaker-independent patterns [Rabiner. L. R, Levinson.S.E, *et. al.*,1979]. In the 1980s, connected word recognition systems were devised based on algorithms that concatenated isolated words for recognition. An important turning point was the transition of approaches from template-based to statistical modeling – especially the Hidden Markov Model (HMM) approach [Rabiner. L. R, 1989]. HMMs were not widely used in speech application until the mid-1980s. From then on, almost all speech research has involved using the HMM technique. In the late 1980s, neural networks were also introduced to problems in speech recognition as a signal classification technique. In spite of these efforts, machines are still far inferior to humans in speech recognition capabilities.

## 1.2 Motivation

The problem addressed in this work concerns the investigation of novel acoustic modeling techniques that exploit the theoretical results of nonlinear dynamics, and applies them to the area of speech processing like pitch detection and speech recognition. In continuous speech recognition system, generally, the word spotting method is adopted. This necessitates large number of isolated word templates, for an efficient recognition system. The main disadvantage of this method is that enormous number of checking is necessary in this system. Many mismatching may occur during checking because of the differences in utterance speed between isolated and continuous speech units. So this method is a failure in many cases. An alternative method is phoneme based word spotting. Phonemes are basic speech units of language. In this method an isolated word to be used as reference pattern is divided into phoneme based segments and a permissible duration range for each segment is set. As a result, the words spotting reference pattern is represented by a sequence of phoneme segments together with minimum and maximum permissible durations and with a typical spectrum for the segment. Phonemes are two types; vowel phoneme and consonant phoneme. Many vowel phoneme recognition systems are reported in literature [Waibel.A. et.al, 1988], [Chandrasekhar. C, 1996], [Chandrasekhar.C and Yegnanarayana B, 1996], [Yegnanarayana B, 1999], and [Sada Siva Sarma. A and Agrawal. S.S, 1996]. However, the average vowel phoneme recognition accuracy reported in literature is very low. So further study in the parameterization of vowels is to be carried out for developing a robust speech recognition system. Many research works are reported on this topic in Indian languages like Hindi, Tamil, Bengali etc. Malayalam speech recognition is still in its infancy. Very few research works are reported so far. So, more basic research works are essential in this area. In the present study Malayalam vowel sounds are modeled using nonlinear speech processing techniques.

As discussed previously, current speech recognition systems typically use frequency domain features, obtained via a frame-based spectral analysis of the speech signal. Such frequency domain approaches are constrained by linearity assumptions incurred by the source-filter model of speech production. The Phase Space of a dynamical system is a mathematical space where the orthogonal co-ordinate directions represent the variables needed to specify the instantaneous state of the system. Takens' theorem states that under certain assumptions, phase space of a dynamical system can be reconstructed through the use of time-delayed versions of the original scalar measurements[Takens.F, 1980]. This new state space is commonly referred to in the literature as a reconstructed phase space (RPS), and has been proven to be topologically equivalent to the original phase space of the dynamical system, as if all the state variables of that system would have been measured simultaneously [Kubin.G, 1995]. A Reconstructed Phase Space can be exploited as a powerful signal processing domain, especially when the dynamical system of interest is nonlinear or even chaotic [Broomhead.D.S. and King.G, 1986], [Kantz.H. and Schreiber.T, 2003]. The RPS representation is capable of preserving the nonlinear dynamics of the signal. The RPS based nonlinear methods concentrate on geometric structure that appear in the RPS. The orbits or trajectories that the row vectors of the trajectory matrix produce these geometric patterns. These orbits can be loosely called attractors. This method addresses the problem in the time domain instead of the frequency domain so that nonlinear information can be captured. The application of RPS for speech recognition is a new path of research and is still in its very early stages. The potential of this method for speech processing motivates the work presented in the thesis.

### 1.3 Outline of the work and Main Results

The intend of the Chapter 2 is to establish a necessary background for the following chapters with a review of previous works. The former part of the chapter contains a brief review of the state of art of nonlinear processing of speech. The later part gives brief survey of research findings reported in the applications of neural network for speech recognition.

In the first session of chapter 3, a brief overview of articulatory and acoustic phonetics has been presented. This provides a framework for later discussions of speech production and the speech signal, especially concentrating on vowel sounds and their characterization. In the second session data acquisition of Malayalam vowel sounds is presented in detail. The personal computers available today possess built in sound card. These cards can be modified for speech processing by incorporating an antialiasing presampling filter appropriately. We have converted a sound card into one suitable for speech processing by incorporating an antialiasing presampling low pass filter at the input for removing the frequency component above 4kHz. A speech database of short vowels in Malayalam is created using the data acquisition system developed, for speech analysis and recognition study described in the following chapters.

In Chapter 4, the field of nonlinear dynamical theory is introduced, including the concepts of chaos and how to measure it. The traditional model of speech production has been shown to have a number of shortcomings and a nonlinear system has been proposed as an alternative. The problem of whether speech (especially vowel sounds) is chaotic has been examined through discussion of previous studies and experiments. Nonlinear invariant parameters for Malayalam vowels are calculated. The major invariant features include attractor dimensions and Kolmogorov entropy. The non-integer attractor dimension and non-zero value of Kolmogorov entropy confirm the contribution of deterministic chaos to the behavior of speech signal.

Though these parameters quantify the chaotic behaviour of the speech signal, as far as recognition application is concerned, we want to get more robust and computationally simpler parameters. In chapter 5, we are focusing to extract a novel parameter from the time domain tool- ‘Phase Space’, in order to capture the nonlinear characteristics of the speech. From the Reconstructed Phase Space of each vowel sound, a promising parameter called Reconstructed Phase Space Distribution Parameter (RPSDP) is extracted. With this parameter we can analyze the geometric structure of the reconstructed attractor. The RPSDPs are found to be similar for same vowel sounds and differs from vowel to vowel. Hence they are further used in the recognition experiments. Here we are presenting an entirely different way of viewing the speech processing problem, and offering an opportunity to capture the nonlinear characteristics of the acoustic structure.

Pitch detection, also referred to as Fundamental frequency ($f_0$) estimation has been a popular research topic for many years, and is still being investigated today. Fundamental frequency, $f_0$ is the lowest frequency component, in the signal, which relates well to most of the other frequency components. In chapter 6, we introduce a general method for pitch estimation using Reconstructed Phase Space in two dimensions. Here methodologies originally developed for analyzing chaotic time series have been successfully applied to pitch determination problem. The proposed new method does not suffer from the limitations of other short-term pitch-estimation techniques. The algorithm is very straightforward and flexible. The results of the simulation experiments show its robust performance on real speech. The experimental results show that the pitch estimated using Reconstructed Phase Space features agrees with that obtained using conventional Pitch Detection Algorithms.

The problem of choosing the optimal time delay and the minimum embedding dimension for the reconstruction of phase space using the method of delays are addressed in chapter 7. From the discussions of the methods of determining proper time delay, it can be concluded that the optimal delay depends upon the details of the time series as well as the dynamics of the underlying system. We developed a simple procedure that quantifies expansion from the identity line of embedding space. Such a procedure may be more useful in the estimation of the proper time delay, since it is the most space filling reconstruction. For determining the minimum embedding dimension we have used Cao's method, which does not contain any subjective parameters except time delay for the embedding and is computationally efficient. This method gives consistent results with Malayalam vowels. With these optimum-embedding parameters, Reconstructed Phase Space Distribution Parameter (RPSDP) is modified as Modified Reconstructed Phase Space Distribution Parameter (MRPSDP).

Chapter 8 deals with the recognition of Malayalam vowels based on the above discussed features, using different classifiers. The credibility of the extracted parameters is tested with the k-NN classifier [Duda.R.O and Hart.P.E, 1973]. A connectionist model based recognition system by means of multi layer feed forward neural network with error back propagation algorithm [Haykin.S, 2004] is then implemented and tested using RPSDP features and MRPSDP features extracted from the vowels. The highest recognition accuracy (92.96%) is obtained with MRPSDP feature using neural network classifier. These results specify the discriminatory strength of the Reconstructed Phase Space derived features for isolated Malayalam vowel classification experiments.

The purpose of the work presented in chapter 9 is to extend the developed nonlinear methods by combing the nonlinear based RPS derived features with the traditional MFCC feature set to achieve a boost in the accuracy of recognition experiments. If the frequency domain features contain different discriminatory information than the RPS derived features, we should get better results. Multi layer feed forward neural network with error back propagation algorithm is implemented and tested with this hybrid feature set extracted from the vowels. An overall recognition accuracy of 96.24% is obtained for the simulation experiments. When the joint feature vector is used as input parameter, there is a significant boost in the recognition accuracy than that obtained when RPS derived feature (92.96%) or MFCC feature (91.68%) alone is used. This result suggests that the frequency domain features and the RPS derived features contain different discriminatory information. The entire system is developed and implemented using MATLAB 7.

Finally chapter 10 concludes this work and suggests a few directions for future research.

# Chapter 2

# Review of Previous Work

## 2.1 Introduction

The last two decades have seen significant advances in human-machine interfaces. Speech and language technology, in particular - speech recognition is one among several areas, which have benefited enormously from these advances (Young, 2001). Initially, the research area of speech recognition was treated as a problem in statistical pattern recognition and classification, using small vocabularies of isolated words or digits recorded in low noise environments. Speech utterances were processed using traditional spectral techniques such as discrete Fourier transforms or filter banks, and direct classification techniques such as template matching were used to make a recognition decision. Some of the earliest systems used statistical pattern matching to do isolated digit recognition [Davis.K.H., Biddulph.R. *et.al*, 1952] or syllable recognition [Forgie.J.W. and Forgie.C.D, 1959],[Olson.H.F. and Belar.H.,1956].

Mathematical models for speech were developed as early as the 1940's [Dudley.H.W, 1940], based on linguistics research that viewed spoken language as the output of the filter system with the impulses from the larynx and vocal folds as the input to the system, the shape of the vocal tract representing the filter parameters, and the speech waveform as the system's output (Source filter Model) [Flanagan.J.L, 1965].

In the late 1960's and early 1970's, speech research became increasingly focused on a few key areas: feature selection and analysis, and template-based classification techniques targeted for speech data. Features such as cepstral coefficients and linear prediction coefficients [Thomas Parsons, 1987], [Rabiner.L and Juang.B.H, 1993] enabled the excitation portion of the speech waveform to be modeled and removed, leaving the vocal tract information relatively intact. Classification techniques such as Dynamic Time Warping (DTW) allowed for a temporally- motivated, non-linear mapping between speech inputs and templates, and resulted in an excellent method for measuring perceptual similarity [Itakura.F, 1975], [Vintsyuk.T.K, 1968]. From mid 1980 onwards, almost all speech research has involved using the Hidden Markov Model (HMM) technique. In the late 1980s, neural networks were also introduced to problems in speech recognition as a signal classification technique.

Source-filter models form the foundation of many speech processing applications such as speech coding, speech synthesis, speech recognition, and speaker recognition technology. Usually, the filter is linear and based on linear predictions. It neglects nonlinear structure known to be present in the speech production mechanism. While this approach has led to great advances in the last 30 years, a fully automatic speech-based interface to products, which would encompass real-time speech processing as well as language understanding, is still considered to be many years away. The replacement of the linear filter with nonlinear operators (models) should enable us to obtain an accurate description of the speech. This in turn may lead to better performance of practical speech processing applications.

This chapter presents a review of previous works in the area of nonlinear speech processing and the applications of neural network for speech recognition and is organized as follows. Section 2.2 provides a summary of research findings in the area of nonlinear speech processing. Section 2.3 gives a review of previous works in the applications of neural network for speech recognition. Finally section 2.4 concludes this review.

## 2.2 Review of Previous works in Nonlinear Speech Processing

Nonlinear methods for speech processing are a rapidly growing area of research. Naturally, it is difficult to define a precise date for the origin of the field, but it is clear that there was a rapid growth in this area, which started in the mid-nineteen eighties. Since that time, numerous techniques were introduced for nonlinear time series analysis, which are ultimately aimed at engineering applications.

Among the nonlinear dynamics community, a budding interest has emerged in the application of theoretical results to experimental time series data analysis in 1980's. One of the profound results established in chaos theory is the celebrated Takens' embedding theorem. Takens' theorem states that under certain assumptions, phase space of a dynamical system can be reconstructed through the use of time-delayed versions of the original scalar measurements. This new state space is commonly referred to in the literature as Reconstructed Phase Space (RPS), and has been proven to be topologically equivalent to the original phase space of the dynamical system.

Packard et al. [Packard.N.H, Crutchfield.J.P, *et.al*, 1980] first proposed the concept of phase space reconstruction in 1980. Soon after, Takens showed that a delay-coordinate mapping from a generic state space to a space of higher dimension preserves topology [Takens.F, 1980]. Sauer and Yorke have modified Taken's theorem to apply for experimental time series data analysis [Sauer T., Yorke.J.A and Casdagli.M, 1991].

Conventional linear digital signal processing techniques often utilize the frequency domain as the primary processing space, which is obtained through the Discrete Fourier Transform (DFT) of a time series. For a linear dynamical system, structure appears in the frequency domain that takes the form of sharp resonant peaks in the spectrum. However for a nonlinear or chaotic system, structure does not appear in the frequency domain, because the spectrum is usually broadband and resembles noise. In the RPS, a structure emerges in the form of complex, dense orbits that form patterns known as attractors. These attractors contain the information about the time evolution of the system, which means that features derived from a RPS can potentially contain more or different information.

The majority of literature that utilizes a RPS for signal processing applications revolves around its use for control, prediction, and noise reduction, reporting both positive and negative results. There is only scattered research using RPS features for classification and /or recognition experiments.

In contrast to the linear source-filter model for speech production process, a large number of research works are reported in the literature to show the nonlinear effects in the physical process. Koizumi.T, Taniguchi.S, *et.al.* in 1985 showed that the vocal tract and the vocal folds do not function independently of each other, but that there is in fact some form of coupling between them when the glottis is open [Koizumi.T, Taniguchi.S, *et.al.*, 1985]. This can cause significant changes in formant characteristics between open and closed glottis cycles. [Brookes.D.M and Naylor.P.A, 1988].

Teager and Teager [Teager.H.M and Teager.S.M, 1989] have claimed that voiced sounds are characterised by highly complex airflows in the vocal tract, rather than well behaved laminar flow. Turbulent flow of this nature is also accepted to occur during unvoiced speech, where the generation of sound is due to a constriction at some point in the vocal tract. In addition, the vocal folds will themselves be responsible for further nonlinear behaviour, since the muscle and cartilage, which comprise the larynx, have nonlinear stretching qualities.

Such nonlinearities are routinely included in attempts to model the physical process of vocal fold vibration, which have focussed on two or more mass models [Ishizaka.K and Flanagan.J.L, 1972], [Koizumi.T, Taniguchi.S, *et.al.*, 1987] [Steinecke.I and Herzel.H, 1995] in which the movement of the vocal folds is modelled by masses connected by springs, with nonlinear coupling. Observations of the glottal waveform reinforce this evidence, where it has been shown that this waveform can change shape at different amplitudes [Shoentgen.J, 1990]. Such a change would not be possible in a strictly linear system where the waveform shape is unaffected by amplitude changes.

Extraction of invariant parameters from speech signal has attracted researchers for designing speech and speaker recognition systems. In 1988, Narayanan.N.K. *et.al.* [Narayanan.N.K. and Sridhar.C.S, 1988] used the dynamical system technique mentioned in the nonlinear dynamics to extract invariant parameters from speech signal. The dynamics of speech signal is experimentally investigated by extracting the second order dimension of the attractor $D_2$ and the second order Kolmogorov entropy $K_2$ of speech signal. The fractal dimension of $D_2$ and non-zero value of $K_2$ confirms the contribution of deterministic chaos to the behavior of speech signal. The attractor dimension $D_2$ and Kolmogorov entropy $K_2$ are then used as a powerful tool for voiced / unvoiced classification of speech signals.

The dimension of the trajectories, or the dimension of the attractor is an important characteristic of the dynamic systems. The estimation of the dimension gives a lower bound of the number of parameters needed in order to model the system. The goal is to find if the system under study occupies all the state space or if it is most of the time in a subset of the space, called attractor. The correlation dimension [Tishby, 1990] is a practical method to estimate the dimension of an empirical temporal series.

There are a large variety of techniques found in the literature of nonlinear methods and it is difficult to predict which techniques ultimately will be more successful in speech processing. However, commonly observed methods in the speech processing literature are various forms of oscillators and nonlinear predictors, the latter being part of the more general class of nonlinear autoregressive methods. The oscillator and autoregressive techniques themselves are also closely related since a nonlinear autoregressive model in its synthesis form forms a nonlinear oscillator if no input is applied. For the practical design of a nonlinear autoregressive model, various approximations have been proposed [Farmer.J.D and Sidorowich.J.D, 1988], [Casdagli.M, Des Jardins. D, *et.al.*, 1992], [Abarbanel.H.D.I, Brown.R, *et.al.*, 1993], [Kubin.G, 1995]. These can be split into two main categories: parametric and nonparametric methods.

Parametric methods are perhaps best exemplified by the polynomial approximation (truncated Volterra series with the special case of quadratic filters [Sicuranza.G.L, 1992], [Mumolo.E and Francescato.D, 1993], [Mumolo.E, Carini.A, *et.al*, 1994], [Thyssen.J, Nielsen.H, *et.al*. 1994], locally linear models [Townshend.B, 1992], [Townshend.B, 1991], [Singer.A.C, Wornell.G.W, *et.al.*1994], [Kumar.A and Gersho.A, 1997], [Ma Y.G.Wei.N, 1998], including threshold autoregressive models [Tong.H, 1990], and state dependent models [Priestley.M.B, 1988]. Another important group of parametric methods is based on neural nets: radial basis functions approximations [Casdagli.M, 1989], [Birgmeier.M, 1995], [Birgmeier.M, 1996], [Diaz de Maria.F, et.al., 1995], [Yee Y.S.P.Haykin, 1995] [Mann.I and McLaughlin, 1998], multi-layer perceptrons [Lapedes.A and Farber.R, 1988], [Reininger.H and Wolf.D, 1990] and recurrent neural nets [Wu.L and Niranjan. M, 1994], [Haykin.S and Li.L, 1995], [Hussain.A, 1996].

Nonparametric nonlinear autoregressive methods also play an important role in nonlinear speech processing. Examples are Lorenz's method of analogues [Lorenz.E.N, 1969], [Bogner.R.E, 1988] [Bogner.R.E and Li.T, 1989], [Casdagli.M, 1989] which may be the simplest of various nearest neighbour methods [Yakowitz.S, 1987], [Farmer.J.D and Sidorowich.J.D, 1988], which also includes nonlinear predictive vector quantization [Gersho.A, 1989] [Wang.S, Paksoy.E, *et.al.*, 1990], [Wu.L, Niranjan.M, 1994], [Gersho.A and Gray.R.M, 1992] or codebook prediction [Singer.A.C, Wornell.G.W, *et.al.*1992], [Kumar.A and Gersho.A, 1997].

Phase space reconstruction is usually the first step in the analysis of dynamical systems. An experimenter obtains a scalar time series from one observable of a multidimensional system. State-space reconstruction is then needed for the indirect measurement of the system's invariant parameters like, dimension, Lyapunov exponent etc. Takens' theorem gives little guidance, about practical considerations for reconstructing a good state space. It is silent on the choice of time delay ($\tau$) to use in constructing m-dimensional data vectors. Indeed, it allows any time delay as long as one has an infinite amount of infinitely accurate data. However, for reconstructing state spaces from real-world, finite, noisy data, it gives no direction [Casdagli. M, Eubank.S, *et. al.*, 1991]. Two heuristics have been developed in the literature for establishing a time lag [Kantz.H and Schreiber.T, 2003]. 1) The first zero of the autocorrelation function and 2) the first minimum of the auto mutual information curve [Fraser.A.M and Swinney.H.L, 1986].

In their work, Andrew M Fraser and Harry L Swinney, the mutual information is examined for a model dynamical system and for chaotic data from an experiment on the Belousov-Zhabotinskii reaction. An N log N algorithm for calculating mutual information (I) is presented. A minimum in ‘I’ is found to be a good criterion for the choice of time delay in Phase Space Reconstruction from time series data. This criterion is shown to be far superior to choosing a zero of the autocorrelation function.

There have been many discussions on how to determine the optimal embedding dimension from a scalar time series based on Taken’s theorem or its extensions [Sauer.T, Yorke.J.A., and Casdagli. M, 1991]. Among different geometrical criteria, the most popular seems to be the method of False Nearest Neighbors [Kennel.M. B, Brown. R, and Abarbanel.H.D.I, 1992]. This criterion concerns the fundamental condition of no self-intersections of the reconstructed attractor.

Work by Banbrook, McLaughlin *et. al.*[Banbrook.M and McLaughlin.S, 1994], Kumar *et. al.* [Kumar.A and Mullick.S.K, 1996], and Narayanan *et. al.* [Narayanan.S.S and Alwan.A.A, 1995] has attempted to use nonlinear dynamical methods to answer the question: “Is speech chaotic?” These papers focused on calculating theoretical quantities such as Lyapunov exponents and Correlation dimension. Their results are largely inconclusive and even contradictory. A synthesis technique for voiced sounds is developed by Banbrook et.al, inspired by the technique for estimating the Lyapunov exponents [Banbrook.M, McLaughlin.S and Mann.I, 1999].

In a work presented by Langi and Kinsner speech consonants are characterised by using a fractal model for speech recognition systems. [Langi.A and Kinsner.W, 1995] Characterization of consonants has been a difficult problem because consonant waveforms may be indistinguishable in time or frequency domain. The approach views consonant waveforms as coming from a turbulent constriction in a human speech production system, and thus exhibiting turbulent and noise like time domain appearance. However, it departs from the usual approach by modeling consonant excitation using chaotic dynamical systems capable of generating turbulent and noise-like excitations. The scheme employs correlation fractal dimension and Takens embedding theorem to measure fractal dimension from time series observation of the dynamical systems. It uses linear predictive coding (LPC) excitation of twenty-two consonant waveforms as the time series. Furthermore, the correlation fractal dimension is calculated using a fast Grassberger algorithm [Grassberger and Procaccia, 1983].

Wei Gang, Lu Yiqing *et. al*. presented a new method for low bit rate speech coding method based on fractal code excited linear prediction [Wei Gang, Lu Yiqing and Quyang Jingzheng, 1996]. Based on the recently developed chaos and fractal theories they introduced new methods for speech signal processing. A novel phase space reconstruction algorithm is proposed for speech signal, the distributions of the maximum Lyapunov exponent and the fractal dimension of speech signal are tested and analyzed statistically. The results of this study indicate that chaos and fractal theories have great potentials in the field of speech signal processing.

The criterion in the False Nearest Neighbor approach for determining optimal embedding dimension is subjective in some sense that, different values of parameters may lead to different results [Cao.L, 1997)]. For realistic time series data, different optimal embedding dimensions are obtained if we use different values of the threshold value. Also with noisy data this method gives spurious results. [Kantz.H and Schreiber.T, 2003].

Lyangyue Cao in 1997 proposed a practical method to determine the minimum embedding dimension from a scalar time series. It does not contain any subjective parameters except for the time delay for the embedding. It does not strongly depend on how many data points are available and it is computationally efficient. Several time series are tested to show the above advantages of the method [Cao.L, 1997)].

Petry *et. al.*[Petry.A, Augusto.D *et.al.*, 2002] and Pitsikalis *et. al.* [Pitsikalis.V and Maragos.P, 2002] have used Lyapunov exponents and Correlation dimension in unison with traditional features (cepstral coefficients) and have shown minor improvements over baseline speech recognition systems. Central to both sets of these papers is the importance of Lyapunov exponents and Correlation dimension, because they are invariant metrics that are the same regardless of initial conditions in both the original and reconstructed phase space. Despite their significance, there are several issues that exist in the measuring of these quantities on real experimental data. The most important issue is that these measurements are very sensitive to noise. Secondarily, the automatic computation of these quantities through a numerical algorithm is not well established and this can lead to drastically differing results. The overall performance of these quantities as salient features remains an open research question.

In addition to these speech analysis and recognition applications, nonlinear methods have also been applied to speech enhancement, speech coding etc. Papers by Hegger *et al.* [Hegger.R, Kantz.H, *et.al.* 2000], [Hegger.R, Kantz.H, *et.al.* 2001] demonstrated the successful application of what is known as local nonlinear noise reduction to sustained vowel utterances.

Michael T. Johnson *et.al.* proposed the implementation of two nonlinear noise reduction methods applied to speech enhancement [Michael T. Johnson, Andrew C. Lindgren, *et.al*, 2003]. The methods are based on embedding the noisy signal in a high-dimensional reconstructed phase space and applying singular value decomposition to project the signal into a lower dimension. The advantages of these nonlinear methods include that they do not require explicit models of noise spectra and do not have the typical ‘musical tone’ side effects associated with traditional linear speech enhancement methods.

In a work presented by Jinjin Ye *et.al*. Principal Component Analysis (PCA) is applied to feature vectors from the reconstructed phase space [Jinjin

Ye, Michael T. Johnson, *et.al.*, 2003]. By using PCA projection, the basis of the feature space is orthogonalized. A Bayes classifier uses the transformed feature vectors to classify phonemes. The results show that the classification accuracy with PCA method surpasses the accuracy using only original features in most cases. PCA projection was implemented in three ways over the reconstructed phase space on both speaker-dependent and speaker-independent data.

Kevin M Lindrebo *et.al.* introduced a method for calculating speech features from third-order statistics of sub band filtered speech signals which are used for robust speech recognition [Kevin M. Indrebo, Richard J. Povinelli, *et.al.*, 2005]. These features have the potential to capture nonlinear information not represented by cepstral coefficients. Also, because the features presented in this method are based on the third-order moments, they may be more immune to Gaussian noise than cepstrals, as Gaussian distributions have zero third-order moments.

Richard J Povinelli *et.al.* introduced a novel approach to the analysis and classification of time series signals using statistical models of reconstructed phase spaces [Povinelli.R.J, Michael T. Johnson, *et.al.*, 2006]. With sufficient dimension, such reconstructed phase spaces are, with probability one, guaranteed to be topologically equivalent to the state dynamics of the generating system, and, therefore, may contain information that is absent in analysis and classification methods rooted in linear assumptions. Parametric and nonparametric distributions are introduced as statistical representations over the multidimensional reconstructed phase space, with classification accomplished through methods such as Bayes maximum likelihood and artificial neural networks (ANNs). The technique is demonstrated on heart arrhythmia classification and speech recognition. This new approach is shown to be a viable and effective alternative to traditional signal classification approaches, particularly for signals with strong nonlinear characteristics.

In a recent study Marcos Faundez-Zanuy compared the identification rates of a speaker recognition system using several parameterizations, with special emphasis on the residual signal obtained from linear and nonlinear predictive analysis [Marcos Faundez-Zanuy, 2007]. It is found that the residual signal is still useful even when using a high dimensional linear predictive analysis. If instead of using the residual signal of a linear analysis a nonlinear analysis is used, both combined signals are more uncorrelated and although the discriminative power of the nonlinear residual signal is lower, the combined scheme outperforms the linear one for several analysis orders.

It is seen that the majority of literature that utilizes the nonlinear techniques for signal processing applications revolves around its use for control, prediction and noise reduction, reporting both positive and negative results. There is only scattered research using these methods for classification or recognition experiments. It is also important to notice that no work has been reported yet in nonlinear speech processing for Malayalam and other

Indian languages. The succeeding session of this chapter is focussed on the review of the applications of neural network for speech recognition.

### 2.3 Review of the applications of neural network for speech recognition

Artificial neural net (ANN) algorithms have been designed and implemented for speech pattern recognition by a number of researchers. ANNs are of interest because algorithms used in many speech recognizers can be implemented using highly parallel neural net architectures and also because new parallel algorithms are being developed making use of the newly acquired knowledge of the working of biological nervous systems. Hutton.L.V compares neural network and statistical pattern comparison method for pattern recognition purpose [Hutton.L.V 1992]. Neural network approaches to pattern classification problems complement and compete with statistical approaches. Each approach has unique strengths that can be exploited in the design and evaluation of classifier systems. Classical (statistical) techniques can be used to evaluate the performance of neural net classifiers, which often outperform them. Neural net classifiers may have advantages even when their ultimate performance on a training set can be shown to be no better than the classical. It is possible to be implemented in real time using special purpose hardware.

Personnaz L.and Dreyfus G presents an elementary introduction to networks of formal neurons [Personnaz L and. Dreyfus.G 1990]. The state of the art regarding basic research and the applications are presented in this work. First, the most usual models of formal neurons are described, together with the most currently used network architectures: static (feedforward) nets and dynamic (feedback) nets. Secondly, the main potential applications of neural networks are reviewed: pattern recognition (vision, speech), signal processing and automatic control. Finally, the main achievements (simulation software, simulation machines, integrated circuits) are presented.

Willian Huang *et.al.* presents some neural net approaches for the problem of static pattern classification and time alignment [Willian Huang *et. al.*, 1988]. For static pattern classification multi layer perceptron classifiers trained with back propagation can form arbitrary decision regions, are robust, and are trained rapidly for convex decision regions. For time alignment, the Viterbi net is a neural net implementation of the Viterbi decoder used very effectively in recognition systems based on Hidden Markov Models (HMMs).

Waibel.A *et. al.* proposed a time delay neural network (TDNN) approach to phoneme recognition, which is characterized by two important properties [Waibel.A *et. al.*, 1988]. Using a three level arrangement of simple computing units, it can represent arbitrary non-linear decision surface. The TDNN learns these decision surfaces automatically using error back propagation. The time delay arrangement enables the network to discover acoustic phonetic features and temporal relationships between them independent of position in time and hence not blurred by temporal shifts in the input. For comparison, several discrete Hidden Markov Models (HMM) were trained to perform the same task, i.e. the speaker dependent recognition of the phonemes “B”, “D” and “G” extracted from varying phonetic contexts.

The TDNN achieved a recognition rate of 98.5% correct compared to 93.7% for the best of HMMs. They show that the TDNN has well known acoustic – phonetic features (e.g., F2-rise, F2-fall, vowel-onset) as useful abstractions. It also developed alternate internal representations to link different acoustic realizations to the same concept**.**

Yoshua Bengio and Renato De Mori used The Boltzmann machine algorithm and the error back propagation algorithm to learn to recognize the place of articulation of vowels(front, center or back), represented by a static description of spectral lines [Yoshua Bengio and Renato De Mori, 1988]. The error rate is shown to depend on the coding. Results are comparable or better than those obtained by them on the same data using hidden Markov Models. They also show a fault tolerant property of the neural nets, i.e. that the error on the test set increases slowly and gradually when an increasing number of nodes fail.

Moore. K.L discussed different types of neural network in his paper entitled “Artificial neural networks” [Moore.K.L, 1992]. Three different tasks for which they are suitable are discussed. They are pattern classification and associative memory, self-organization and feature extraction, and optimization.

Mah. R.S.H and Chakravarthy.V examine the key features of simple networks and their application to pattern recognition. [Mah. R.S.H and Chakravarthy.V, 1992]. Beginning with a three-layer back propagation network, the authors examine the mechanisms of pattern classification. They relate the number of input, output and hidden nodes to the problem features and parameters. In particular, each hidden neuron corresponds to a discriminant in the input space. They point out that the interactions between number of discriminant, the size and distribution of the training set, and numerical magnitudes make it very difficult to provide precise guidelines. They found that the shape of the threshold function plays a major role in both pattern recognition, and quantitative prediction and interpolation. Tuning the sharpness parameter could have a significant effect on neural network performance. This feature is currently under-utilized in many applications. For some applications linear discriminant is a poor choice.

Janssen. R.D.T, Fanty. M and Cole. R.A developed a phonetic front-end for speaker-independent recognition of continuous letter strings [Janssen. R.D.T, Fanty.M and Cole.R.A, 1991]. A feedforward neutral network is trained to classify 3 msec speech frames as one of the 30 phonemes in the English alphabet. Phonetic context is used in two ways: first, by providing spectral and waveform information before and after the frame to be classified, and second, by a second-pass network that uses both acoustic features and the phonetic outputs of the first-pass network. This use of context reduced the error rate by 50%. The effectiveness of the DFT and the more compact PLP (perceptual linear predictive) analysis is compared, and several other features, such as zerocrossing rate, are investigated. A frame-based phonetic classification performance of 75.7% was achieved.

Ki-Seok-Kim and Hee-Yeung-Hwang present the result of the study on the speech recognition of Korean phonemes using recurrent neural network models conducted by them [Ki-Seok-Kim and Hee-Yeung-Hwang, 1991]. The results of applying the recurrent multi layer perceptron model for learning temporal characteristics of speech phoneme recognition, is presented. The test data consist of 144 vowel+consonant+vowel (VCV) speech chains made up of 4 Korean monothongs and 9 Korean plosive consonants. The input parameters of the artificial neural network model used are the FFT coefficients, residual error and zero crossing rates. The baseline model showed a recognition rate of 91% for vowels and 71% for plosive consonants of one male speaker. The authors obtained better recognition rates from various other experiments compared to the existing multilayer perceptron model, thus showing the recurrent model to be better suited to speech recognition. The possibility of using the recurrent models for speech recognition was experimented upon by changing the configuration of this baseline model.

Ahn.R and Holmes.W.H propose a voiced / unvoiced / silence classification algorithm of speech using 2-stage neural networks with delayed decision input [Ahn .R and Holmes.W.H, 1996]. This feed forward neural network classifier is capable of determining voiced, unvoiced and silence in the first stage and refining unvoiced and silence decisions in the second stage. Delayed decision from the previous frame's classification along with preliminary decision by the first stage network, zero crossing ratio and energy ratio enable the second stage to correct the mistakes made by the first stage in classifying unvoiced and silence frames. Comparisons with a single stage classifier demonstrate the necessity of two-stage classification techniques. It also shows that the proposed classifier performs excellently.

Sunilkumar.R.K and Narayanan.N.K investigated the potential use of zerocrossing based information of the signal for Malayalam vowel recognition [Sunilkumar.R.K. 2002]. A vowel recognition system using artificial neural network is developed. The highest recognition accuracy obtained for normal speech is 90.62%.

Dhananjaya.N, Guruprasad.S, *et.al.* proposed a method for detecting speaker changes in a multi speaker speech signal [Dhananjaya.N, Guruprasad.S, *et.al.*, 2004]. The statistical approach to a point phenomenon (speaker change) fails when the given conversation involves short speaker turns (< 5 sec duration). They used auto associative neural network (AANN) models to capture the characteristics of the excitation source that present in the linear prediction (LP) residual of speech signal. The AANN models are then used to detect the speaker changes.

Xavier Domont, Martin Heckmann**,** *et.al.* proposed a feed forward neural network for syllable recognition [Xavier Domont, Martin Heckmann**,** *et.al.,* 2007]. The core of the recognition system is based on a hierarchical architecture initially developed for visual object recognition. In this work, they showed that, given the similarities between the primary auditory and visual cortexes, such a system can successfully be used for speech recognition. Syllables are used as basic units for the recognition. Their spectrograms, computed using a Gammatone filter bank, are interpreted as images and subsequently feed into the neural network after a preprocessing step that enhances the formant frequencies and normalizes the length of the syllables.

In a recent work by Ana I. Garcia Moral, Ruben Solera Urena**,** *et.al.* the solutions provided in the past for Artificial Neural Network are recalled and applied them to Support Vector Machines (SVM), performing a comparison between them [Ana I. Garcia Moral, Ruben Solera Urena, *et.al.*, 2007]. Support Vector Machines are state-of-the-art methods for machine learning but share with more classical ANN the difficulty of their application to temporally variable input patterns. Preliminary results are encouraging.

### 2.4 Conclusion

This chapter provides a summary about the recent advances, new trends and important contributions in the area of nonlinear speech processing and the applications of neural network for speech recognition. Nonlinear signal processing techniques have several potential advantages over traditional linear signal processing methodologies. They are capable of recovering the nonlinear dynamics of the signal of interest possibly preserving natural information. They are not constrained by strong linearity assumptions. Despite these facts, the use of nonlinear signal processing techniques also have disadvantages as well, which is why they have not been widely used in the past. Primarily, they are not as well understood as conventional linear methods. Salient features of Reconstructed Phase Space for classification or recognition have yet to be firmly established, and this work is clearly in the early stages, which is what motivates this research pursuit. Moreover, researchers have just begun to study nonlinear signal processing techniques for a variety of engineering tasks with mixed success. The use of nonlinear methodologies especially as it applies to speech recognition is truly in its infancy with very little work published, most of which has been in the last five years.

## Chapter 3

# Speech Characteristics and Data Acquisition

### 3.1 Introduction

Utterances in any language can be analyzed into a sequence of abstract units called phonemes. Subsets of these phonemes can be grouped into categories according to the type of phonation involved, that is, based on the behaviour of the vocal tract organs to create the particular speech sound. There are many ways in which speech sounds can be grouped. Our primary school categories of vowel and consonant capture the most basic contrast in speech sounds. In phonetics, a vowel is a sound in spoken language that is characterized by an open configuration of the vocal tract so that there is no build-up of air pressure above the glottis and a consonant is a sound, that is characterized by a closure or stricture of the vocal tract sufficient to cause audible turbulence.

The personal computers available today possess built in sound card. The purpose of this card is not the accurate sound feature measurement, but the digitization of speech and music. These cards can be modified for speech processing by incorporating an antialiasing presampling filter appropriately. We have converted a sound card into one suitable for speech processing by incorporating an antialiasing presampling low pass filter at the input.

This chapter is organized as follows. Section 3.2 presents a brief introduction to phonetics, which aims to define the type of phonation involved in the production of different phonemes. Section 3.3 concentrates on an overview of Malayalam vowel sounds. Following this, speech data acquisition and the need for antialiasing filter in data acquisition system is presented in section 3.4. Finally section 3.5 concludes this chapter.

### 3.2 Phonological description of Speech

The field of phonetics includes the study of speech production and the acoustics of the speech signal, and provides a way to effectively describe speech. It can be broadly classified into articulatory phonetics and acoustic phonetics. Articulatory phonetics deals with the articulatory aspects of speech sounds. That is, articulatory phoneticians are interested in how the different structures of the vocal tract, called the articulators (tongue, lips, jaw, palate, teeth etc.), interact to create the specific sounds. Acoustic phonetics is a subfield of phonetics which deals with acoustic aspects of speech sounds. Acoustic phonetics investigates properties like the mean squared amplitude of a waveform, its duration, its fundamental frequency, or other properties of its frequency spectrum, and the relationship of these properties to other branches of phonetics.

#### 3.2.1 Articulatory Phonetics

The process of air being expelled from the lungs and pushing through the vocal tract produces speech signals. The resulting sound pressure wave radiates out from the lips. The various organs involved in speech production process are shown in Figure 3.1. According to their positioning, a large variety of sounds can be produced. In order to discuss these sounds unambiguously, they are categorised into a series of distinct types like, nasals, plosives, fricatives etc. according to how they are produced.

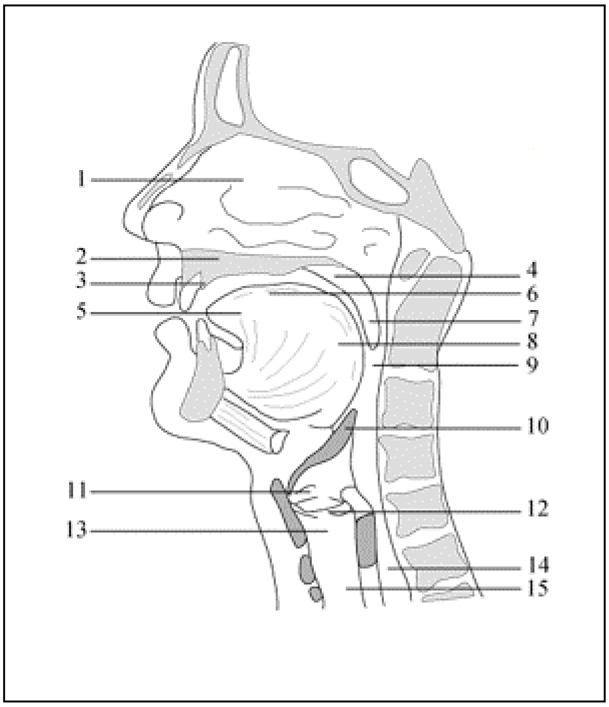


**Fig. 3.1** The human vocal organs. (1) Nasal cavity, (2) Hard palate, (3) Alveoral ridge, (4) Soft palate (Velum), (5) Tip of the tongue (Apex), (6) Dorsum, (7) Uvula, (8) Radix, (9) Pharynx, (10) Epiglottis, (11) False vocal cords, (12) Vocal cords, (13) Larynx, (14) Esophagus, and (15) Trachea.

The larynx is at the base of the vocal tract and mainly comprises two bands of muscle and tissue called the vocal cords or folds. All air from the lungs must pass through the vocal folds, and they can obstruct its passage to a greater or lesser extent. In terms of speech production, the vocal folds can operate in three ways:

- vibrating in a pseudo–periodic manner to create *voiced* sounds. The frequency of this vibration is called the fundamental frequency, and corresponds to the tone heard by a listener which is called pitch;
- not vibrating for *unvoiced* sounds;
- stopped or closed to produce a *glottal stop*, the glottis being the gap between the vocal cords;

The different articulatory organs (*e.g.* tongue, lips, soft–palate) in the vocal tract can be positioned so as to modulate the flow of air through the tract in different ways (close, narrow and open).

- Closure. As well as the glottal stop, the vocal tract may be closed at other places such as at the lips, or between the tongue and hard palate. If the velum is lowered, then air can flow out through the nose creating a *nasal* sound. However, if it is raised, there is no way for the air to escape. Therefore the pressure in the vocal tract increases, and when the closure is removed the air bursts out creating a *plosive* sound.
- Narrowing. If rather than completely closing the vocal tract, two speech organs are instead brought close together, then the air flow through them becomes turbulent and produces *fricative* sounds. The narrowing can occur at any point in the vocal tract.
- Open. With the speech organs sufficiently open so that no turbulence is produced in the airflow, *vowel* sounds are generated. These sounds are always voiced, and it is mainly the position of the highest part of the

tongue that determines the vowel produced. This leads to a widely used description in which vowels are specified according to which part of the tongue is highest (front, central, back) and how high it is (close, mid, open).

### 3.2.2 Acoustic Phonetics

Now we consider how the actual speech waveforms themselves can be linked to the description of sounds explained above, which is based entirely on the mechanics of their production. Figures 3.2(a) and (b) show the time domain and frequency domain representations of the Malayalam vowel അ/Λ/ respectively.

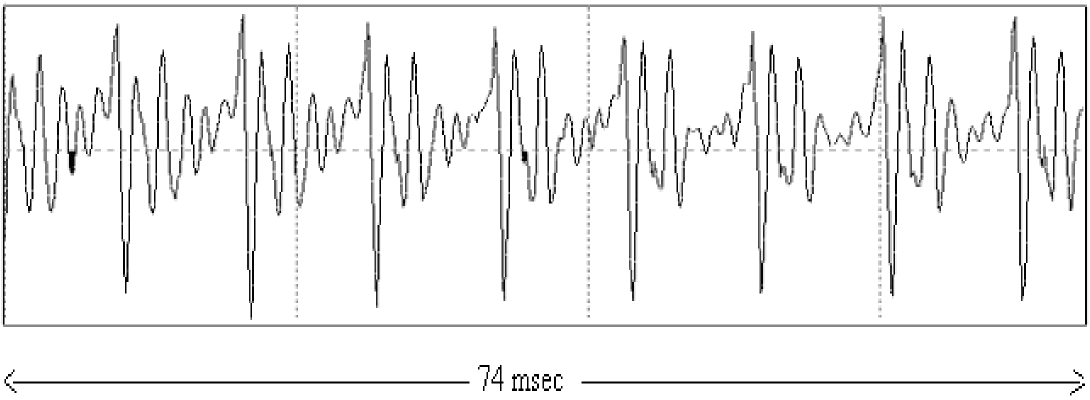


**Fig.3.2(a)** Time domain representation of the Malayalam vowel അ/Λ/.

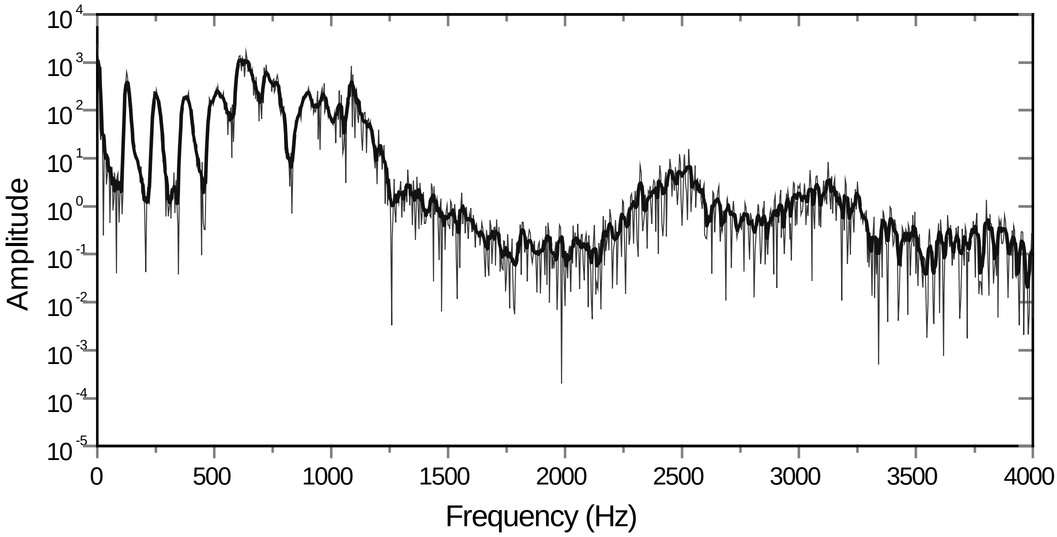


**Fig.3.2(b)** Frequency spectrum of the Malayalam vowel അ/Λ/.

The spectrum, which is limited to 4 kHz, shows a number of peaks at differing levels. These are due to the pseudo–periodic excitation signal and the subsequent modulation by the vocal tract. The first peak, at approximately 125 Hz, is the fundamental frequency and the other peaks are harmonics of this. There are also areas of higher energy within the spectrum, corresponding to the resonant frequencies of the vocal tract. These are called formants, and are labeled incrementally as first formant, second formant, third formant *etc (*F1, F2, F3). In the frequency range shown, three formants are visible at approximately 650 Hz, 1100 Hz and 2500 Hz. Formant frequencies provide a useful method to characterise vowels acoustically, since they depend upon where the vocal tract is constricted and by how much – *i.e.* tongue position [Rabiner.L.R and Schafer.R.W, 1978]. There are at least five formants present in a typical vowel, but in terms of using them for specification, only the first two or three are required. Figure 3.3 shows the listing of Formant frequencies in the Malayalam vowel sound അ/Λ/ over a time period of 183.5 msec.

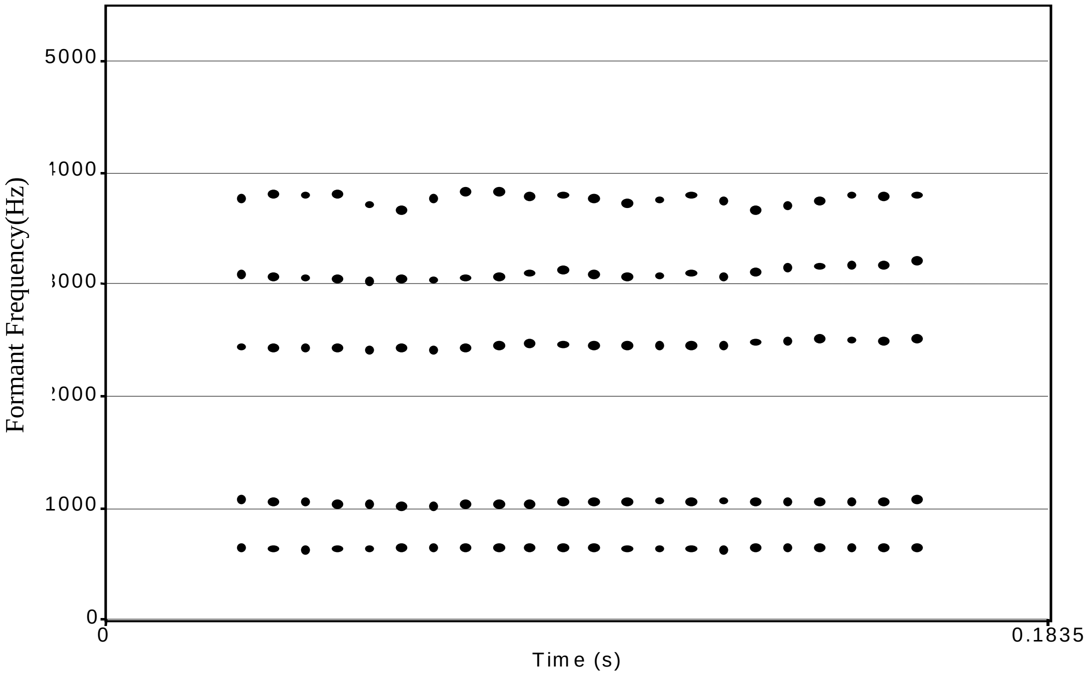


**Fig.3.3** Formant Frequencies Malayalam vowel അ/Λ/.

### 3.3 Vowel Sounds in Malayalam

Vowel sounds are the most interesting class of sounds in any language. The most practical speech recognition systems rely heavily on vowel recognition to achieve high performance [Rabiner.L.R, Levinson.S.E, *et. al.*, 1979], [Rabiner.L.R and Schafer.R.W, 1978], [Rabiner.L.R and Juang.B.H, 1993]. Vowels are produced by exciting a fixed vocal tract with quasi-periodic pulses of air caused by vibration of the vocal cords. Conventional methods used to classify vowels are articulatory configuration required to produce sounds, typical waveform plots, typical spectrogram plots and formant frequency analysis [Gimson.A, 1972], [Rabiner L.R and

Schafer.R.W, 1978]. In this work, we concentrate on the study of nonlinear properties of speech sounds for speech recognition applications. To this end, Malayalam vowel sounds are used for the analysis.

Malayalam is the language spoken predominantly in the state of Kerala, in southern India. It is one of the 23 official languages of India, spoken by around 37 million people. A native speaker of Malayalam is called a ‘Malayali’. Malayalam is also spoken widely in Lakshadweep, Mahe (Mayyazhi), Kodagu (Coorg) Kanyakumari and Dakshina Kannada. The language belongs to the family of Dravidian languages. The language is closely related to Tamil. However, Malayalam has a script of its own, covering all the symbols.

Generally Malayalam phonemes can be classified into vowel sounds and consonant sounds. There are 15 vowels and 36 consonants. The vowels in Malayalam.  are shown in table 3.1 .

<table>
<tr><th colspan="5">Malayalam Vowels</th></tr>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td>അ/Λ/</td><td>ആ/Λ□/</td><td>ഇ/I/</td><td>ഈ/I□/</td><td>ഉ/u/</td></tr>
<tr><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td></tr>
<tr><td>ഊ/u□/</td><td>ഋ/<i>eru</i>/</td><td>എ/ae/</td><td>ഏ/ae□/</td><td>ഐ/ai/</td></tr>
<tr><td>11</td><td>12</td><td>13</td><td>14</td><td>15</td></tr>
<tr><td>ഒ/o/</td><td>ഓ/o□/</td><td>ഔ/au/</td><td>അം/am/</td><td>അഃ/ah/</td></tr>
</table>

**Table 3.1** Vowels in Malayalam

Analyzing carefully, we will realize that, there are only about 8 basic sounds in this list of 15. In other words, half of them are just modifications of these eight. The 8 basic vowels are :.

അ/Λ/, ഇ/ I/, ഉ/u/, ഋ/*eru*/, എ/ae/, ഒ/o/, അം/am/, and അഃ/ah/

Among these 8 basic units, ഋ/*eru*/ is very rarely used. As far as the definition for a vowel sound given in the earlier session is concerned, 'അം/am/' cannot be included in the vowel list, since there is a closure of vocal track during its production. Influence of Sanskrit is very prominent in formal Malayalam. Malayalam has also borrowed a limited amount of Sanskrit words. The unit 'അഃ/ah/' mostly occur only in words accepted from Sanskrit.

Considering these facts, for the study presented in this thesis, we limit our analysis to the five basic vowel units viz. **അ**/Λ/ (as in **A**RUNACHAL), **ഇ/I/ (as** in **E**NGLAND), **എ**/ae/ **(**as in **A**NYONE), **ഒ**/o/ (as in **O**RRISA), and **ഉ**/u/ (as in **U**TTERPREDESH).

The International Phonetic Association (IPA) has produced a set of phonetic symbols to define all of the individual speech sounds (called phonemes) in terms of their place of articulation; for vowel sounds, the tongue position is used. The IPA vowel chart for Malayalam is shown in Figure 3.4.

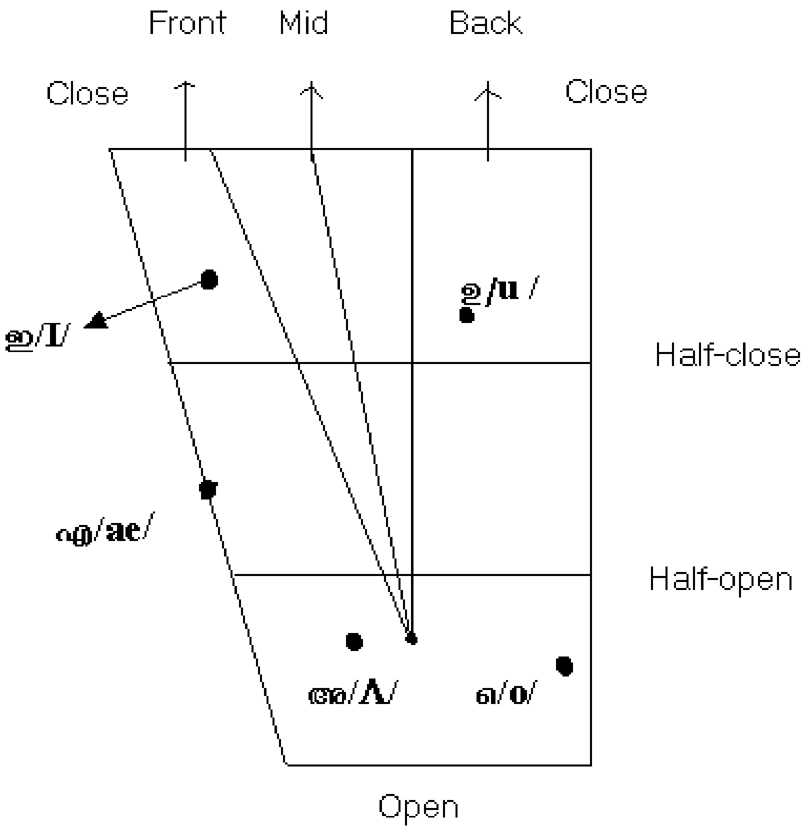


**Fig.3.4** IPA vowel chart for Malayalam vowels

In language terminology, in terms of tongue and hump position, tongue and hump height and typical spectrogram studies, the vowels, ഇ**/I/** and എ**/ae/** are classified as front vowels, ഉ/u/ and ഒ/o/ are classified as back vowels and അ/Λ/ is classified as mid vowel [Velayudhan. S, 1971],

The vowel ഇ**/I/** is a front vowel articulated by raising the rear part of front of the tongue ( i.e. the part of the tongue nearer to center than to the front) in the direction of hard palate, just above the half-close position. The vowel എ**/ae/** is also a front vowel. During the articulation of this vowel, the front of the tongue is raised in the direction of the hard palate to a height between half-close and half-open. The vowel ഒ/o/ is a back vowel and for the articulation of this vowel, the back of the tongue is just above the fully open position. Vowel ഉ/u/ is also back vowel. During the articulation of this vowel the front part of the back of the tongue (i.e. the part nearer to the center than the back of the tongue) is raised in the direction of the soft palate to a height just above half-close position. The vowel അ/Λ/ is a mid vowel. During the articulation of this vowel the center of the tongue is raised in the direction of the roof of the mouth. Figure 3.5 shows the tongue position during the production of front vowel ഇ**/I/,** back vowel ഉ/u/ and the mid vowel അ/Λ/

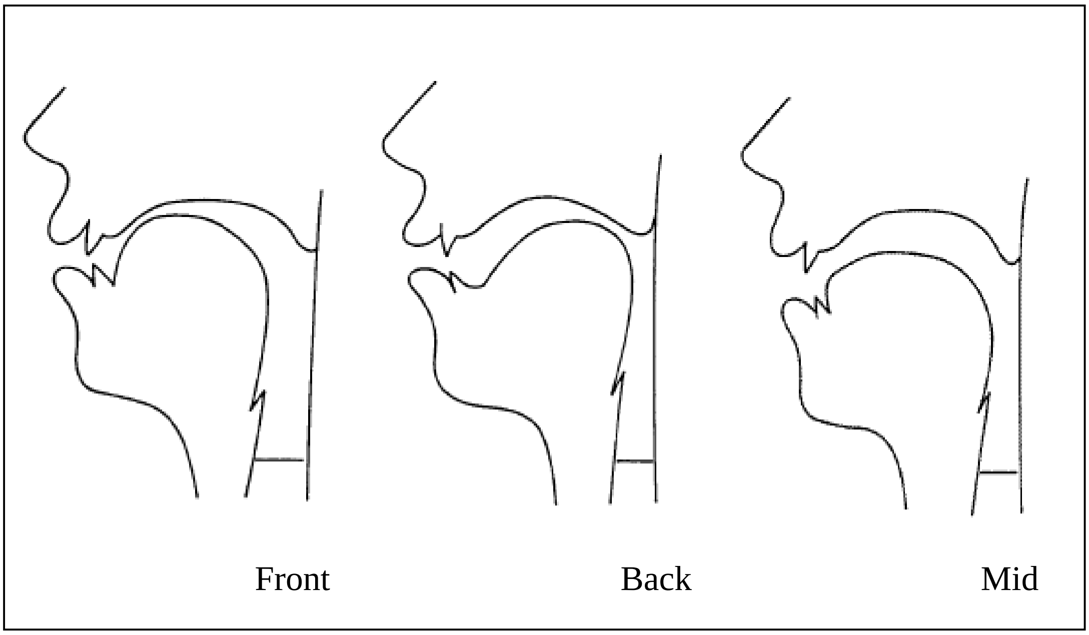


**Fig.3.5** Tongue position during the production of front, back and mid vowels

### 3.4 Speech Data Acquisition

The purpose of the sound card, coming with the personal computers available today is not the accurate sound feature measurement, but the digitization of speech and music. By incorporating an antialiasing presampling filter appropriately, these cards can be modified for speech processing. For the acquisition of data used in this work, we have converted a sound card into one, suitable for speech processing by incorporating an antialiasing presampling low pass filter at the input. The performance of the system with such a modified sound card is comparable to the dedicated high cost systems available in the market.

### 3.4.1 Need for anti aliasing filters

According to the Sampling Theorem, any signal can be accurately reconstructed from values sampled at uniform intervals as long as it is sampled at a rate at least twice the highest frequency present in the signal. Failure to satisfy this requirement will result in *aliasing* of higher-frequency components, meaning that these components will appear to have frequencies lower than their true values. That is if we can exactly *reconstruct* the analog signal from the samples, we must have done the sampling *properl*y. Figure 3.6 shows several sinusoids before and after digitization. The continuous line represents the analog signal entering the ADC, while the square markers are the digital signal leaving the ADC. In fig 3.6(a), the analog signal is a constant DC value, a cosine wave of *zero* frequency. Since the analog signal is simply a series of straight lines between each of the samples, all of the information needed to reconstruct the analog signal is contained in the digital data.

The sine wave shown in Fig 3.6(b) has a frequency of 0.09 times of the sampling rate. This might represent, for example, a 90 cycle/second sine wave being sampled at 1000 samples/second. Expressed in another way, there are 11.1 samples taken over each complete cycle of the sinusoid. These samples properly represent the analog signal because no other sinusoid, or combination of sinusoids, will produce this pattern of samples. These samples correspond to only one analog signal, and therefore the analog signal can be exactly reconstructed. In  fig 3.6(c), the situation is made more difficult by increasing the sine wave's frequency to 0.31 of the sampling rate. This results in only 3.2 samples per sine wave cycle. These samples properly represent the analog waveform. The samples are a unique representation of the analog signal. In fig 3.6(d), the analog frequency is pushed even higher to 0.95 of the sampling rate, with a mere 1.05 samples per sine wave cycle. Here the samples are so sparse that they don't even appear to follow the general trend of the analog signal. These samples do not properly represent the data. The samples represent a *different* sine wave from the one contained in the analog signal. In particular, the original sine wave of 0.95 frequency misrepresents itself as a sine wave of 0.05 frequency in the digital signal. This phenomenon of sinusoids changing frequency during sampling is called **aliasin**g. Just as a criminal might take on an assumed name or identity (an *alia*s), the sinusoid assumes another frequency that is not its own. Since the digital data is no longer uniquely related to a particular analog signal, an unambiguous reconstruction is impossible. This is an example of *improper samplin*g. This line of reasoning leads to a milestone in Digital Signal Processing, the *sampling theorem*. This is called the *Shannon’s* sampling theorem, or the

*Nyquist* sampling theorem, after the authors of 1940s papers on the topic. The sampling theorem indicates that a continuous signal can be properly sampled, *only if it does not contain frequency components above half of the sampling rat*e. For instance, a sampling rate of 2000 samples/second requires the analog signal to be composed of frequencies below 1000 cycles/second. If frequencies above this limit *are* present in the signal, they will be *aliased* to frequencies between 0 and 1000 cycles/second.

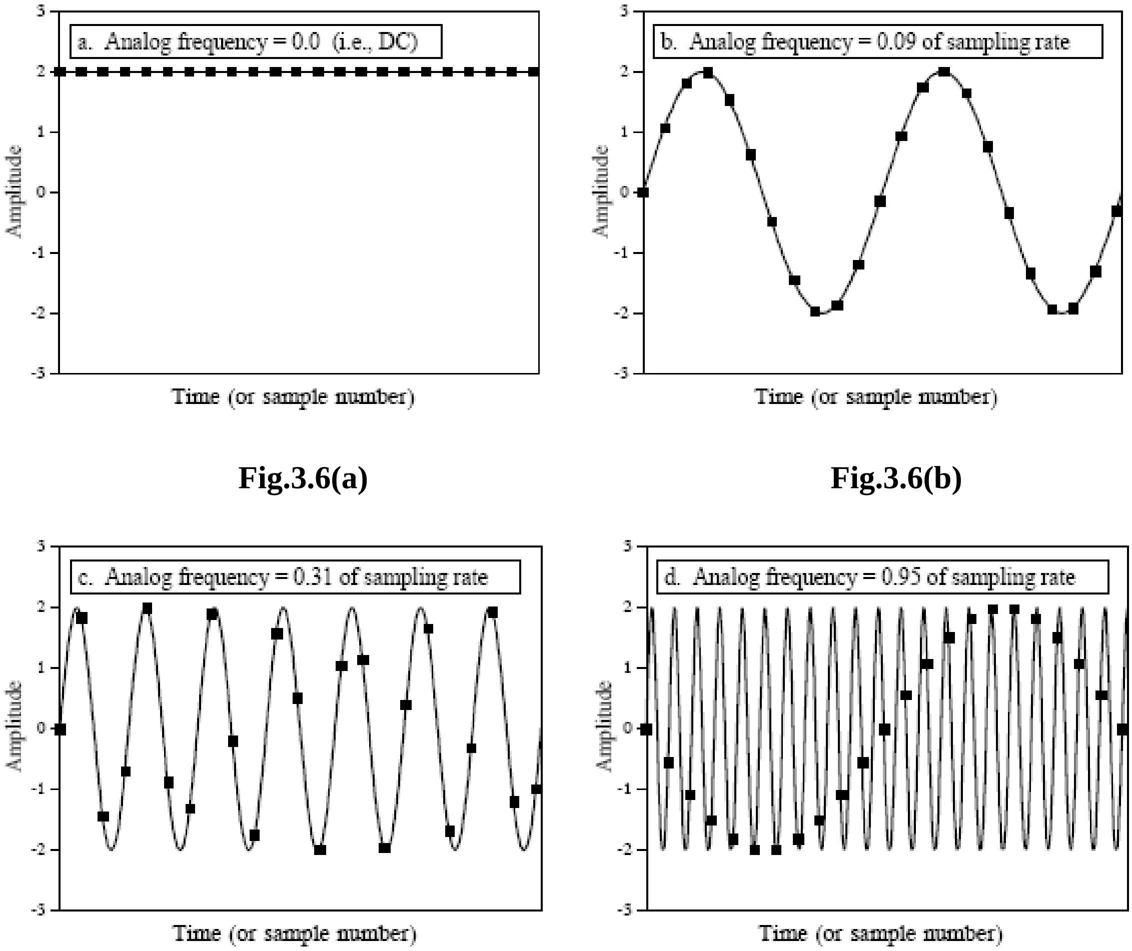


**Fig.3.6(a)** **Fig.3.6(b)**

**Fig.3.6(c)** **Fig.3.6(d)**

**Fig.3.6(a-d)**: Illustration of aliasing in a sinusoidal signal

From the above discussion we conclude that, any signal can be faithfully reproduced from its samples if it is sampled at a sampling frequency, greater than or equal to twice the highest frequency present in the signal. For this, the signal must be a band limited one. Otherwise higher frequency components would be mapped back into the base band and thereby distorts the low frequency information. To avoid this problem of aliasing, it is highly essential to apply a band limiting low-pass filter to the signal, prior to the sampling stage, to remove any frequency components above the "folding" or Nyquist frequency (half the sampling frequency). Such anti-aliasing filters are commonly built into the analog interface chips and *codecs*, which convert analog input signals into digital form for processing, by a digital signal processor. Most of the sound blaster cards available now a days in the market are general purpose cards. It has to handle both speech and music. This card consists generally of one or more A/D and D/A converters. They do not have front-end antialiasing filter circuits to band limit the speech data. So these cards cannot be directly used for speech data acquisition. We can convert such a card into one suitable for speech processing by incorporating an antialiasing filter before the analog to digital converter. We have constructed an 8th order Butterworth low pass filter that can convert our multimedia system to low cost data acquisition system.

### 3.4.2. Speech Processing System

Figure 3.7 shows the block diagram of a typical speech processing system. Before encountering the analog-to-digital converter, the input signal is filtered with an 8th order Butterworth low-pass filter with a cut off frequency of 4 kHz, to remove all frequencies above the Nyquist frequency (since sampling frequency is 8 kHz). This is done to prevent aliasing during sampling, and is correspondingly called an *antialias filter*. The system equipped with antialiasing filter is now suitable for acquisition of speech data. We used the data acquisition system supported by Windows based multimedia system. The recording part of this software consists of option for selecting the sampling frequency and bits per sample for quantization. The sound files are recorded in the software as Windows wave format (.wav). This wave file can be accessed directly by Mathlab 7, the language used for implementing the software part in this work. On the other end, the digitized signal is passed through a digital-to-analog converter and another low-pass filter. This output filter is called a *reconstruction filter*.

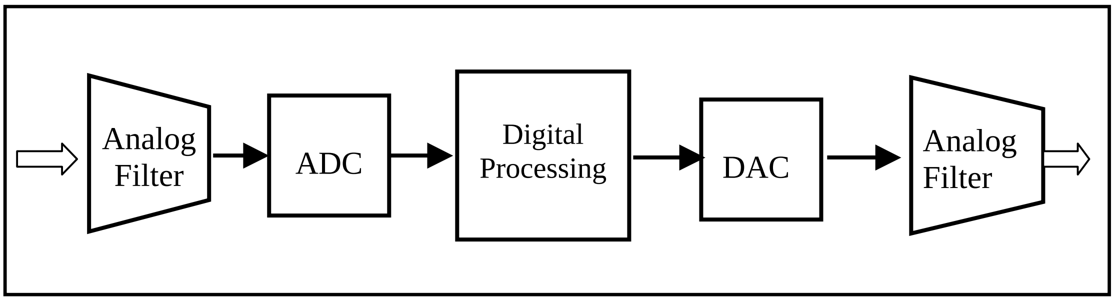


**Fig.3.7** Block diagram of Digital Speech Processing system

### 3.4.3. Speech Database Creation

The speech signal, low pass filtered to 4 kHz and sampled at 8 kHz is digitized using 16-bit A/D converter. A speech database is created using the above data acquisition system. As explained in the earlier session, for the analysis of nonlinear properties, we have used the five basic short vowel units in Malayalam viz: അ/Λ/, ഇ**/I/**, ഉ/u/, എ**/ae/**, and ഒ/o/. The database created consists of all the five short vowels in Malayalam. Each vowel is uttered 500 times by the author at different occasions, digitized and stored in separate data files. The file name of the data reveals the vowel identity, speaker identity, and repetition. The structure of the file name is given below

XX , YYY… ,ZZZ

The first XX indicates the vowel identity. For example, for the vowel ഉ /u/, this part is UX and for എ**/ae/** this is EA. The second part, YYY indicates the repetition of the vowels. For example, 001,002 etc. A provision is also included to create multi speaker (both male and female) database for future research work. The third part, ZZZ indicates the speaker identity. For example, M01, F01 represent the first male speaker and first female speaker respectively. The time domain representations of the five vowel units from the created database are given in figures 3.8(a-e)

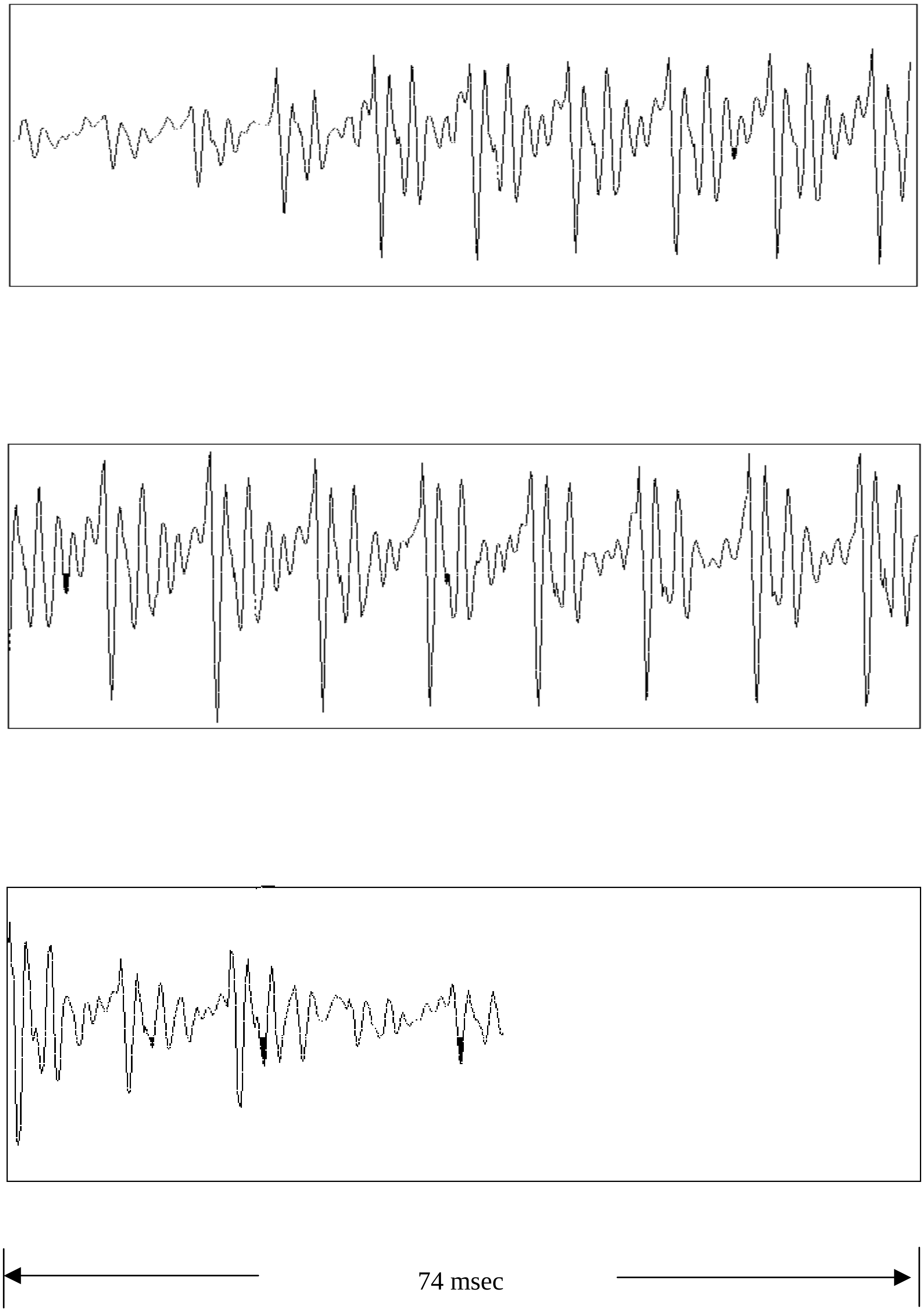


**Fig.3.8(a)**. Speech waveform of vowel അ/Λ/ - (Data file: AX001M01)

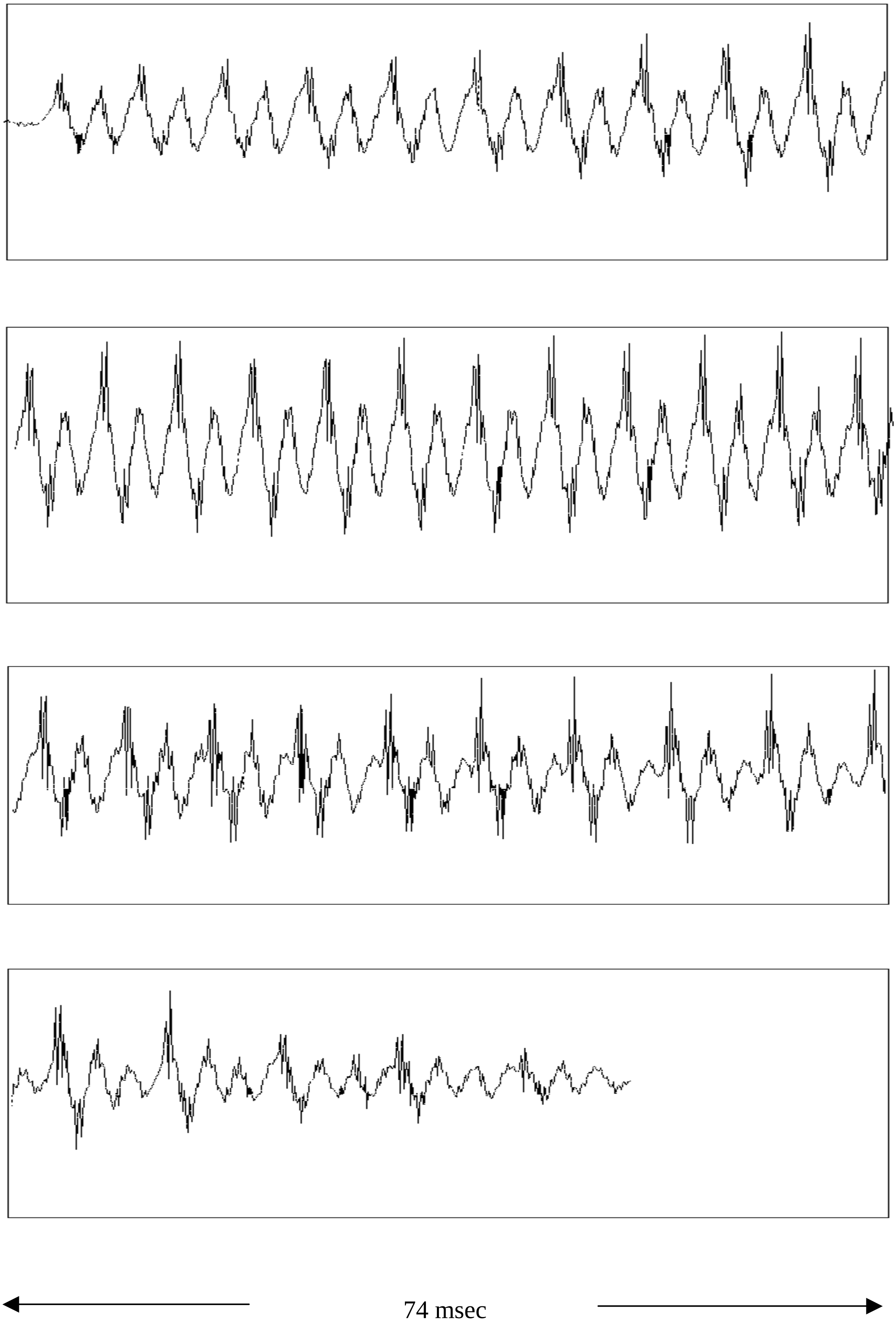


**Fig.3.8(b)**. Speech waveform of vowel ஐ**/I/ -** (Data file: EX001M01)

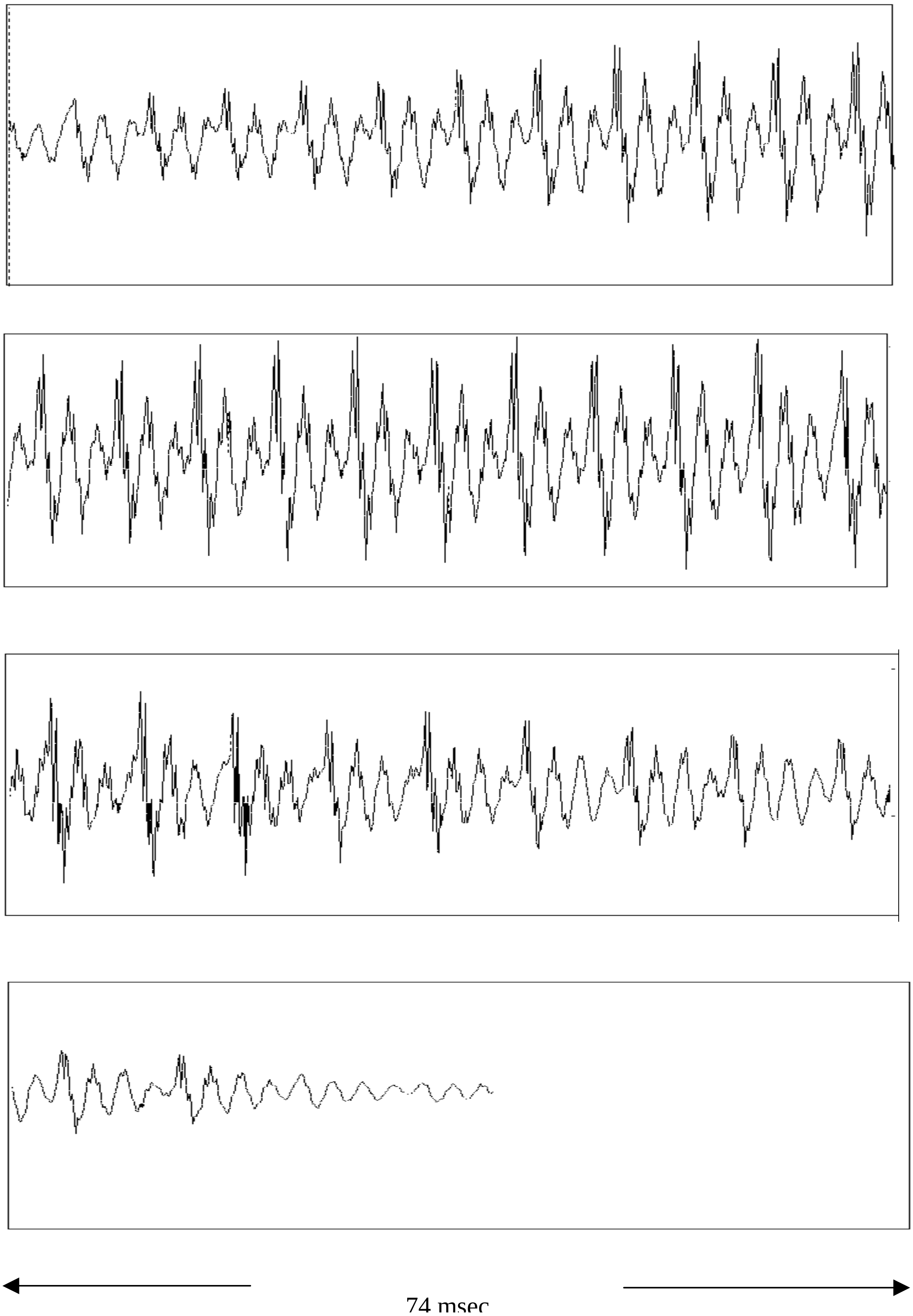


**Fig.3.8(c)**. Speech waveform of vowel ഐ**/ae/ -** (Data file: EA001M01)

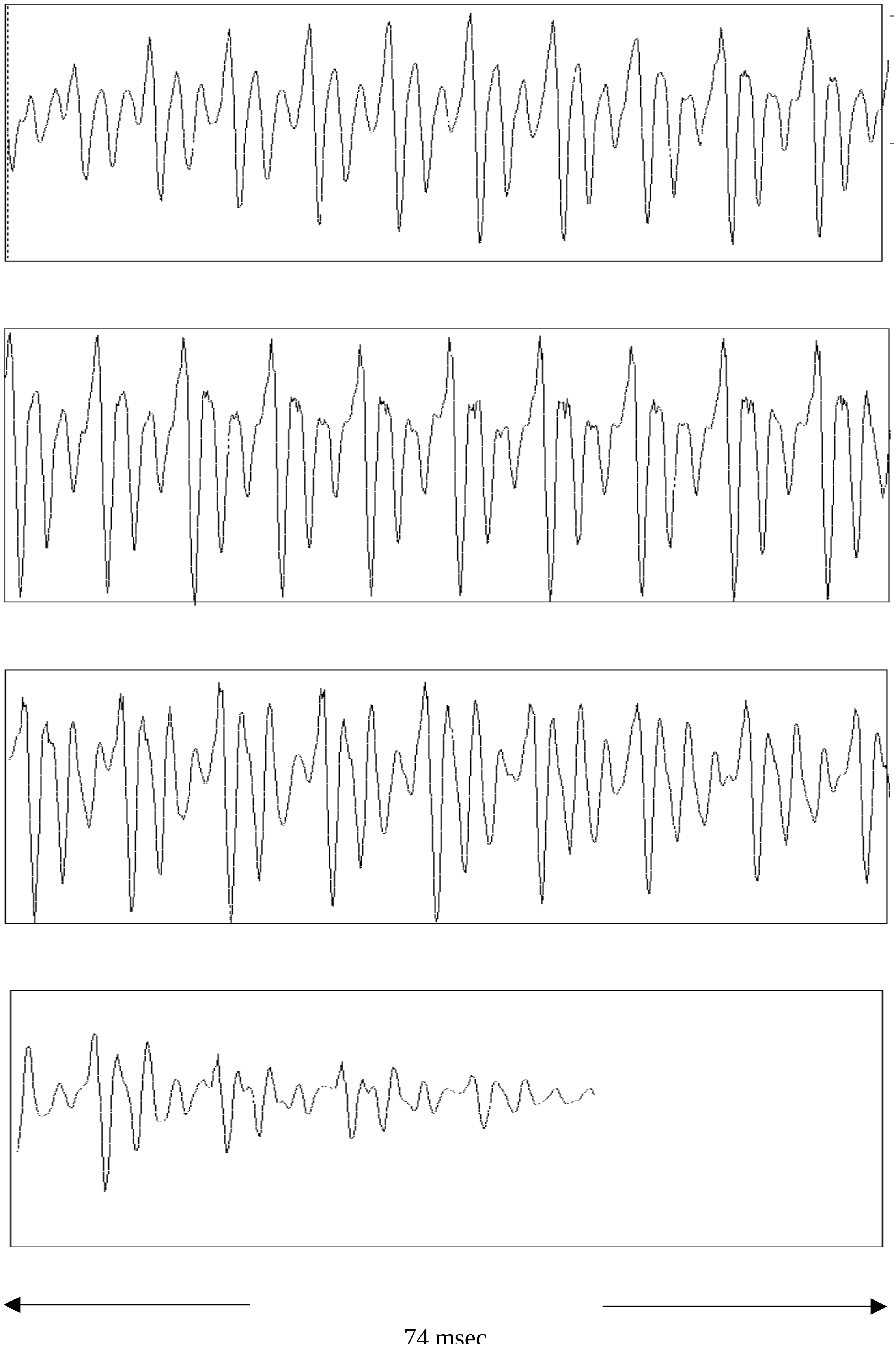


**Fig.3.8(d)**. Speech waveform of vowel ഒ/o/ **-** (Data file: OX001M01)

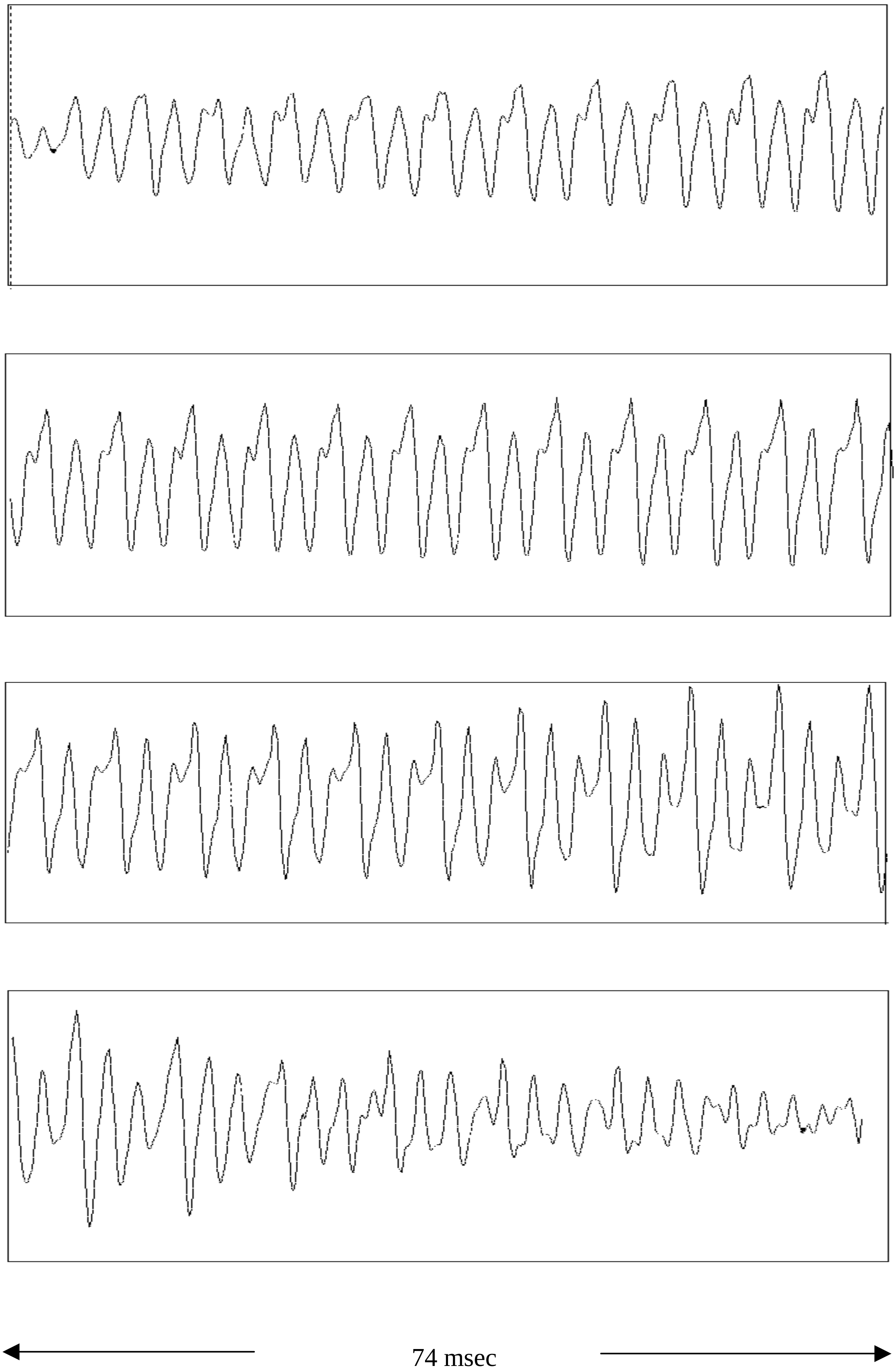


**Fig.3.8(e)**.Speech waveform of vowel உ/u/ **-** (Data file: UX001M01)

### 3.5 Conclusion

In this chapter a brief overview of articulatory and acoustic phonetics has been presented. This provides a framework for later discussions of speech production and the speech signal, especially concentrating on vowel sounds and their characterization. Most of the sound blaster cards available in the market do not have front-end antialiasing filter circuit to band limit the speech signal. So this card cannot be directly used for speech data acquisition. Here an 8th order analog antialiasing presampling filter with a low pass cut off 4 kHz is connected prior to the ADC card for removing the frequency component above 4kHz. A speech database of short vowels in Malayalam is created using the data acquisition system developed, for speech analysis and recognition study described in the following chapters.

# Chapter 4
# Nonlinear dynamical Invariants of Speech

## 4.1 Introduction

For several decades, the traditional approach to speech modeling has been the linear (source-filter) model. This model forms the foundation of many speech-processing applications such as speech coding, speech synthesis, speech recognition, and speaker recognition technology. Here the true nonlinear nature of speech production are approximated via the standard assumptions of linear acoustics and one dimensional (1-D) plane wave propagation of the sound in the vocal tract. There is strong theoretical and experimental evidence for the existence of important nonlinear 3D fluid dynamics phenomena during the speech production that cannot be accounted for by the linear model. Examples of such phenomena include modulations of the speech airflow and turbulence. Teager and Teager present several physical measures that show turbulences in the airflow [Teager H. M. and Teager S. M, 1989]. We can view the linear model only as a first order approximation to the true speech acoustics, which also contain second-order and nonlinear structure. The ‘standard’ speech features used in automatic speech recognition (ASR) are based on short-time smoothed cepstra stemming from the linear model. Therefore adding new robust nonlinear information should give promising results.

The dynamics of a system can be studied by extracting invariant parameters from the experimental time series data. In this chapter we investigate the features, known as invariants of non-linear dynamical system, that measure the non-linearity in a signal. We analyze three popular measures: Box counting dimension, Correlation dimension and Kolmogorov entropy. These measures quantify the presence (and extent) of chaos in the underlying system that generated the speech.

This chapter is organized as follows. Firstly the theory of dynamical systems is presented to provide a background for subsequent work. Following this, speech production mechanism is demonstrated to be a nonlinear process, giving the motivation to pursue speech processing from this perspective. Then a detailed description is presented for the invariant parameters – Box counting dimension, Correlation dimension and Kolmogorov entropy. Finally these parameters are evaluated for speech signals.

## 4.2 Nonlinear Dynamics

Emerging from the theoretical development of nonlinear dynamics, the study of chaotic dynamics in deterministic systems has become very popular in the past few decades. Perhaps it is because of the amazing findings that the study of chaotic systems has delivered. Interest in chaos (or more generally nonlinear dynamics) grew rapidly after 1963, when Lorenz published its implications on weather prediction.

It would be natural to think that if a system is deterministic, its behavior should be easily predicted. But there are systems where their behavior turns out to be non-predictable: not because of lack of determinism, but because of the complexity of the underlying dynamics that require an infinite precision that is unable to be computed. This can be seen in systems where very similar initial conditions yield very different behaviors, even though the systems obey deterministic laws. No matter how much precision we have, the most minimal differences will tend in the long time to very different results, or the system is highly sensitive to initial conditions. The extreme 'sensitivity to initial conditions' mathematically present in the systems despite the deterministic laws is called dynamical instability, or simply chaos. It occurs in mechanical oscillators such as pendula or vibrating objects, in rotating or heated fluids, in laser cavities, in electronic circuits and in some chemical reactions.

If prediction becomes impossible, it is evident that a chaotic system can resemble a stochastic system ( a system subject to random external forces). However, the source of the irregularity is quite different. For chaos, the irregularity is part of the intrinsic dynamics of the system, not because of unpredictable outside influences. Necessary conditions for chaotic motion are that (a) the system has at least three independent dynamical variables, and (b) the equations of motion contain a nonlinear term that couples several of the variables [Ott. E and Sauer. T, 1994].

### 4.2.1 Dynamical Systems theory

Dynamical theory provides a framework with which nonlinear deterministic systems can be examined. Important steps in the analysis are to find the system dimension; reconstruct from available data a representation of the system; investigate if chaos is present and determine the predictability of the system. The theory is now well enough established that a number of excellent tutorial papers [Eckmann.J.P and Ruelle.D, 1985], [Crutchfield.J, Farmer.J, *et. al.*, 1986], [ Parker.T.S and Chua.L.O, 1987], [Kearney. M. J. and Stark. J, 1992], and books [Hilborn. R, 1994], [Ott.E, Sauer.T *et.al.*, 1994], [Alligood.K, Sauer.T *et. al.*, 1997] exist, which give a thorough background to the subject. The following will be limited to an attempt to show some basic principles which will lead on to subsequent work presented in this thesis.

A dynamical system may be defined, for the continuous time case, by the equation:

$$\dot{\mathrm{X}} \quad = \mathcal{F}(\mathrm{X}) \tag{4.1}$$

where X is a vector in ‘d’ dimensional space, $\mathcal{F}$ is some function (linear or nonlinear) operating on X , and $\dot{\mathrm{X}}$ is the time derivative of X . This system is deterministic, in that it is possible to completely specify its evolution, or flow of trajectories in the d dimensional space, given the initial starting conditions.

Now consider a simple system that of a point mass on an ideal spring, that obeys Hooke’s Law. The position, x(t), of this continuous time system is given by

$$\mathrm{x(t)} = \mathrm{x}_0 \cos\omega \mathrm{t} + \frac{\dot{\mathrm{x}}_0}{\omega}\sin\omega \mathrm{t} \tag{4.2}$$

where

$$\omega = \sqrt{\frac{k}{m}}$$

$\omega$ is the angular frequency of oscillations, k is the spring constant, m is the mass of the particle. $x_0$ and $\dot{x}_0$ are the initial conditions at time t= 0.

The velocity of the system is found by differentiating the equation 4.2

$$\dot{x}(t) = -\omega x_0 \sin\omega t + \dot{x}_0 \cos\omega t \tag{4.3}$$

The position and the velocity completely specify the system.

Plotting x against $\dot{x}$ gives the phase portrait, as shown in Figure 4.1. This plot is also called the phase space or state space of the system, since it can be used to specify the state of the system at any moment in time.

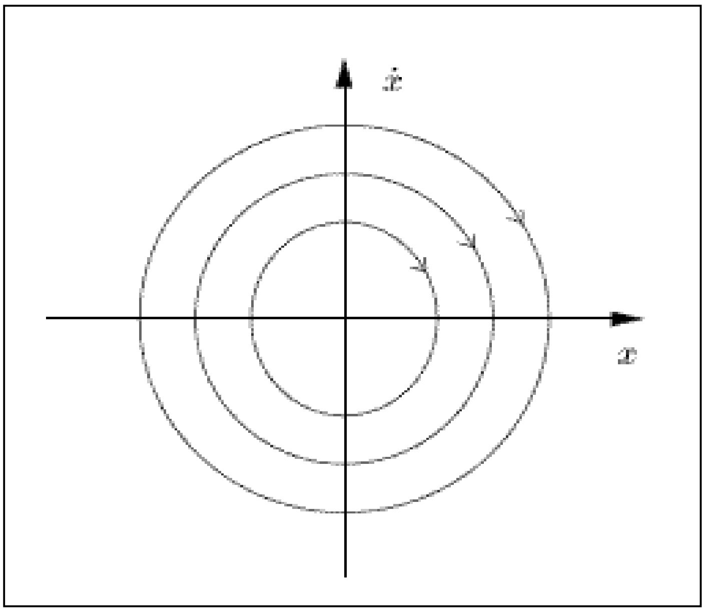


**Fig. 4.1** Phase portrait for a point mass on an ideal spring

If the spring was not ideal, then energy would be dissipated from the system through time. This would lead to a phase portrait of the type shown in Figure 4.2(a). When all energy has been dissipated, the particle will be at rest, corresponding to the origin of the phase portrait. Hence this system has an *attractor* at the origin, since any damped harmonic oscillator will converge towards it. This is known as a point attractor. In contrast, the addition of some driving force will lead to a closed curve, or limit cycle attractor, as seen in Figure 4.2(b).

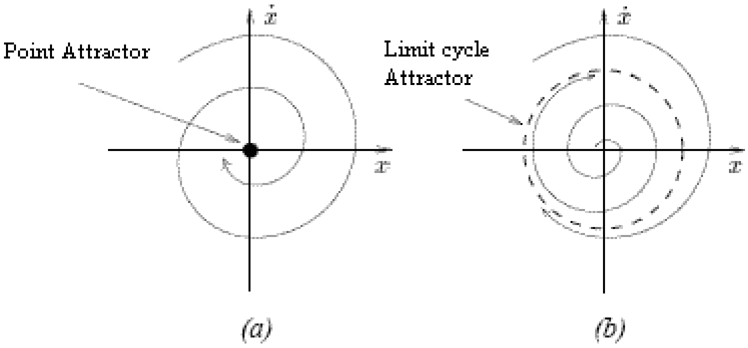


**Fig.4.2** (a) Point attractor and (b) Limit cycle attractor for a harmonic oscillator

Alternatively, for discrete time, the dynamical system can be defined as a map:

$$X_{n+1} = \mathcal{G}\,(X_n) \qquad (4.4)$$

where $X_n$ is again the d length vector, at time step n, and $\mathcal{G}$ is the operator function. Given the initial state, $X_0$, it is possible to calculate the value of $X_j$ for any $j > 0$.

A simple example of a map is the quadratic map

$$X_{n+1} = \alpha - X_n^2 \qquad (4.5)$$

where $\alpha$ is a control parameter. Plotting iterations of $X_{n+1}$ against $X_n$ gives the evolution of the system, as seen in Figure 4.3. This type of plot is known as a phase portrait. Here knowledge of the variable $X_n$ completely specifies the system. Detailed description of Phase space is given in the next chapter.

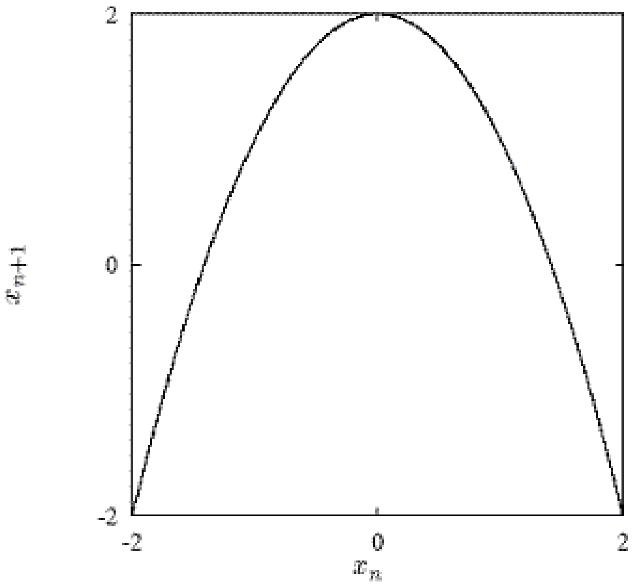


**Fig.4.3** Phase portrait of the quadratic map with $\alpha = 2$

It is somewhat surprising to find that certain dynamical systems, with system equations not much more complicated than the previous examples, can exhibit extremely complex behavior (Eg. Lorenz equations). If a variable is available for observation, the resulting time series may seem entirely random; such systems were often called random and unpredictable. The discovery of chaos has changed this viewpoint dramatically.

How can a deterministic dynamical system exhibit such behavior? The answer lies in the sensitivity to initial conditions [Kearney.M.J and Sark.J, 1992]. In the phase portrait of such systems, two trajectories that are close to each other at some time will eventually become separated. If a system is allowed to evolve from two sets of initial conditions that are very close together (to machine precision), then after some time there will be exponential divergence for the trajectories, hence leading to differing behavior of the system. This phenomenon, known as sensitive dependence on initial conditions, is very common to chaotic systems. Just a small change in the initial conditions can drastically change the long-term behavior of a system. Such a small amount of difference in a measurement might be considered experimental noise, background noise, or an inaccuracy of the equipment. Such things are impossible to avoid in even the most isolated lab.

### 4.3 Nonlinear Dynamical Aspects of Speech

Speech generation has classically been modelled as a linear system, which provides a convenient and simple mathematical formulation. However, a number of nonlinear effects are present in the physical process, which limit the effectiveness of the linear model. An improved approach may be to view speech as an output of a nonlinear dynamical system.

Speech production is an extremely complex process. It can be described in basic terms as follows. The lungs generate air pressure and this pressure wave is modulated as it flows through the larynx. Essentially, the larynx is made up of two almost symmetric masses known as the vocal folds, which are capable of closing completely together and move apart.

While moving apart there creates a triangular opening called the glottis. During normal respiration and the production of unvoiced sounds, air passes freely through the glottis. When the vocal folds vibrate in a quasi-periodic manner then voiced sounds are produced. The frequency of this excitation is known as the fundamental frequency. The resulting glottal waveform excites the vocal tract, which is the region extending from the larynx to the lips. Different configurations of the vocal tract will result in different modulations of the glottal waveform and thus produce specific sounds [Rabiner.L.R and Juang.B.H, 1993].

The speech production process is generally modelled by the linear source–filter speech model, as shown in Figure 4.4 [Fant.G, 1960]. There is a hard switch between voiced and unvoiced sounds. The glottal waveform for voiced sounds is generated from an impulsive periodic signal, with period equal to the pitch period, whereas the generation of unvoiced sounds is modelled by white noise. These signals are then applied to a slowly time varying linear filter whose transfer function represents the contribution of the vocal tract, followed by another filter to represent the radiation from the lips.

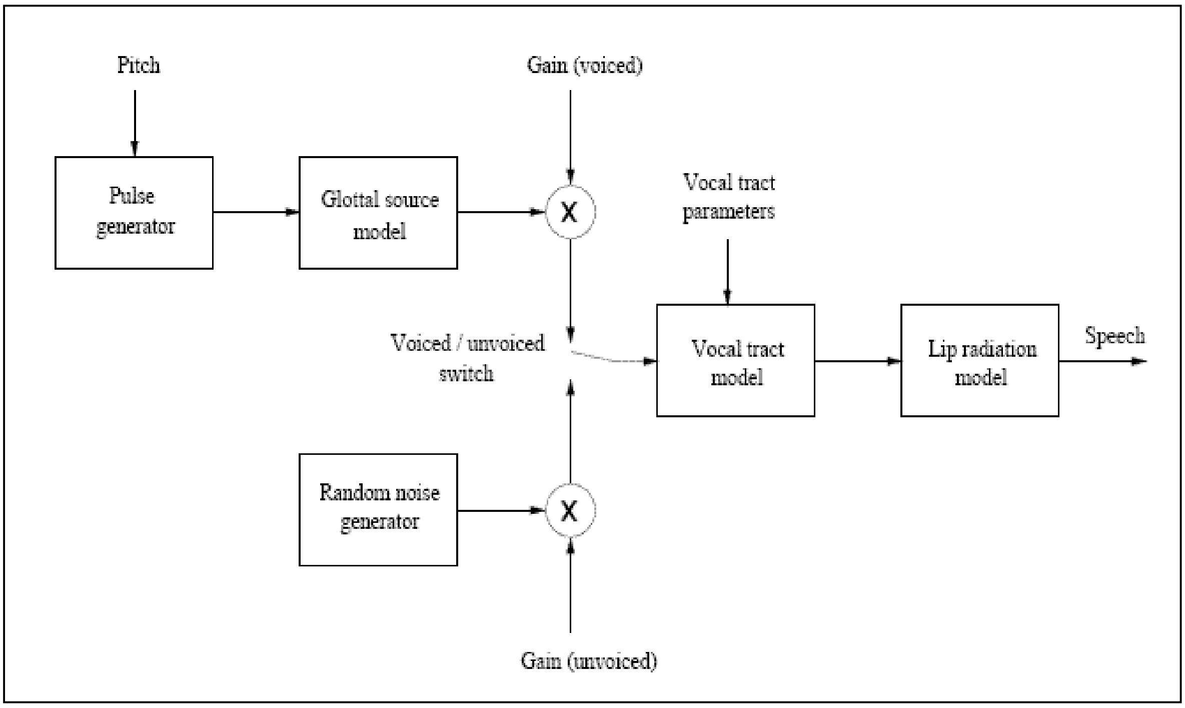


**Fig.4.4** Linear Source-filter model of Speech Production.

In order to arrive at this simplified model, a number of major assumptions are made. These include:

- The vocal tract and speech source are uncoupled – thus allowing source-filter separation
- Airflow through the vocal tract is laminar
- The vocal folds vibrate in an exactly periodic manner during voiced speech production
- The configuration of the vocal tract will only change slowly

But today a number of nonlinear effects in the speech production process are known. Firstly, it has been accepted for some time that the vocal tract and the vocal folds do not function independently of each other, but that there is in fact some form of coupling between them when the glottis is open [Koizumi.T, Taniguchi.S, *et.al.*, 1985]. This can cause significant changes in formant characteristics between open and closed glottis cycles. [Brookes.D.M and Naylor.P.A, 1988]. More controversially, Teager and Teager [Teager.H.M and Teager.S.M, 1989] have claimed (based on physical measurements) that voiced sounds are characterised by highly complex air flows in the vocal tract, rather than well behaved laminar flow. Turbulent flow of this nature is also accepted to occur during unvoiced speech, where the generation of sound is due to a constriction at some point in the vocal tract. In addition, the vocal folds will themselves be responsible for further nonlinear behaviour, since the muscle and cartilage, which comprise the larynx, have nonlinear stretching qualities. Such nonlinearities are routinely included in attempts to model the physical process of vocal fold vibration, which have focussed on two or more mass models [Ishizaka.K and Flanagan.J.L, 1972], [Koizumi.T, Taniguchi.S, *et.al.*, 1987] [Steinecke.I and Herzel.H, 1995] in which the movement of the vocal folds is modelled by masses connected by springs, with nonlinear coupling. Observations of the glottal waveform reinforce this evidence, where it has been shown that this waveform can change shape at different amplitudes [Shoentgen.J, 1990]. Such a change would not be possible in a strictly linear system where the waveform shape is unaffected by amplitude changes.

Taking into account the evidence presented here for the nonlinearities in speech, the source-filter model clearly has drawbacks. The assumptions made imply a loss of information, which although allowing a simple, mathematically tractable linear model, the full speech signal dynamics can never be properly captured. If instead of choosing a linear model, the nonlinear option is pursued, then it is not necessary to make many of the assumptions that are required in the linear case. So Speech is modelled as the output of a nonlinear dynamical system, which only has a low number of active degrees of freedom. So it is possible to model them using a nonlinear dynamical approach by:

$$X_{i+1} = \mathcal{F}(X_i)$$

where $\mathcal{F}$ is some nonlinear mapping between previous samples $X_i$ and the next sample $X_{i+1}$. By Taken's theorem, there will be a one–to–one mapping between $\mathcal{F}$ and the underlying nonlinear system. Therefore correctly modeling $\mathcal{F}$ implies that the dynamics of speech production have been captured.

The fact that the speech signal is produced by a nonlinear dynamical system that often generates small or large degrees of turbulence, motivated our study of nonlinear aspects of speech processing. In the following section we investigate the features, known as invariants that measure nonlinearities in a signal.

## 4.4 Invariant Characteristics of the Dynamical System

Suppose we are analyzing a signal with the help of Fourier spectrum, where the source of signal is linear. We see a spectral peak at some frequency. Then we know that if we were to simulate the same system at a different time with a different forcing strength, we would see the spectral peak in the same location, possibly with a different integrated power. That is the Fourier frequency is an invariant of the system evolution. The phase associated with that frequency depends on the time at which the measurements begin and the power under the peak depends on the strength of the forcing..

Since chaotic motion produces continuous, broadband Fourier spectra, we clearly have to replace narrowband Fourier features with other invariant features or characteristics of the system for the purpose of identification and classification. The major invariant features include attractor dimensions and Kolmogorov entropy.

### 4.4.1 Fractal Dimension of the Attractor

Informally the attractor 'A' of a dynamical system is the subset of phase space toward which the system evolves. An initial condition $x_0$, that is sufficiently near the attractor will evolve in time so that $\phi(x_0)$ comes arbitrarily close to the set A as $t \rightarrow \infty$. That is attractors are defined as a broad subset of the Phase Space, in which the orbits or trajectories reside as $t \rightarrow \infty$ [Ott. E, 1993]. Fractal dimensions are characteristics of the geometric figure of the attractor and relate to how the points on the attractor are distributed in phase space. If the attractor *A* is a fractal, then the attractor is said to be strange. A strange attractor is having non-integer dimension in phase space. To understand the possibility of non-integer dimension requires a more sophisticated concept of dimension than that associated with lines, surfaces and solids.

To characterize the structure of the underlying strange attractor from an observed time series, it is necessary to reconstruct a phase space from the time series. This reconstructed phase space captures the structure of the original system's attractor (the true state-space that generated the observable). The process of reconstructing the phase space that underlies the system's attractor is commonly referred to as embedding.

The simplest method to embed scalar data is the method of delays. In this method, the pseudo phase-space is reconstructed from a scalar time series, by using delayed copies of the original time series as components of the Reconstructed Phase Space (RPS). It involves sliding a window of length d through the data to form a series of vectors, stacked row-wise in the matrix. Each row of this matrix is a point in the reconstructed phase-space. Let $\{x_n\}$ represent the row vectors. The row vectors can be compiled into a matrix called trajectory matrix and is represented as:

$$\mathbf{X} \quad = \quad \begin{bmatrix} x_1 & x_{1+\tau} & x_{1+2\tau} & \cdots\cdots & x_{1+(d-1)\,\tau} \\ x_2 & x_{2+\tau} & x_{2+2\tau} & \cdots\cdots & x_{2+(d-1)\,\tau} \\ x_3 & x_{3+\tau} & x_{3+2\tau} & \cdots\cdots & x_{3+(d-1)\,\tau} \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ x_N & x_{N+\tau} & x_{N+2\tau} & \cdots\cdots & x_{N+(d-1)\tau} \end{bmatrix}$$

where d is the embedding dimension and τ is the embedding delay. Detailed study on various aspects of Reconstructed Phase Space is presented in the next chapter.

There are several ways to generalize dimension to the fractional case, like capacity or box counting dimension, correlation dimension, information dimension, Lyapunov dimension etc. In the present study, we have analyzed box counting and correlation dimensions of speech signal.

### 4.4.1.1 Box counting Dimension

There are many ways to define the dimension, d(A), of a set A. One approach is the Box counting dimension or capacity dimension $d_B$. Consider a one-dimensional figure such as a straight line or curve of length L. This line can be ‘covered by N(ε) one dimensional boxes of size ε. If L is the length of the line then,

$$N(\varepsilon) = L\,(1/\varepsilon) \qquad (4.6)$$

Similarly, a two dimensional square of side L can be covered by $N(\varepsilon) = L^2(1/\varepsilon)^2$ boxes. For a three dimensional cube, the exponents would be 3, and so on for higher dimensions. In general,

$$N(\varepsilon) = L^{d}\,(1/\varepsilon)^{d} \tag{4.7}$$

Taking logarithms one obtains

$$d = \frac{\log N(\varepsilon)}{\log L + \log(1/\varepsilon)} \tag{4.8}$$

and in the limit of small ε, the term involving L becomes negligible. The box counting dimension is defined as

$$d_B = \lim_{\varepsilon \to 0} \frac{\log N(\varepsilon)}{\log(1/\varepsilon)} \tag{4.9}$$

Some of the first numerical estimates of fractal dimension were obtained with the box-counting method [Ott.E, 1993]. In this section some of the practical issues that arise in the implementation of box-counting algorithms are discussed. To compute the box-counting dimension, break up the embedding space into a grid of boxes of size ε. Count the number of boxes N(ε), inside which, at least one point of the attractor lies. The box-counting dimension is formally defined by the limit in Eq. (4.9). There are some obstacles, however, in applying Eq. (4.9) directly to an experimentally observed signal.

In practical situations, only a finite resolution is available, so the limit $\varepsilon \to 0$ cannot be taken. A natural and direct approximation is just to apply Eq. (4.9) directly but with the smallest ε available. That is,

$$d_B \approx \frac{\log N(\varepsilon)}{\log(1/\varepsilon)} \tag{4.10}$$

An equivalent approach is to regard $d_B$ as the slope of the log N versus log(1/ ε) curve for ε is very small.

In computing the box-counting dimension, one either counts or does not count a box according to whether there are some points or no points in the box. No provision is made for weighting the box count according to how many points are inside a box. In other words, the geometrical structure of the fractal set is analyzed but the underlying measure is ignored. For experimental data another type of dimension is more efficient to compute than the box-counting dimension. This is the correlation dimension ($d_C$) [Grassberger and Procaccia ,1983].

### 4.4.1.2 Correlation dimension.

Suppose that many points are scattered over a set. The typical number of neighbors of a given point will vary more rapidly with distance from that point if the set has high dimension than otherwise. The correlation dimension $d_C$ measures the variation of the average fraction of the neighboring points with distance. It may be computed from the correlation integral $C_d(R)$ in the d dimensional phase space defined by: [Grassberger and Procaccia, 1983].

$$C_d(R) = \lim_{N \to \infty} \left[ \frac{1}{N^2} \sum_{i,j=1}^{N} H(R - |x_i - x_j|) \right] \qquad (4.11)$$

where $x_i$ and $x_j$ are points on the attractor, H(y) is the Heaviside function ( H(y) = 1 if $y \geq 0$ and 0 if $y < 0$), and N is the number of points randomly chosen from the entire data set. The Heaviside function simply counts the number of points within a radius R of the point denoted by $x_i$ , and $C_d(R)$

gives the average fraction of points within R. The correlation dimension $d_C$is defined by the variation of $C_d(R)$ with R: [Grassberger and Procaccia ,1983].

$$C_d(R) = R^{dc} \text{ as } R \to 0 \tag{4.12}$$

Therefore $$d_C = \lim_{R \to 0} \frac{\log[C_d(R)]}{\log R} \tag{4.13}$$

When R approaches the size of the phase space, $C_d(R)$ saturates at unity since all points are then included in the range R. On the other hand when R is smaller than the spacing between the data points, only one point lies in the range, and $C_d(R)$ levels off at $1/N^2$

The quantities $d_B$ and $d_C$ are not equivalent. The box-counting dimension depends only on whether small elements of phase space contain any points and does not take into account the differing numbers of points in the various elements. That is, small scale variations of the density of points are ignored. On the other hand, the correlation dimension does include this effect. Because of these differences, $d_C$ is called metric dimension and $d_B$ is called frequency dimension.

One reason for the popularity of the correlation algorithm is that it is easy to implement. One computes the correlation integral $C_d(R)$ merely by counting distances. However, there are a variety of practical issues and potential pitfalls that come with making an estimate from finite data. The limit $N \to\infty$ in the calculation of $C_d(R)$ will get us into trouble whenever we have a finite sample: N is limited by the sample size. The meaningful choice for 'R' is limited by the inevitable lack of near neighbors at small length scales. Practically $d_C$ is calculated as the slope of log $C_d(R)$ versus log R curve.

### 4.4.2 Kolmogorov Entropy

The concept of the entropy is fundamental for the study of statistical mechanics and thermodynamics. Entropy is a thermodynamic quantity describing the amount of disorder in the system. One can generalize this concept to characterise the amount of information stored in a more general probability distribution. This is in part what information theory is concerned with. The theory has been developed since 1940s and the main contributions came from Shannon, Renyi, and Kolmogorov.

Information theory provides an important approach to time series analysis. The complex appearance of the various graphical representations of chaotic behaviour naturally leads to the question of the relationship between statistical mechanics and chaos. One way to connect these phenomena is to apply the concept of entropy to a chaotic system. Consider a hypothetical statistical system for which the outcome of a certain measurement is located on the unit interval. If the interval is subdivided into M subintervals, we can associate a probability $p_i$ with the i$^{\text{th}}$ subinterval containing a particular range of possible outcomes. The entropy of the system is then defined as

$$\mathrm{S} = -\sum_{\mathrm{i}=1}^{\mathrm{M}} p_i \log_{\mathrm{e}} p_i$$

This quantity may be interpreted as a measure of the amount of disorder in the system or as the information necessary to specify the state of the system. If the subintervals are equally probable so that $p_i$ = 1/M for all i, then the entropy reduces to S = $\log_e$M, which can be shown to be its maximum value. Conversely, if the outcome is known to be in a particular subinterval, then S = 0, the minimum value. When S = $\log_e$M, the amount of further information needed to specify the result of a measurement is at a maximum. On the other hand, when S = 0, no further information is required. The entropy is also interpreted as ‘missing’ information.

In a dynamical system the entropy is defined through a partition of the phase space with the probability $p_i$ given by the integral of the invariant measure ρ(x) (probability density) over the $i^{\text{th}}$ element. Define the partition β = {$B_i$}; $i$ = 1…….. M with $B_i$ non-empty, non-intersecting sets that cover the attractor. Then the probability $p_i$ of finding a point in the box $B_i$ is $\int \rho(\mathrm{x})\mathrm{dV_i}$, where $\mathrm{dV_i}$ is an arbitrary small elemental phase space volume. The entropy of the partition of the dynamical system is computed using

$$\mathrm{S} = -\sum_{\mathrm{i}=1}^{\mathrm{M}} p_i \log_{\mathrm{e}} p_i$$

This tells us about the uncertainty coming from the “random” aspect of the dynamics. This quantity is not immediately useful, since it depends on the scheme of partitioning (e.g. the box size) as well as intrinsic properties of the attractor. Two related quantities have been defined to give intrinsic properties. One is the scaling of the entropy as the box size of the partition is reduced. This defines the “information density” of the attractor and is called Information dimension of the attractor $d_I$. The information density is a static property of the attractor [Farmer.J.D and Naturforsch. Z, 1982].

The second quantity tells us how the uncertainty or information of the system evolves in time, or under iteration for a map. This is known as the Kolmogorov *or* Kolmogorov-Sinai *or* metric entropy (K) [Farmer.J.D,1982]. The Kolmogorov Entropy is defined as,

$$K = \lim_{M \to \infty} \frac{1}{M} S(\beta_M)$$

Consider a dynamical system can be defined by an iterative map:

$$X_{n+1} = \mathcal{G}(X_n)$$

where $X_n$ is the d length vector, at time step n. Suppose we know the initial value $X_0$ to a certain precision. The Kolmogorov entropy tells us how the precision of our prediction for the $n^{th}$ iterates $X_n$ decreases with *n*, due to the “sensitive dependence on initial conditions”. A positive value of *K* may be used to define the existence of chaos.

Numerically, the Kolmogorov entropy can be estimated as the second order Renyi entropy ( $K_2$) and can be related to the correlation integral of the reconstructed attractor as [Kantz.H and Schreiber.T, 2003]:

$$C_d(R) \sim \lim_{\substack{R \to 0 \\ d \to \infty}} R^D \exp(-\tau d K)$$

Where D is the fractal dimension of the system’s attractor, d is the embedding dimension and τ is the time delay used for attractor reconstruction.

This leads to the relation, $K \sim \frac{1}{\tau} \lim_{\substack{R \to 0 \\ d \to \infty}} \ln \frac{C_d(R)}{C_{d+1}(R)}$

### 4.4.3 Simulation Experiments and Results

In this work we extracted box-counting dimension for Malayalam vowels അ/Λ/, ഇ/I/, എ/ae/, ഒ/o/, and ഉ/u/. The acoustic data from each phoneme is embedded into a reconstructed phase space using time delay embedding with a delay of one sample and dimension two. Determination of more scientific time delay and dimension are described in the next chapter.

Figures 4.5(a) to 4.5(e) show the log-log plot of N(ε) vs. 1/ε for vowels അ/Λ/, ഇ/I/, എ/ae/, ഒ/o/, and ഉ/u/ with different speech samples. Box counting dimension $d_B$ is calculated as the slope of log N versus log(1/ ε) curve. Calculated value of $d_B$ is tabulated in table 4.1.

Using the speech data the correlation integral $C_d(R)$ for several ‘R’ with respect to each particular dimension ‘d’ is computed. Fig.4.6(a) shows a log-log plot of $C_d(R)$ vs R for vowel അ/Λ/ and Fig. 4.6(b) that for vowel ഇ/I/. It may be noted that with the increase in dimensions the slope at the linear portions of these curves converges to a limiting value. Fig.4.7 shows the slope ‘m’ in the linear range of the different curves vs the dimension ‘d’. The limiting value of the slope corresponds to the correlation dimension $d_C$ of the attractor. Table 4.2 shows the calculated value of $d_C$ for different vowels

Figures 4.8(a-b) represent the plot showing mean value of $\log[C_d(R) / C_{d+1}(R)]$ obtained from the linear range of the curves in Fig.4.6(a-b) for vowels അ/Λ/ and ഇ/I/. The Kolmogorov entropy K is obtained as the product of the limiting value of $\log[C_d(r) / C_{d+1}(r)]$ and ( $1/\tau$ ) and is tabulated in table 4.3.

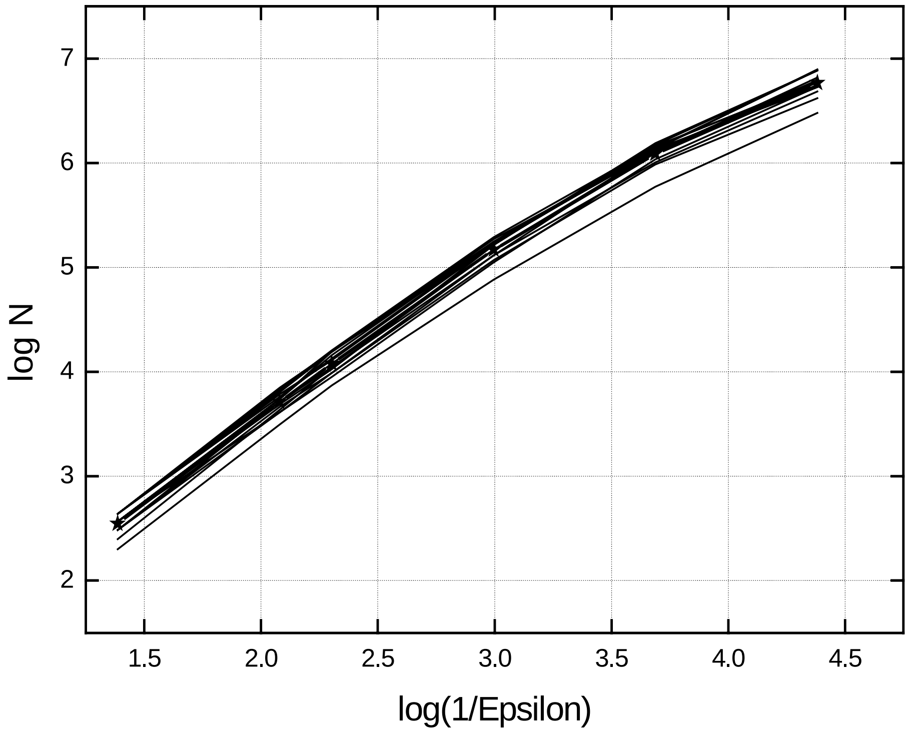


**Fig.4.5(a)**.log N versus log(1/ ε) plot for vowel അ/Λ/

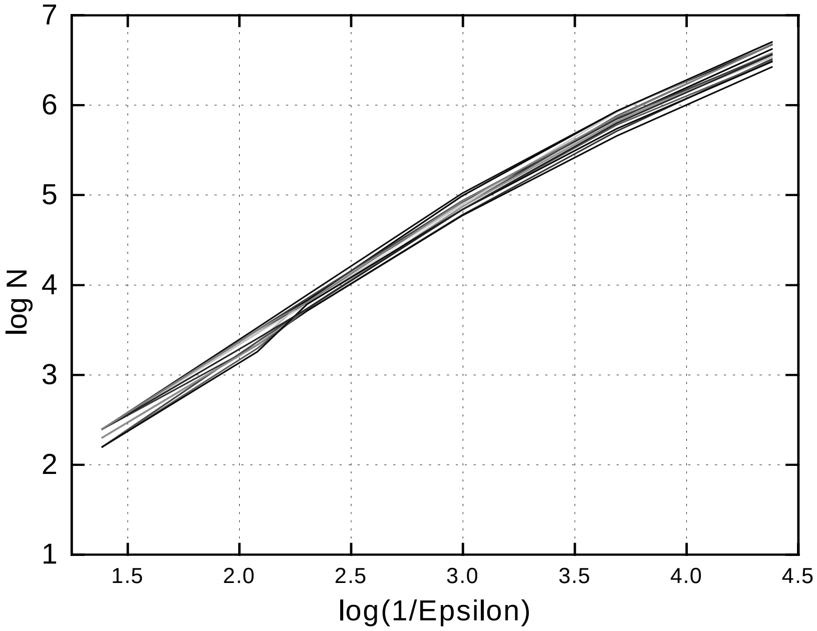


**Fig.4.5(b)**.log N versus log(1/ ε) plot for vowel ഇ/I/

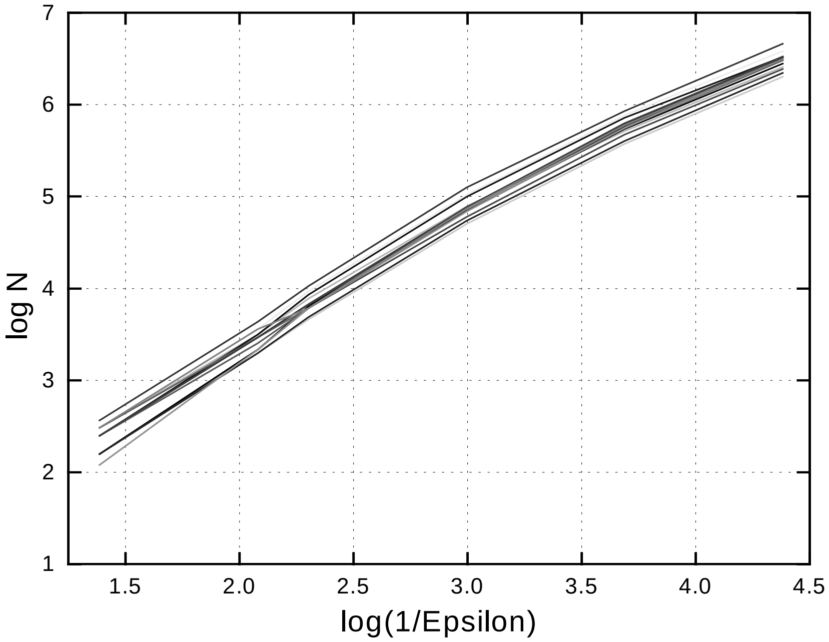


**Fig.4.5(c)**.log N versus log(1/ ε) plot for vowel ഐ/ae/

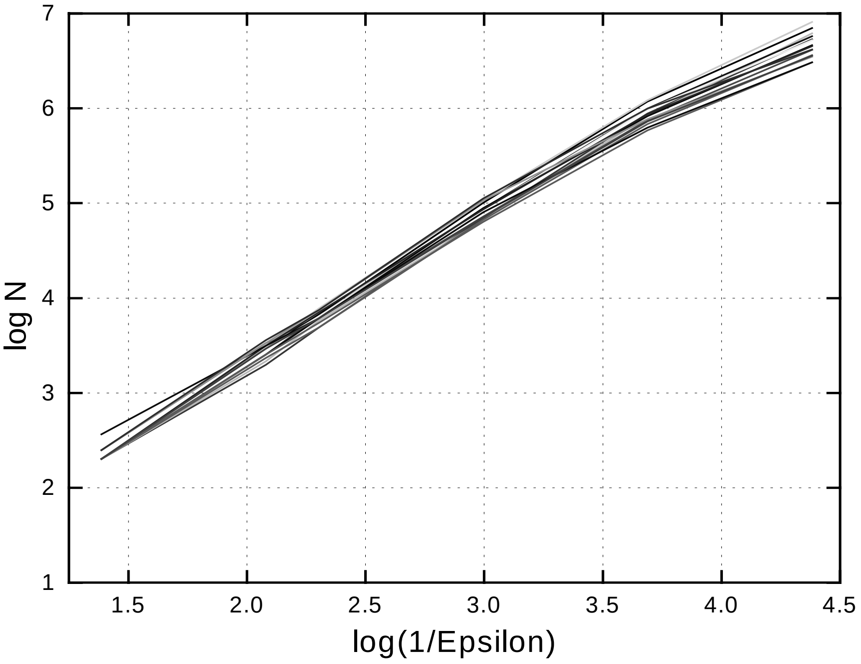


**Fig.4.5(d)**.log N versus log(1/ ε) plot for vowel ഒ/o/

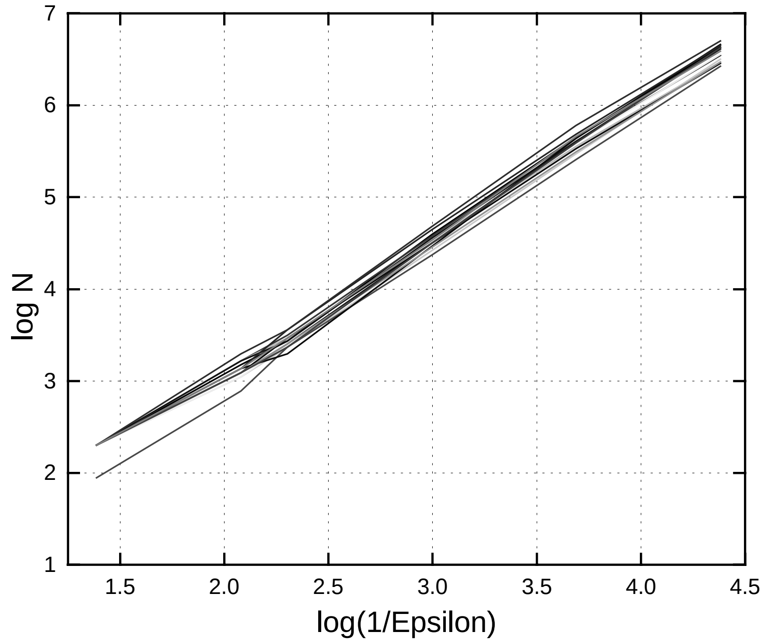


**Fig.4.5(e)**.log N versus log(1/ ε) plot for vowel ഉ/u/

| Vowel | Dimension $d_B$ |
|---|---|
| അ/Λ/ | 1.4208 |
| ഇ/I/ | 1.4371 |
| എ/ae/ | 1.3850 |
| ഒ/o/ | 1.4653 |
| ഉ/u/ | 1.4630 |

**Table 4.1.** Box counting dimension for Malayalam vowels

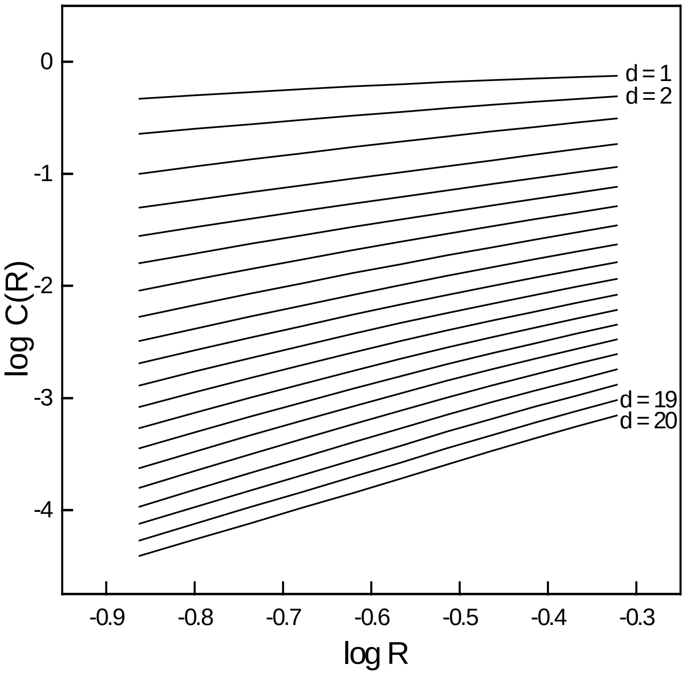


**Fig.4.6(a)**. Log – log plot of C(R) versus the distance R for vowel അ/Λ/

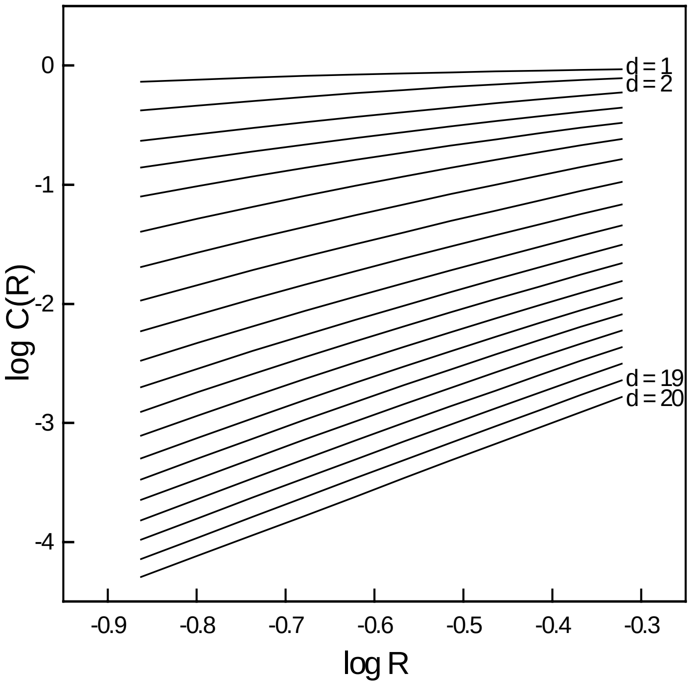


**Fig.4.6(b)**. Log – log plot of C(R) versus the distance R for vowel ഇ/I/

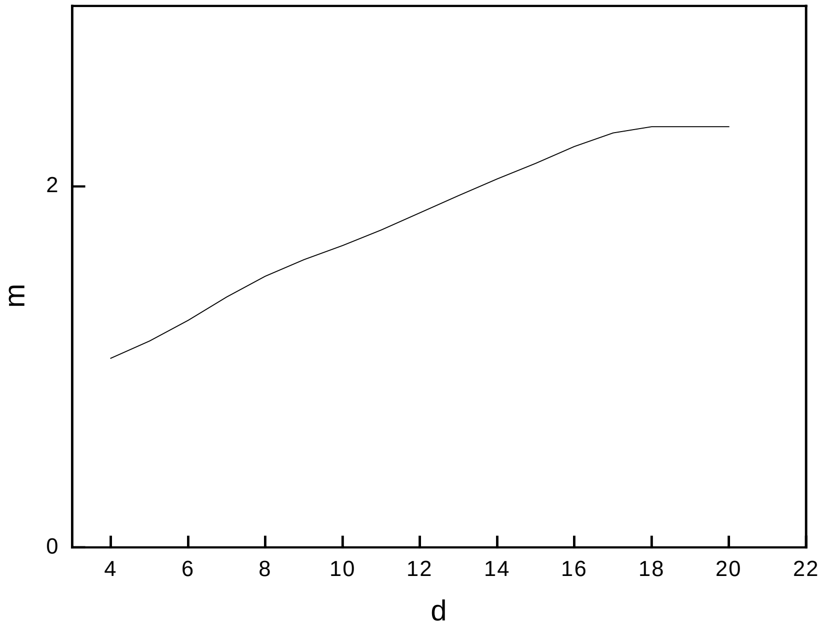


**Fig.4.7(a)**. The slope 'm' of different curves in Fig.4.6(a) versus the dimension 'd'

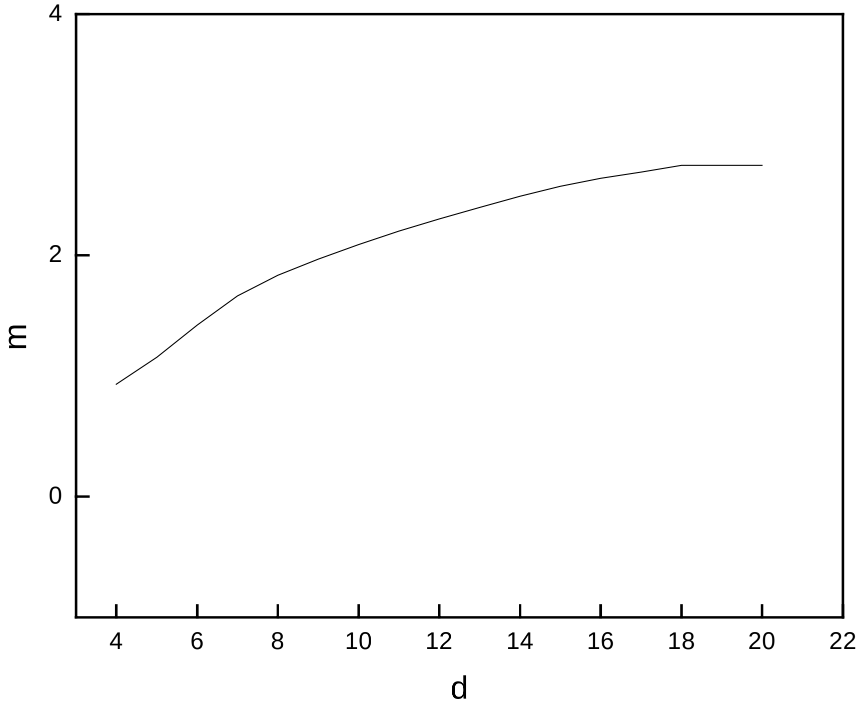


**Fig.4.7(b)**. The slope 'm' of different curves in Fig.4.6(b) versus the dimension 'd'

| Vowel | Dimension $d_C$ |
|---|---|
| അ/Λ/ | 2.33 |
| ഇ/I/ | 2.79 |
| എ/ae/ | 2.84 |
| ഒ/o/ | 2.80 |
| ഉ/u/ | 2.69 |

**Table 4.2.** Correlation dimension for Malayalam vowels

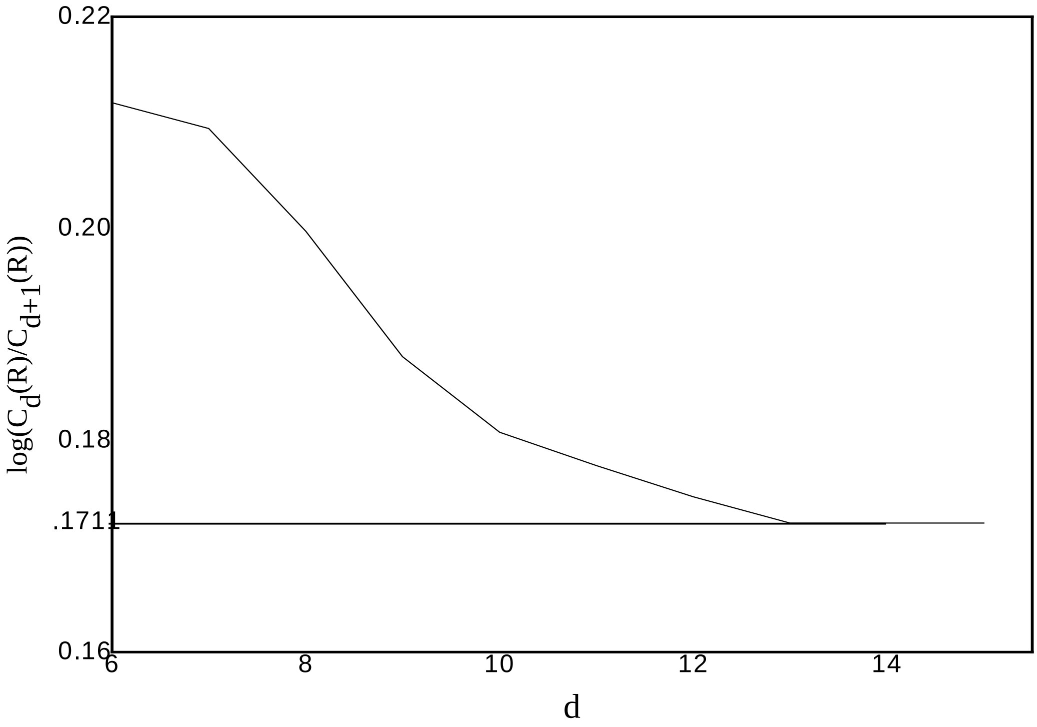


**Fig.4.8(a).**Mean value of log[$C_d$(R)/$C_{d+1}$(R) ] as a function of d for vowel അ/Λ/

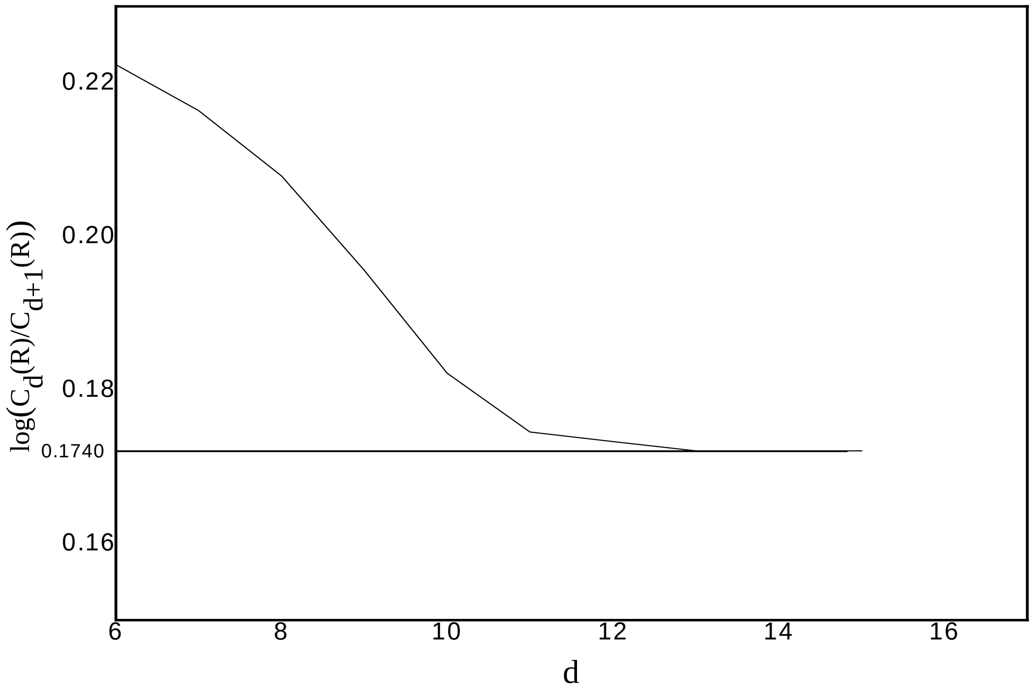


**Fig.4.8(b).** Mean value of log [$C_d$(R)/$C_{d+1}$(R) ] as a function of d for vowel ഇ/I/

| Vowel | Kolmogorov Entropy (K) |
|---|---|
| അ/Λ/ | 1367 |
| ഇ/I/ | 1392 |
| എ/ae/ | 1217 |
| ഒ/o/ | 1650 |
| ഉ/u/ | 1540 |

**Table 4.3.** Kolmogorov Entropy for Malayalam vowels

### 4.5 Conclusion

This chapter has introduced the field of nonlinear dynamical theory, including the concepts of chaos and how to measure it. The traditional model of speech production has been shown to have a number of short–comings and a nonlinear system has been proposed as an alternative. The problem of whether speech (especially vowel sounds) is chaotic has been examined through discussion of previous studies and experiments. Nonlinear invariant parameters for Malayalam vowels are calculated. The non-integer attractor dimension and non-zero value of Kolmogorov entropy confirm the contribution of deterministic chaos to the behavior of speech signal. Though these parameters quantify the chaotic behaviour of the speech signal, as far as recognition application is concerned, we want to go for more robust and computationally simple parameters. Phase space is a tool to analyze the underlying dynamics of a system. It can be exploited as a powerful signal processing domain, especially when the dynamical system of interest is nonlinear or chaotic. In the subsequent chapters we are focusing to extract a novel parameter from this time domain tool in order to capture the nonlinear characteristics of the speech.

# Chapter 5

# Phase Space Features for Speech Modeling

## 5.1 Introduction

The Phase Space of a dynamical system is a mathematical space with orthogonal co-ordinate directions representing each of the variables needed to specify the instantaneous state of the system. In Mathematics and Physics, Phase Space is the space in which all possible states of a system are represented, with each possible state of the system corresponding to one unique point in the phase space. For mechanical systems, the phase space usually consists of all possible values of position and momentum variables. A plot of position and momentum variables as a function of time is called a phase diagram. The number of state variables determines the dimension of the system. These systems are typically treated as deterministic, i.e. if the state of the system at time $t_0$ is known, then the state of the system at any time $t_1$ is completely predictable.

Unfortunately, the entire state space of almost all real systems cannot be observed, if only one state variable is available. It would seem that accurate characterization of the system is impossible in this case, especially if the dimensionality and nonlinearity of the system are high. However, with the use of a transformation on the observable variable known as a time-delay embedding, more information about the system is available than one might expect. Takens' theorem states that under certain assumptions, phase space of a dynamical system can be reconstructed through the use of time-delayed versions of the original scalar measurements [Takens.F, 1980]. This new state space is commonly referred to in the literature as a reconstructed phase space (RPS), and has been proven to be topologically equivalent to the original phase space of the dynamical system, as if all the state variables of that system would have been measured simultaneously [Kubin.G, 1995]. A Reconstructed Phase Space can be exploited as a powerful signal processing domain, especially when the dynamical system of interest is nonlinear or even chaotic [Broomhead.D.S and King.G, 1986], [Kantz.H and Schreiber.T,2003].

As explained in the earlier chapter, RPS can be used to estimate the dynamical invariants like attractor dimension, Kolmogorov Entropy etc. of the system [Abarbanel.H.D.I, 1996]. Recently this approach has been taken in the realm of speech signal processing by many researchers [Narayanan.N.K and Sridhar.C.S, 1988], [Narayanan.N.K, 1999], [Lindgren.A.C, Johnson.M.T, *et.al.*, 2003]. In this thesis, we use the distribution or density of the RPS as a basis for modeling the speech phonemes. To this end we introduce a new Phase Space based parameter called Reconstructed Phase Space Distribution Parameter (RPSDP), which can be effectively used for the recognition applications.

This chapter is organized as follows. In the first session a general introduction and properties of the Phase Space of a dynamical system is presented. Following this the reconstruction of Phase Space from a scalar time series is demonstrated based on Taken’s embedding theorem. Then application of Reconstructed Phase Space (RPS) in the area of signal processing and automatic speech recognition is explained. Finally the parameter extraction (Reconstructed Phase Space Distribution Parameter) from the RPS of Speech signal is presented.

## 5.2 Phase Space of a Dynamical System

For a purely deterministic system, once its present state is fixed, the states at all future times are determined as well. Thus it will be important to establish a vector space called Phase Space or State Space for the system, such that specifying a point in this space specifies the state of the system and vice versa. Then the information about the dynamics of a system can be obtained by studying the various features of the corresponding phase space distribution. This concept is illustrated in detail in the following session.

The state of a particle moving in one dimension is specified by its position (x) and velocity (v). Its phase space is a plane. On the other hand a particle moving in three dimensions would have a six dimensional phase space with three position and three velocity directions. In phase space, momenta can be used instead of velocities.

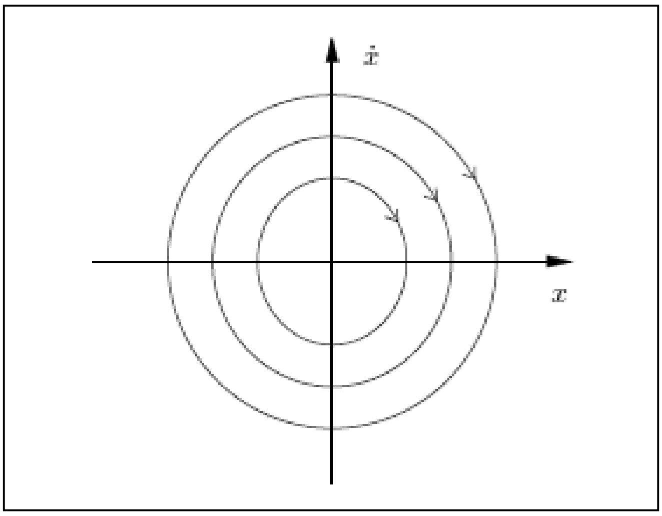


**Fig.5.1:** Phase trajectory of a conservative system

Figure 5.1 is a two dimensional phase diagram for an energy conserving system. The energy increases with the square of the radius of the trajectory. An important feature of the trajectory is that two trajectories corresponding to similar energies will pass very close to each other, but the orbits will not cross each other. This *non crossing* property derives from the fact that past and future states of a deterministic mechanical system are uniquely prescribed by the system state at a given time. A crossing of trajectories at time t would introduce ambiguity into past and future states, thereby rendering the system indeterminate.

Another important feature of the phase space of conservative (constant energy) systems is the preservation of areas. For a dissipative system, area of phase trajectory will not be a constant. This property leads to a classification of dynamical systems into two categories - conservative or dissipative, depending upon whether the phase volumes stay constant or contract, respectively. Figure 5.2 is the phase trajectory of a dissipative system.

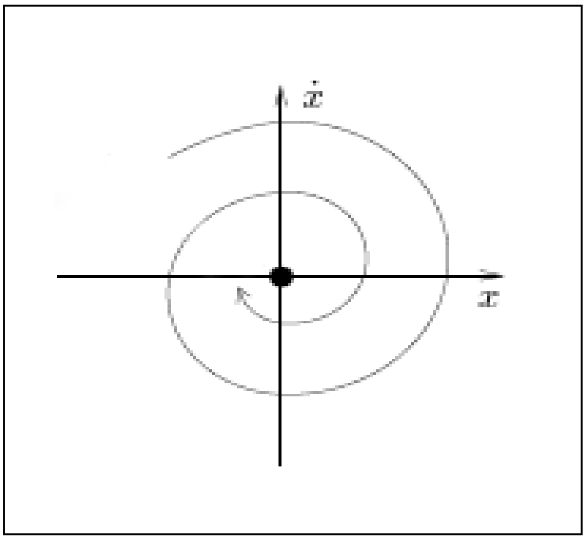


**Fig.5.2:** Phase trajectory of a dissipative system

### 5.3 Reconstructed Phase Space Using Time delay Embedding

Having stressed the importance of Phase Space for the study of dynamical systems, we have to face the first problem: in the case of a speech signal, what we have is not a Phase Space object but a time series, only a sequence of scalar measurements. We therefore have to convert the observations into state vectors. This is the important problem of Phase Space Reconstruction, which is technically solved by the method of time delay embedding.

One of the profound results established in chaos theory is the celebrated Takens' embedding theorem. Takens' theorem states that under certain assumptions, phase space of a dynamical system can be reconstructed through the use of time-delayed versions of the original scalar measurements.

This new state space is commonly referred to in the literature as Reconstructed Phase Space (RPS), and has been proven to be topologically equivalent to the original phase space of the dynamical system.

Packard et al. [Packard.N.H, Crutchfield.J.P, *et.al*, 1980] first proposed the concept of phase space reconstruction in 1980. Soon after, Takens showed that a delay-coordinate mapping from a generic state space to a space of higher dimension preserves topology [Takens.F, 1980].

Let M be a compact manifold of topological dimension d. Then for pairs (P,y) where $P : M \rightarrow M$ and $y : M \rightarrow R$, then the map

$$\Phi_{P,y}(x) = (y(x), y(P(x), \ldots\ldots\ldots\ldots y(P^{m-1}(x)))$$

This map is an embedding in $R^m$ , provided that $m \geq 2d$. This result guarantees that for an observable y from a smooth mapping with Phase Space dimension d. There is one to one correspondence between the mapped points $(y_i, y_{i+1}, \ldots.. y_{i+m-1}\ )$ and the points lying on the original Phase Space for $m \geq 2d$. That is Taken proved that, for a given uncorrupted signal of infinite length, this transformation is topologically equivalent [Takens.F, 1980].

This theorem provides important theoretical justification for the use of RPS’s for system identification and pattern classification. Because the topology of the RPS is identical to the topology of the underlying system’s phase space, we can expect the shape and density of the RPS attractor to provide valuable information of the system that generates a signal.

According to Taken's embedding theorem a Reconstructed Phase Space can be produced for a measured state variable - $\mathbf{x_n}$, n = 1,2 ,3,4 ….N, via the method of delays by creating vectors given by

$$\mathbf{x_n} = [\ x_n \ \ x_{n+\tau}\ x_{n+2\tau}\ \dots\ x_{n+(d-1)\,\tau}\ ]$$

where d is the embedding dimension and $\tau$ is the chosen time delay value. The row vector, $\mathbf{x_n}$, defines the position of a single point in the RPS. The row vectors then can be compiled into a matrix called a trajectory matrix to completely define the dynamics of the system and create a Reconstructed Phase Space, as

$$\mathbf{X} \quad = \quad \begin{bmatrix} x_1 & x_{1+\tau} & x_{1+2\tau} & \dots\dots\dots\ x_{1+(d-1)\,\tau} \\ x_2 & x_{2+\tau} & x_{2+2\tau} & \dots\dots\dots\ x_{2+(d-1)\,\tau} \\ x_3 & x_{3+\tau} & x_{3+2\tau} & \dots\dots\dots\ x_{3+(d-1)\,\tau} \\ \bullet & & & \\ \bullet & & & \\ \bullet & & & \\ x_N & x_{N+\tau} & x_{N+2\tau} & \dots\dots\dots\ x_{N+(d-1)\tau} \end{bmatrix}$$

A speech signal with amplitude values can be treated as a dynamical system with a one dimensional time series data. Based on the above theory, this study investigates a method to model a Reconstructed Phase Space for Malayalam vowels, through the use of time delay versions of the original scalar measurements. Here trajectory matrices $\mathbf{X}_1$ with embedding dimension d = 2 and time delay $\tau$ = 1 and $\mathbf{X}_2$ with embedding dimension d = 3 and time delay $\tau = 1$ are constructed by considering the speech amplitude values $\mathbf{x_n}$ as one dimensional time series data. The matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ thus obtained are given below.

$$\mathbf{X_1} = \begin{bmatrix} x_1 & x_2 \\ x_2 & x_3 \\ x_3 & x_4 \\ \cdot & \\ \cdot & \\ \cdot & \\ x_N & x_{N+1} \end{bmatrix}$$

$$\mathbf{X_2} = \begin{bmatrix} x_1 & x_2 & x_3 \\ x_2 & x_3 & x_4 \\ x_3 & x_4 & x_5 \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ x_N & x_{N+1} & x_{N+2} \end{bmatrix}$$

By plotting the row vectors of the trajectory matrix, a visual representation of the system dynamics becomes evident as shown in the figures 5.3(a-e) and 5.4(a-e).

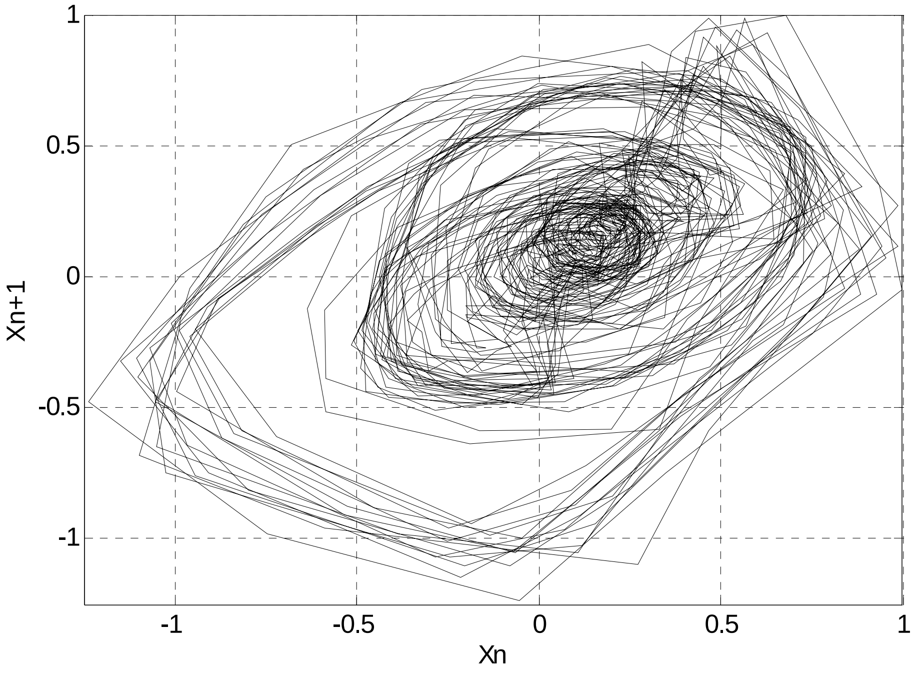


**Fig. 5.3(a)**:Two dimensional Reconstructed Phase Space for Malayalam Vowel അ/Λ/

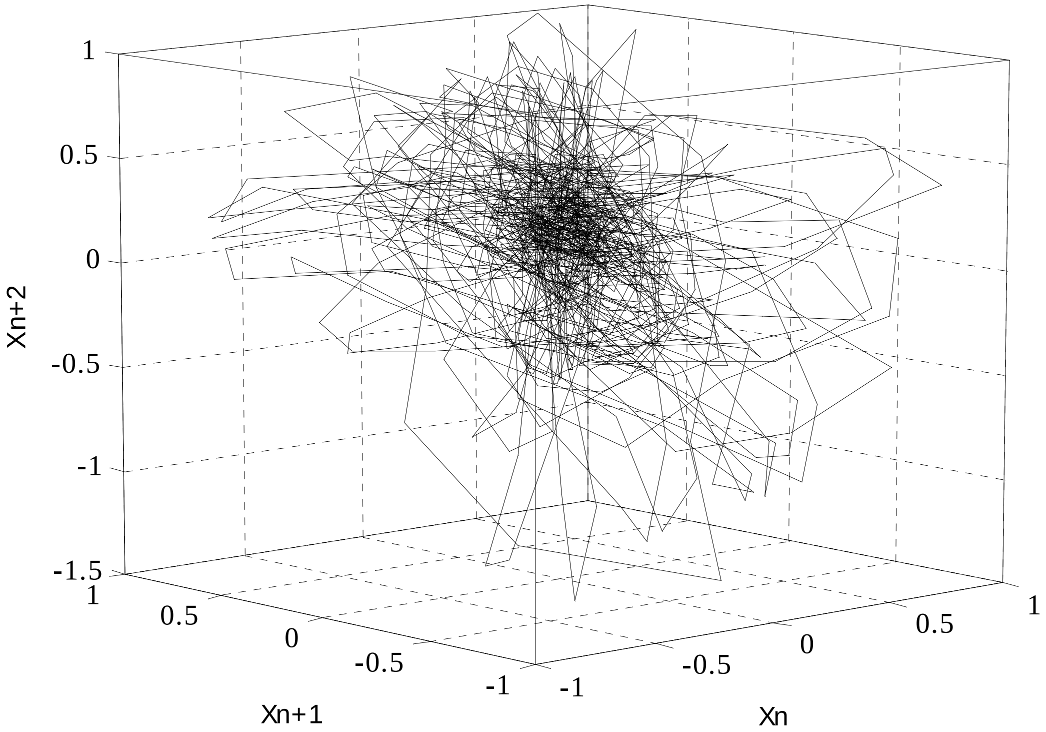


**Fig.5.4(a):** Three dimensional Reconstructed Phase Space for Malayalam Vowel അ/Λ/

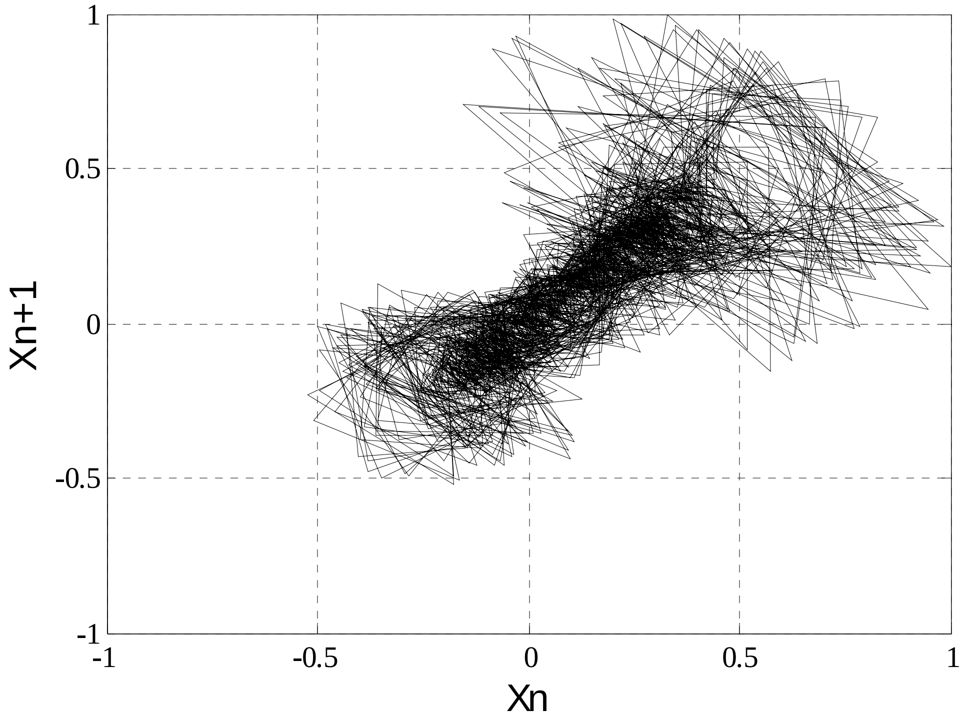


**Fig. 5.3(b)**: Two dimensional Reconstructed Phase Space for Malayalam Vowel **ഇ/I/**

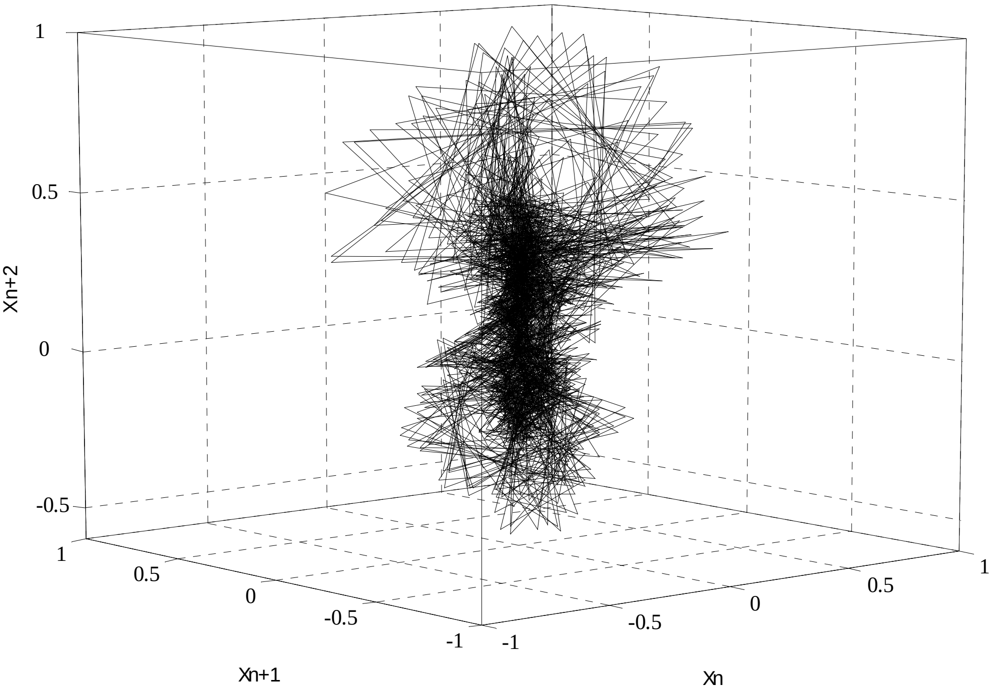


**Fig. 5.4(b):** Three dimensional Reconstructed Phase Space for Malayalam Vowel **ഇ/I/**

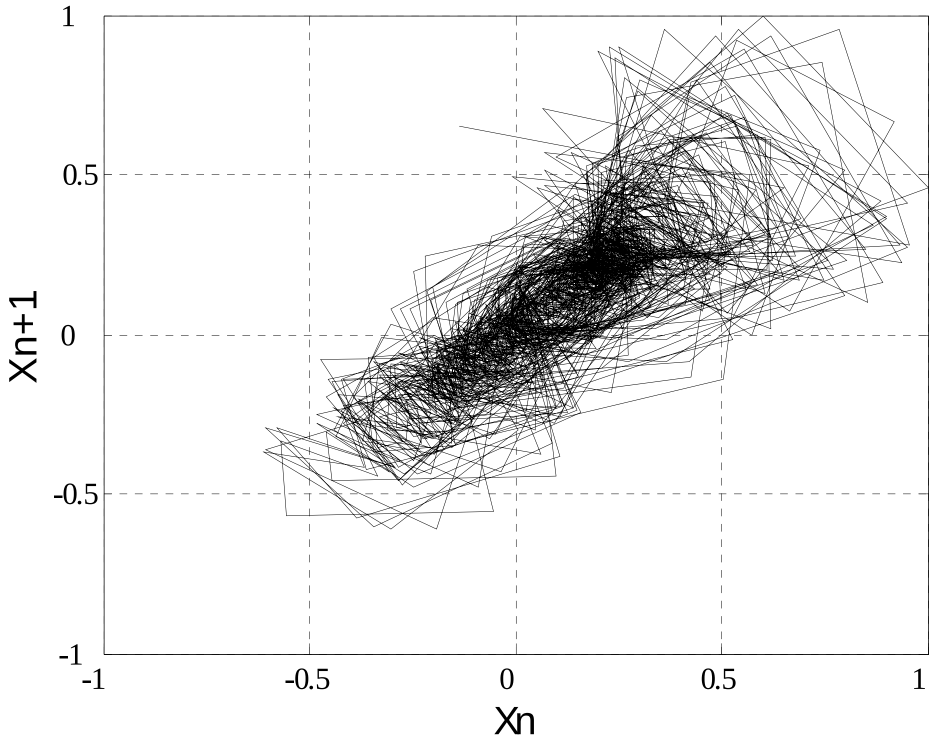


**Fig. 5.3(c):** Two dimensional Reconstructed Phase Space for Malayalam Vowel **ഏ/ae/**

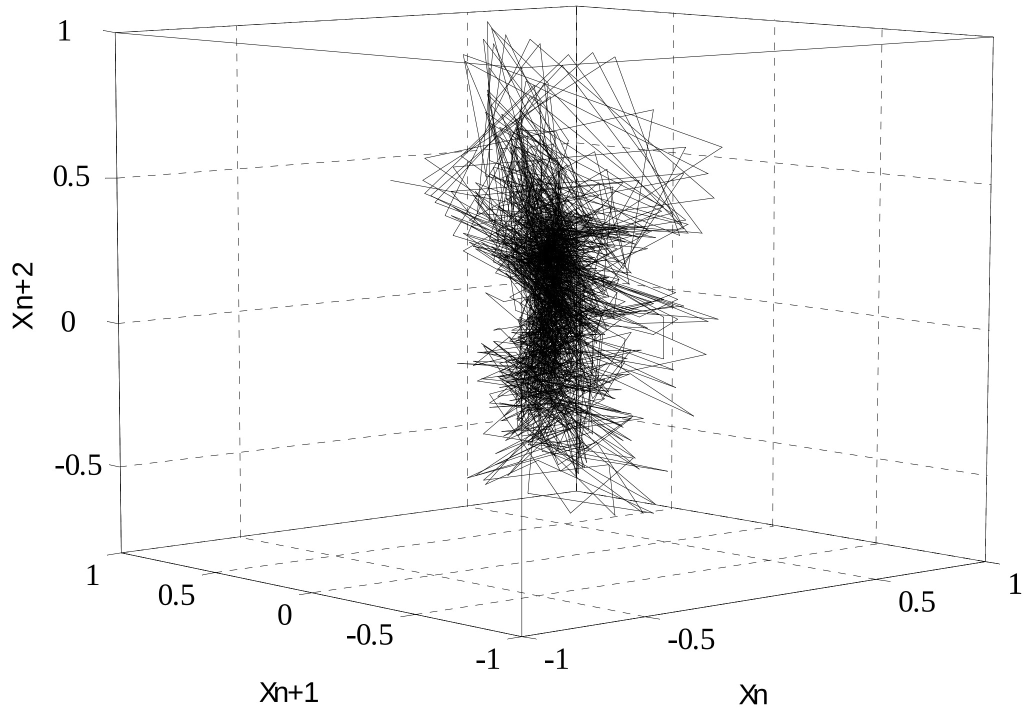


**Fig.5.4(c):** Three dimensional Reconstructed Phase Space for Malayalam Vowel **ഏ/ae/**

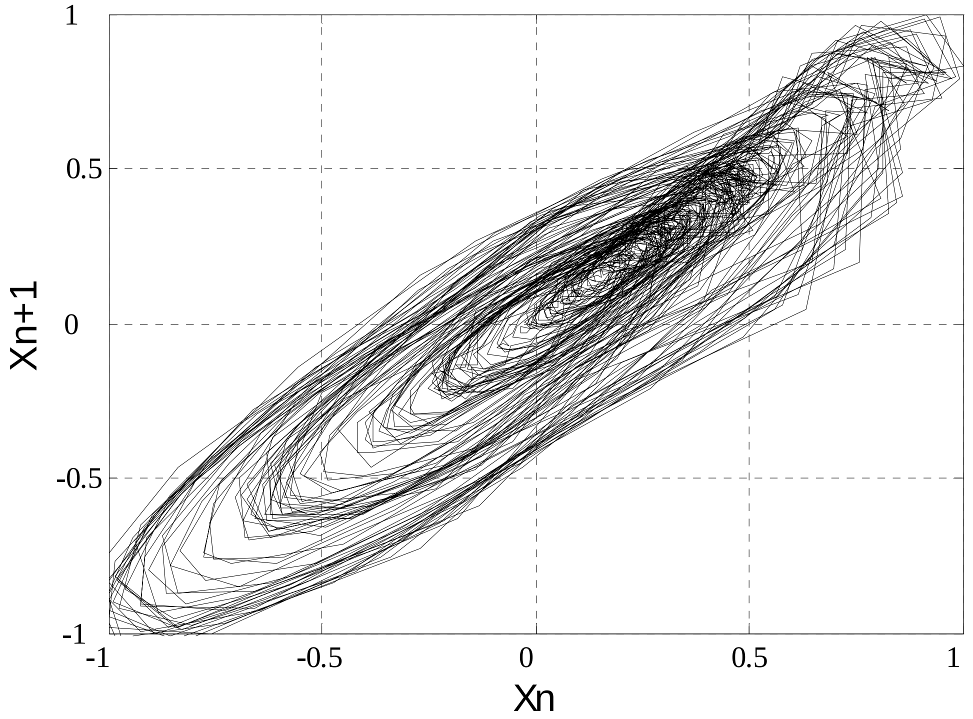


**Fig. 5.3(d):** Two dimensional Reconstructed Phase Space for Malayalam Vowel **ഒ/o/**

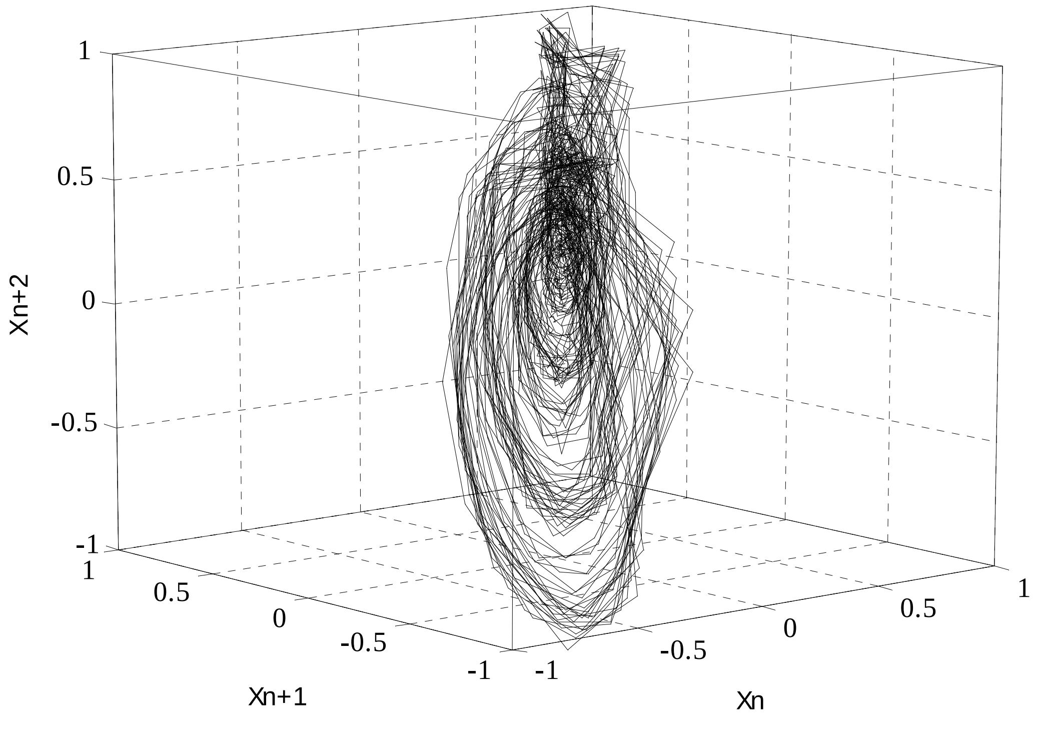


**Fig. 5.4(d):** Three dimensional Reconstructed Phase Space for Malayalam Vowel **ഒ/o/**

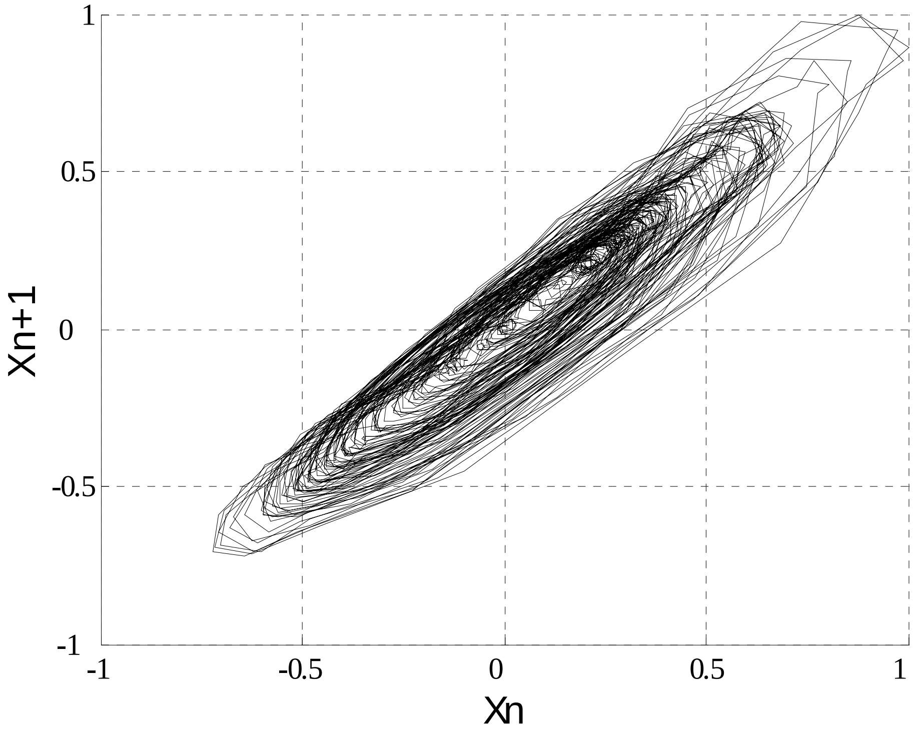


**Fig. 5.3(e):** Two dimensional Reconstructed Phase Space for Malayalam Vowel ഉ/**u**/

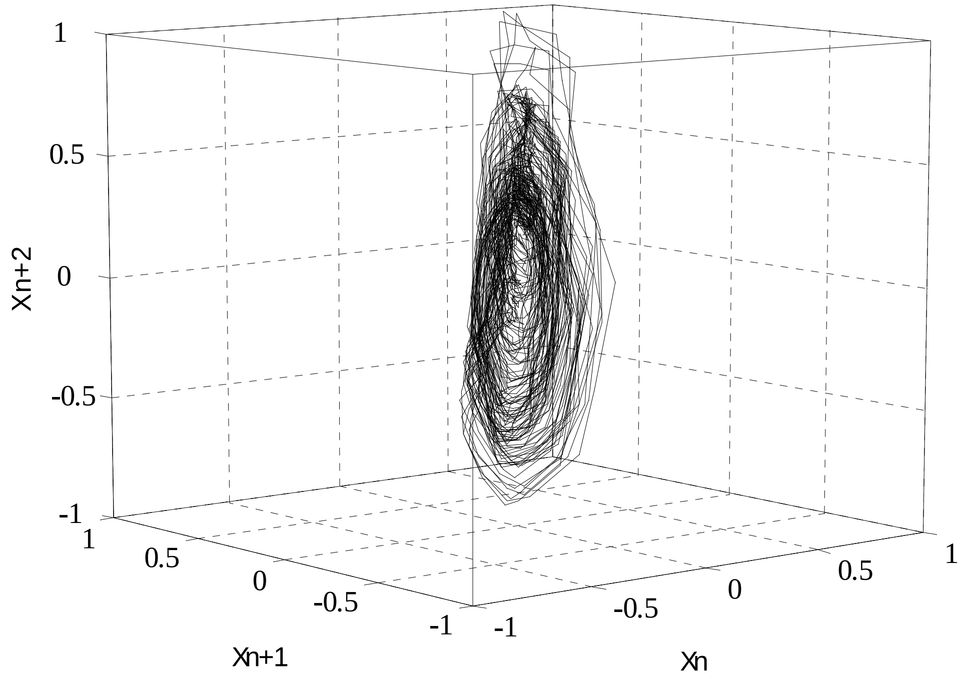


**Fig. 5.4(e):** Three dimensional Reconstructed Phase Space for Malayalam Vowel ഉ/**u**/

## 5.4 Application of RPS in Signal Processing

Because RPS's preserve the dynamical information in a system, they have advantages over other methods for signal classification. Many classical methods make assumptions about the systems in question, and in doing so, we ignore information that does not fit with those assumptions. While data reduction is often desirable for practical purposes, the reduction of data can remove important features. RPS's do not remove any information from the signals; the entire signal can be reconstructed from an RPS.

Unlike traditional linear signal classification methods, RPS's are able to capture nonlinear information that may be present in signals. Classic methods such as LPC analysis and cepstral analysis work by assuming a linear model and learning parameters to fit the linear model [Deller.J.R, Hansen.J.H.L, *et.al.*, 2000]. These approaches work very well if the signals to be analyzed are generated by linear systems, but are fundamentally flawed if that is not the case. These methods are popular for speech signal analysis, based on the linearity assumption.

Dynamical invariant features extracted from RPS's have been used for classification of speech signals. Lyapunov exponents [Narayanan.S.S and Alwan.A.A, 1995], [Kumar.A and Mullick.S.K, 1996] and correlation dimension [Pitsikalis.V and Maragos.P, 2003] have been used as features in addition to cepstral coefficients by appending them to the standard feature vectors. While this has been shown to increase accuracy, the improvements have been relatively small.

In addition to invariant features, RPS’s can be statistically modelled. This approach has been successfully applied to heart arrhythmia classification [Roberts.F.M, Povinelli.R.J, *et.al.*, 2001] as well as motor fault detection [Povinelli.R.J, Bangura.J.F, *et.al.*, 2002]. Instead of extracting features from the RPS’s, such as Lyapunov exponents and various dimensions, the full attractors are modelled. In this case, there is no data reduction, which is important, as the attractors may contain information not preserved by the dynamical invariants. To accomplish this, a statistical model is built over the attractor, describing the natural dimension of the RPS.

## 5.5 Application of RPS in ASR

Conventional speech signal processing techniques are based on linear systems theory, where the fundamental processing space is the frequency domain. Traditional acoustic approaches assume a source-filter model where the vocal tract is modeled as a linear filter. Although the features based on these approaches have demonstrated excellent performance over the years, they are, nevertheless, rooted in the strong linearity approximations of the underlying physics. Performance of the current speech recognition systems which rely on the linear system theory are far inferior to humans, and there are many factors that severely degrade recognition performance. As an alternative to the traditional techniques, interest has emerged in studying speech as a nonlinear system [Teager.H.M and Teager.S.M, 1990] [Banbrook.M and McLaughlin.S, 1994], [Banbrook.M and McLaughlin.S, 1999]. State of the art speech recognition systems typically use Cepstral coefficient features, obtained via a frame-based spectral analysis of the speech signal. However, recent  work in Phase Space Reconstruction Techniques for nonlinear modeling of time-series signals has motivated investigations into the efficacy of using dynamical systems models in the time-domain for speech recognition. In theory, reconstructed Phase Spaces capture the full dynamics of the underlying system, including nonlinear information not preserved by traditional spectral techniques, leading to possibilities for improved recognition accuracy. The potential benefit of using this information is still under investigation.

One disadvantage of using RPS for speech recognition is seen in the time complexity for classification of phonemes. Because cepstral coefficients are frame-based, a feature vector is computed and used for statistical analysis every 10 milliseconds, where as RPS have a feature vector at every sample. Due to the amount of data, time complexity is far greater for RPS based approach to speech recognition. For RPS based methods to become useful, this issue must be solved.  In this chapter we introduce a new method to extract a useful feature vector from the Reconstructed Phase Space, called RPS Distribution  Parameter (RPSDP)

## 5.6  Reconstructed Phase Space Distribution

In this session Reconstructed Phase Space Distribution plot (scatter graph) in two dimensions is constructed for Malayalam vowels as follows. The normalized 'N' sample values for the Malayalam vowel sounds is the scalar time series $\mathbf{x_n}$, where n = 1,2,…………N. For each speech signal, a trajectory matrix is formed with dimension d = 2 and time delay τ = 1. (More scientific choice for ‘d’ and ‘τ’ is explained in chapter 7). Now a scatter plot named Reconstructed Phase Space Distribution is generated by plotting the row vectors of the above constructed trajectory matrix. (ie by plotting $x_n$ versus $x_{n+1}$). Figures 5.5(a-e) shows the Reconstructed Phase Space Distribution plot for five Malayalam phonemes അ/Λ/, **ഇ/I/**, **എ/ae/**, **ഒ/o/**, and **ഉ/u/**.

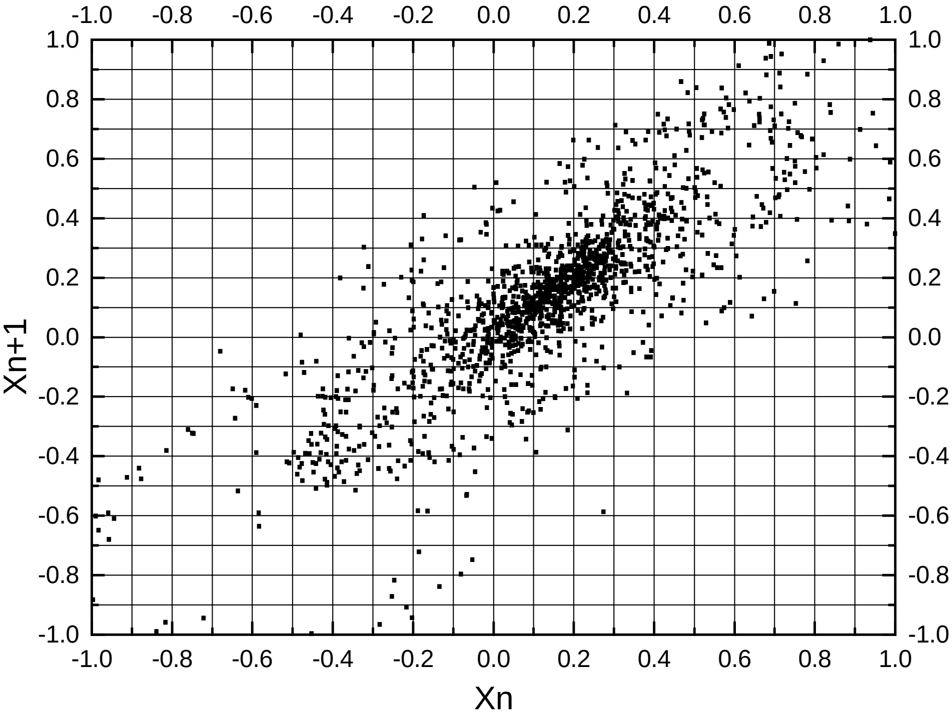


**Fig. 5.5(a):** Reconstructed Phase Space Distribution for Malayalam Vowel അ/Λ/

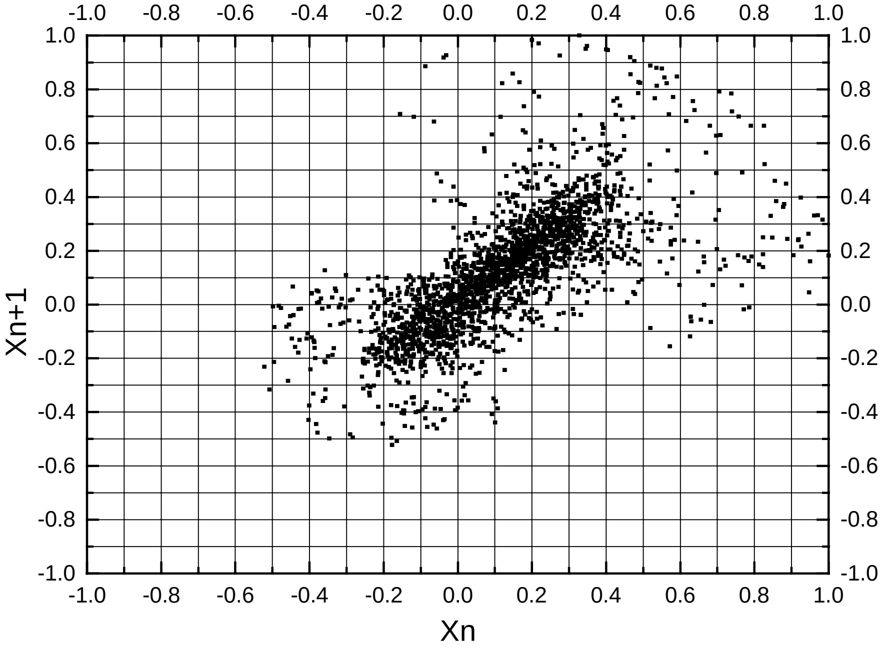


**Fig. 5.5(b)**: Reconstructed Phase Space Distribution for Malayalam Vowel ഇ/I/

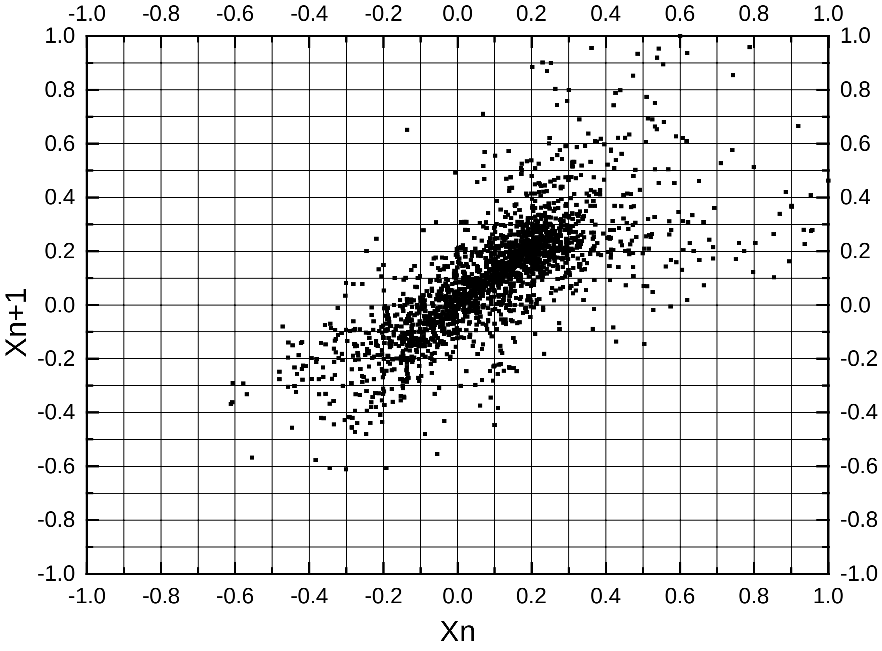


**Fig. 5.5(c)**: Reconstructed Phase Space Distribution for Malayalam Vowel ൡ/ae/

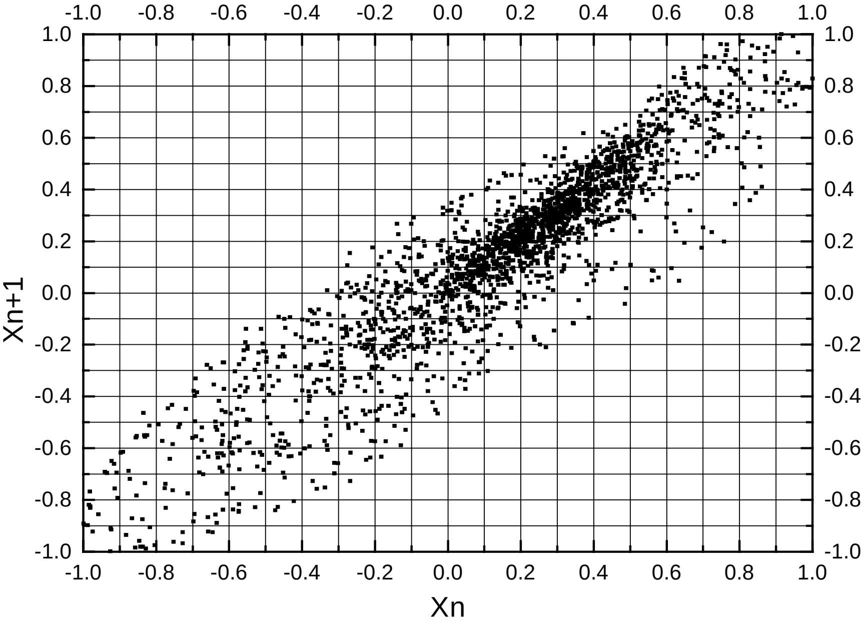


**Fig. 5.5(d)**: Reconstructed Phase Space Distribution for Malayalam Vowel **ഒ/o/**

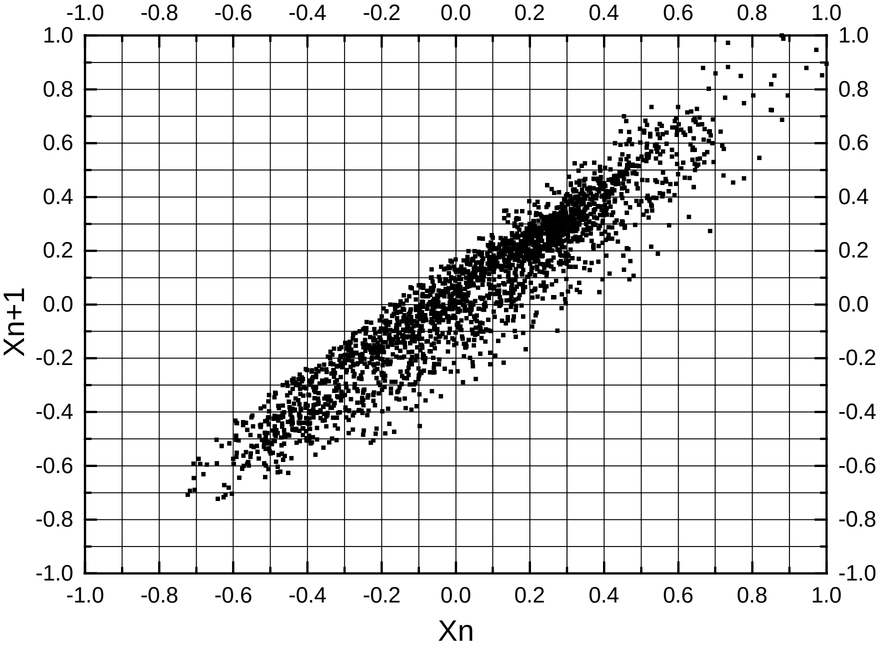


**Fig. 5.5(e)**: Reconstructed Phase Space Distribution for Malayalam Vowel **ഉ/u/**

## 5.7 Reconstructed Phase Space Distribution Parameter

The complexity of RPS based approach due to the excess amount of data can be reduced by extracting useful parameters from the Reconstructed Phase Space. Here RPS is divided into 400 locations, and the number of Phase Space Points distributed in each location is calculated as follows.

RPS is divided into grids with 20 x 20 boxes. The box defined by co-ordinates (-1, .9) , (-.9, 1) is taken as location 1. Box just right side to it is taken as location 2 and it is extended towards X direction, with the last box in the row (.9,.9),(1, 1) as location 20. This is repeated for the next row, taking the starting box as location 21 and repeated for all other rows. The Reconstructed Phase Space Distribution Parameter (RPSDP) is calculated by estimating the number of Phase Space points distributed in each location.

This can be mathematically represented as follows. The Reconstructed Phase Space Distribution Parameter for location 'i' in two dimension can be defined as :

$$(\mathrm{RPSDP})_i = \sum_{n=1}^{N} \eta(\,[x_n, x_{n+\tau}], i\,)$$

where $\eta(\,[x_n, x_{n+\tau}], i\,) = 1$, if the Phase Space point defined by the row vector $[x_n, x_{n+\tau}]$ is in the location 'i'.

$= 0$, otherwise.

Generally RPSDP for location ‘i’ in d dimension can be defined as :

$$(\mathrm{RPSDP})_i = \sum_{n=1}^{N} \eta([x_n, x_{n+\tau,} x_{n+2\tau} \ldots\ldots x_{n+(d-1)\tau}], i)$$

where $\eta([x_n, x_{n+\tau,} x_{n+2\tau} \ldots\ldots x_{n+(d-1)\tau}], i) = 1$, if the Phase Space point defined by the row vector $[x_n, x_{n+\tau,} x_{n+2\tau} \ldots x_{n+(d-1)\tau}]$ is in the location ‘i’.

= 0, otherwise.

Figures 5.6(a-e) show the Reconstructed Phase Space Distribution Parameter versus locations for the vowels അ/Λ/, **ഇ/I/**, **എ/ae/**, **ഒ/o/**, and **ഉ/u/**.

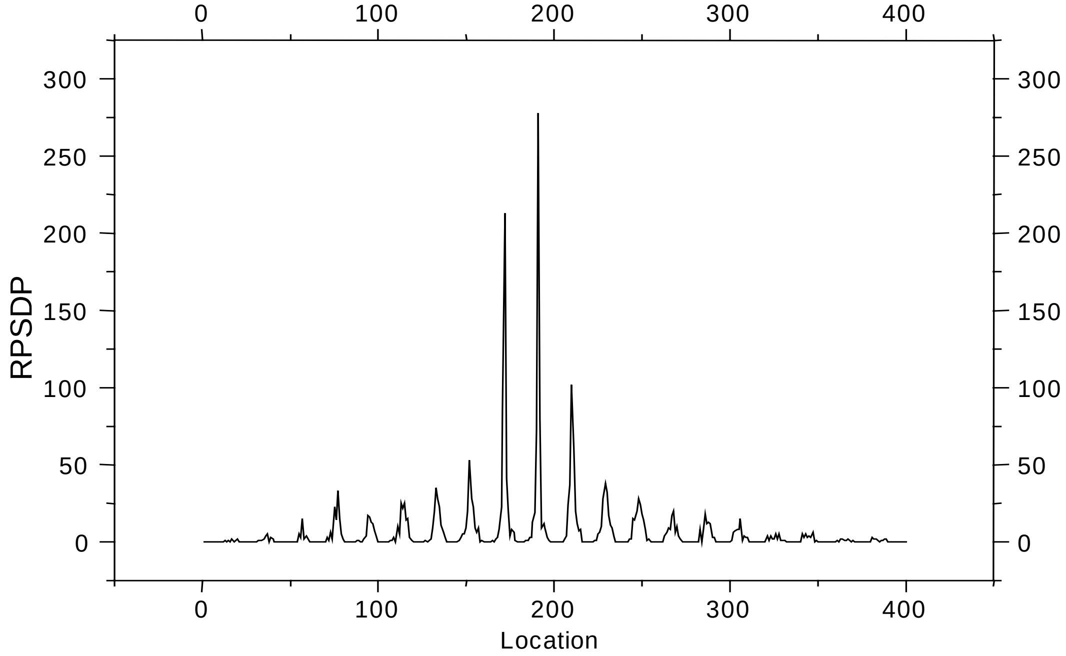


**Fig. 5.6(a)** :Reconstructed Phase space Distribution Parameter (Vowel അ/Λ/)

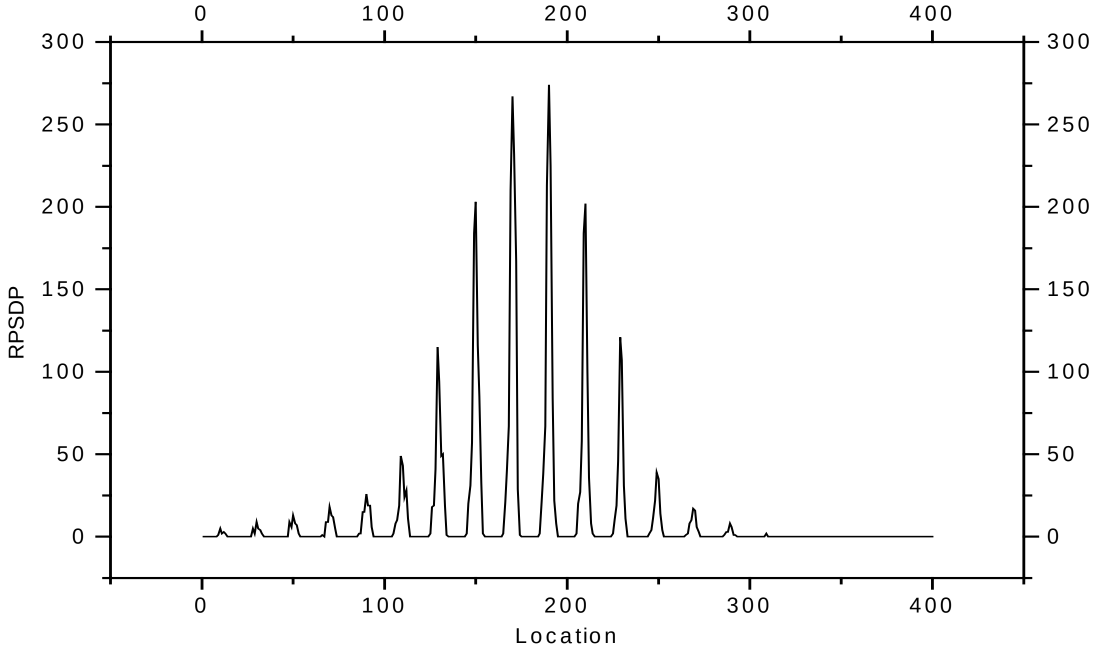


**Fig. 5.6(b)** :Reconstructed Phase space Distribution Parameter (Vowel **ഇ/I/**)

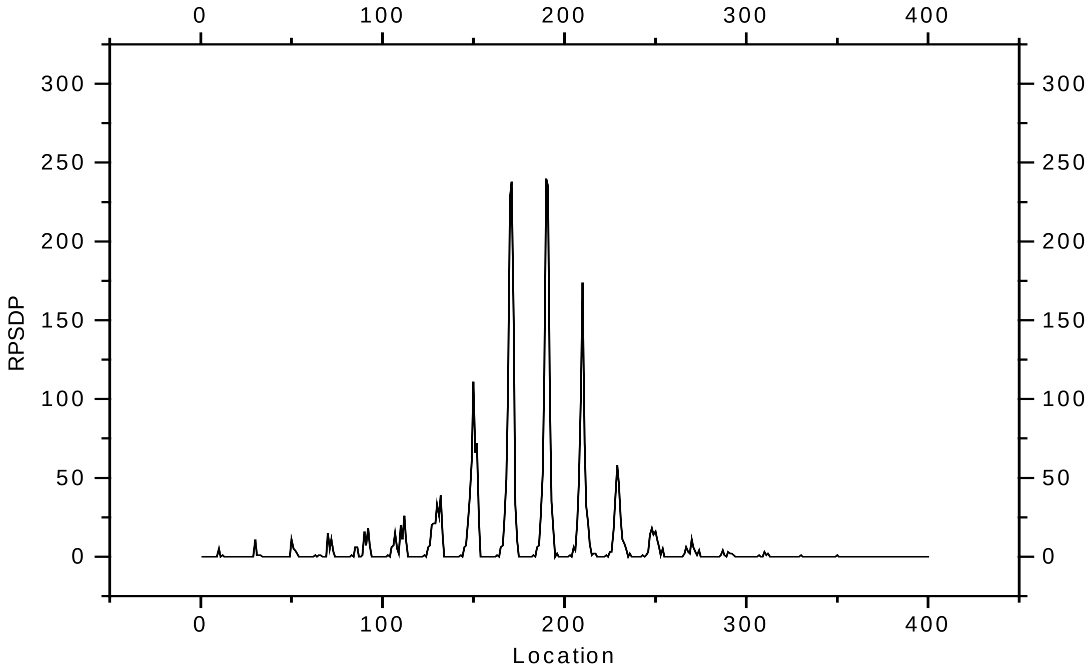


**Fig. 5.6(c):**Reconstructed Phase space Distribution Parameter (Vowel **എ/ae/**)

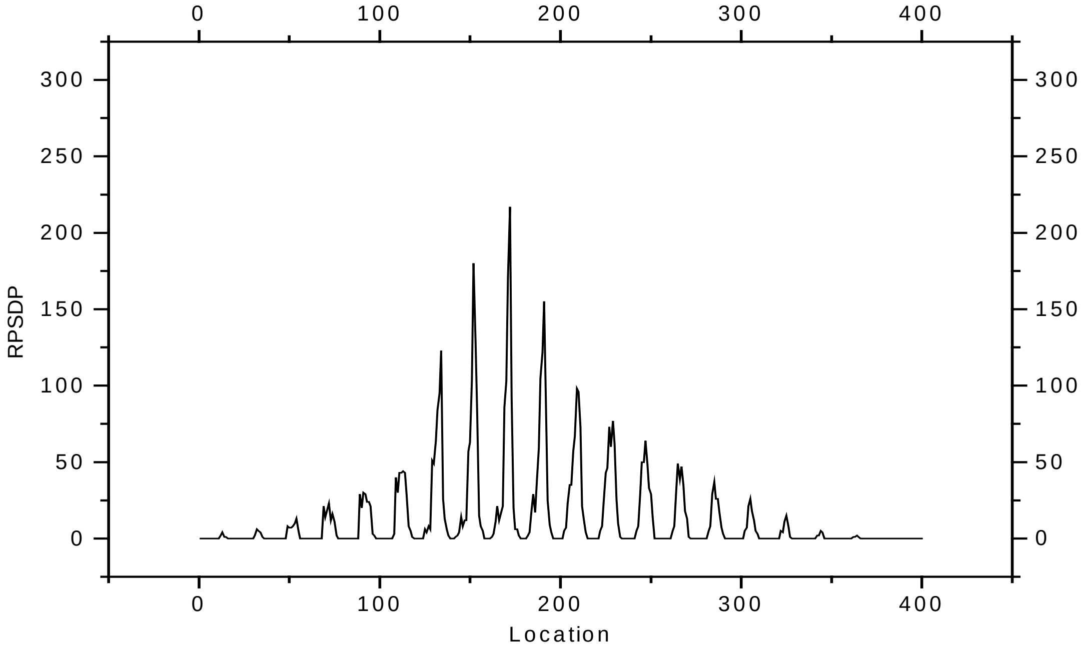


**Fig. 5.6(d):** Reconstructed Phase space Distribution Parameter (Vowel ഒ/**o**/)

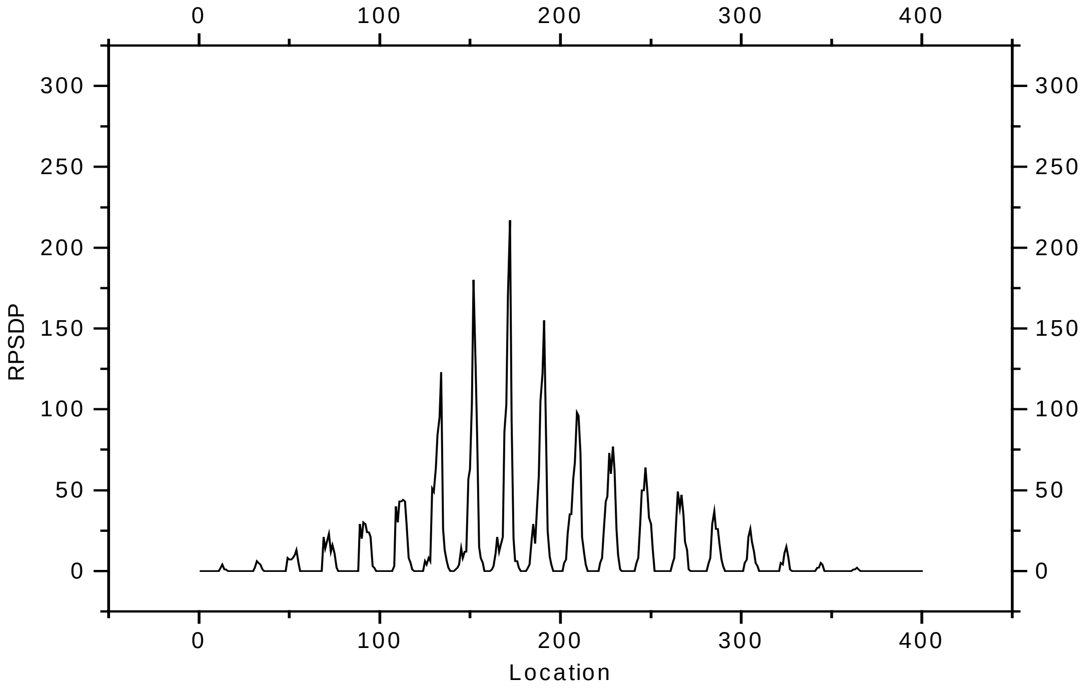


**Fig. 5.6(e):** Reconstructed Phase space Distribution Parameter (Vowel ഉ/**u**/)

This operation is repeated for the same vowel uttered at different occasions. Figures 5.7(a-e) show the Reconstructed phase space point distribution graphs for each vowel uttered at different occasions. The graph thus plotted for different vowels shows the identity for a vowel as regard to pattern. Therefore this technique can be effectively utilized for speech recognition applications

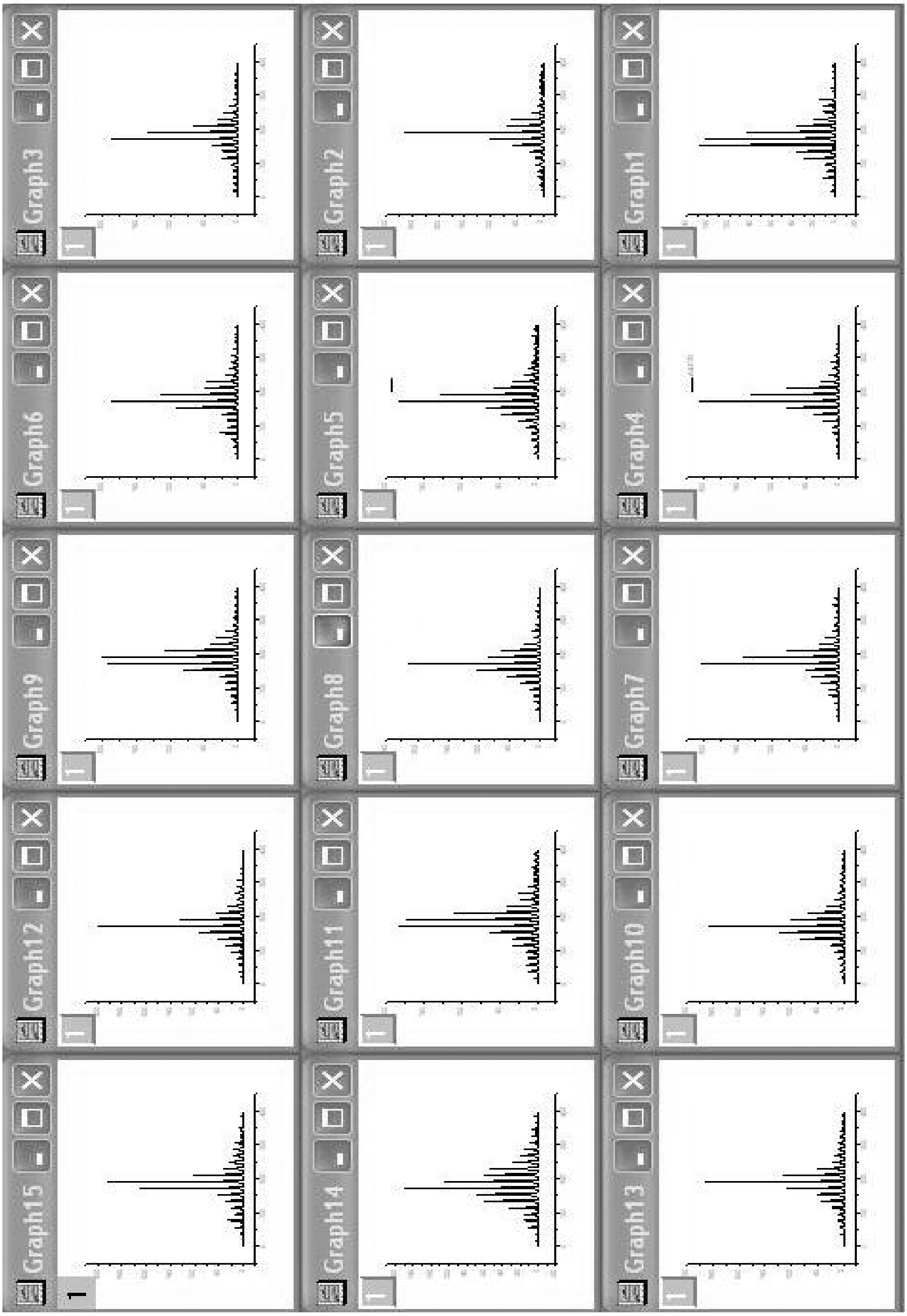


**Fig. 5.7(a):**Reconstructed Phase Space Distribution Parameter for 15 repeated utterances (Vowel അ / Λ/)

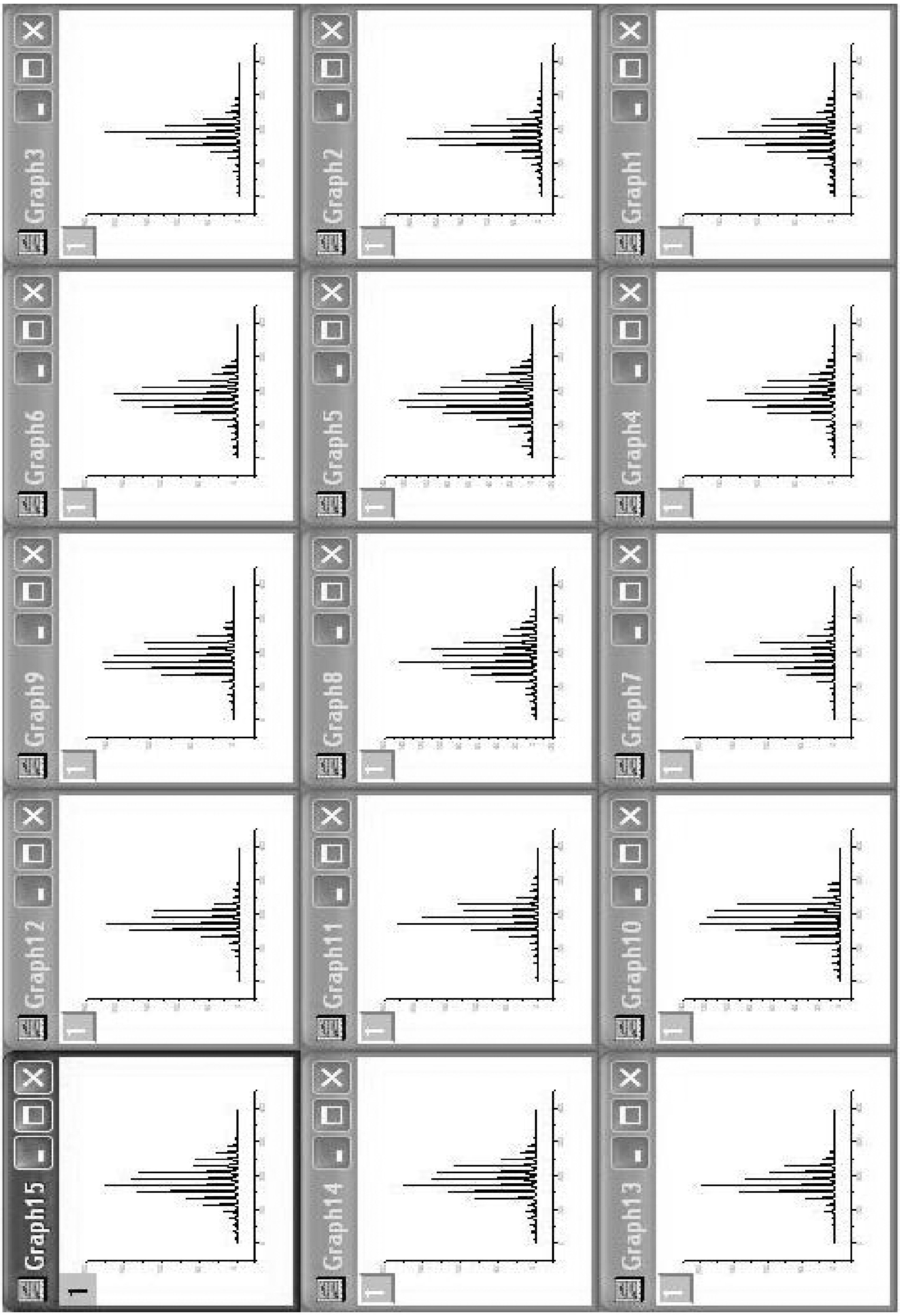


**Fig. 5.7(b):** Reconstructed Phase Space Distribution Parameter for 15 repeated utterances (Vowel ഇ / **I** /  )

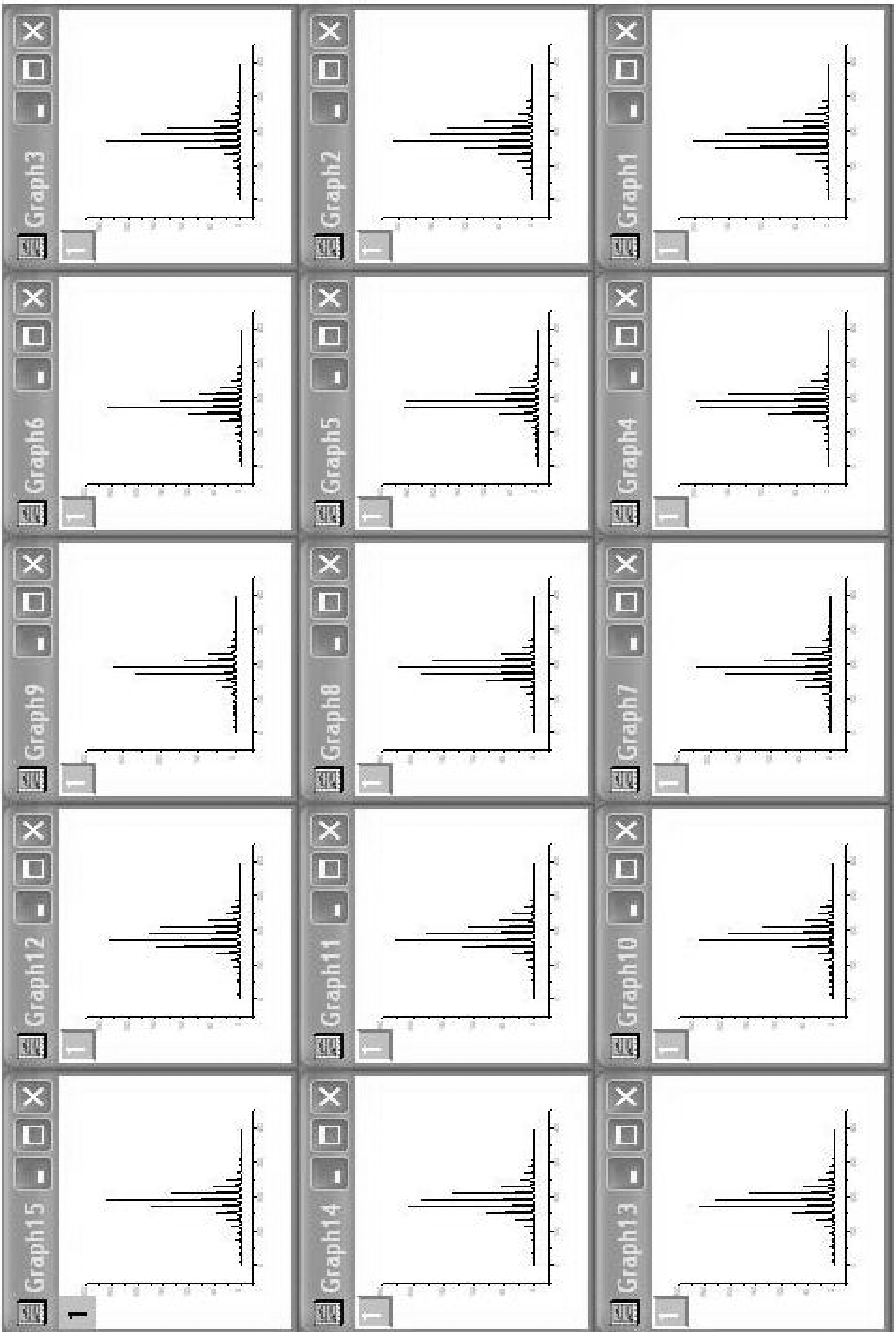


**Fig. 5.7(c):** Reconstructed Phase Space Distribution Parameter for 15 repeated utterances (Vowel ഐ /**ae**/ )

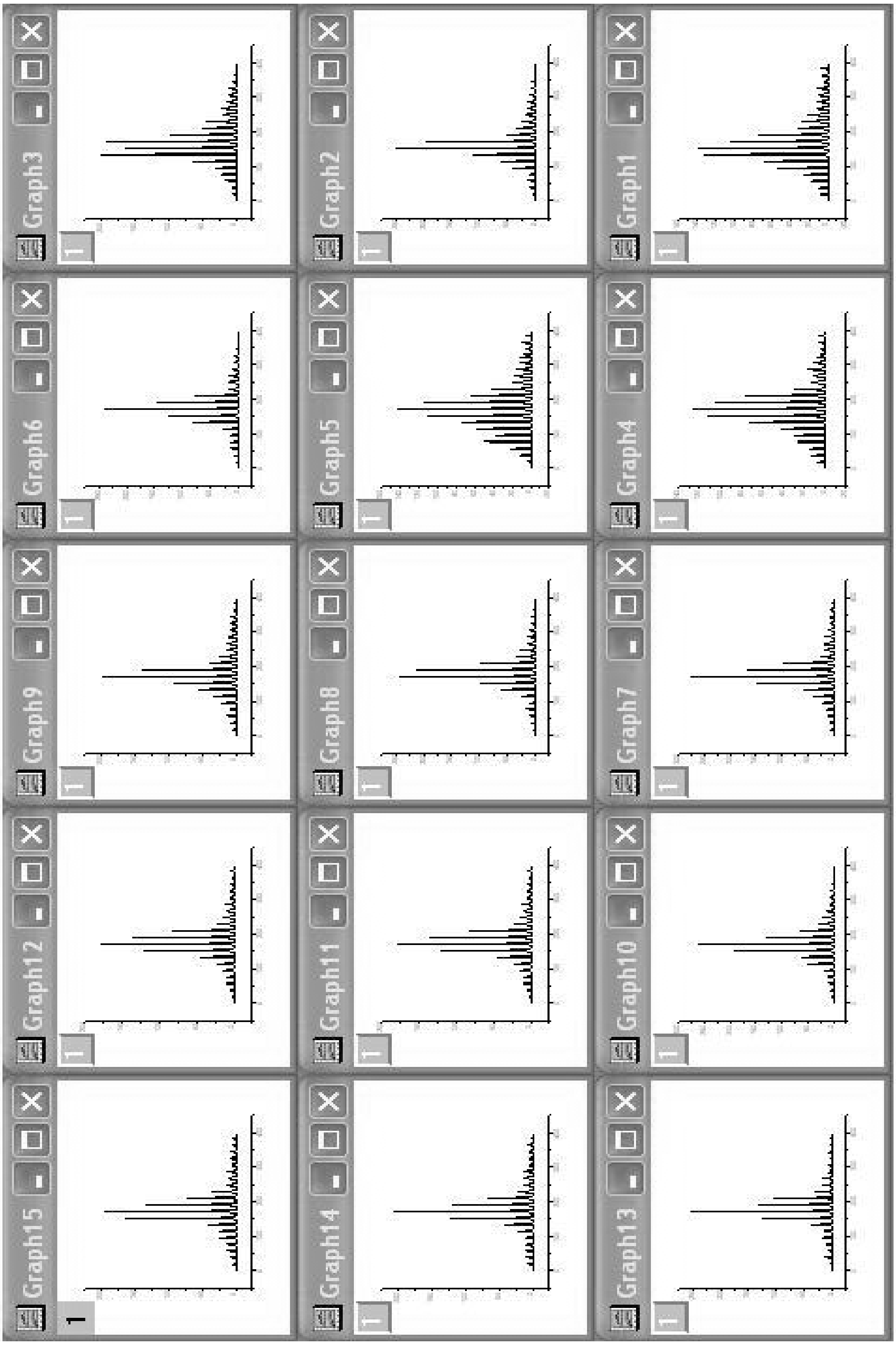


**Fig. 5.7(d):** Reconstructed Phase Space Distribution Parameter for 15 repeated utterances (Vowel **ഒ/o/**)

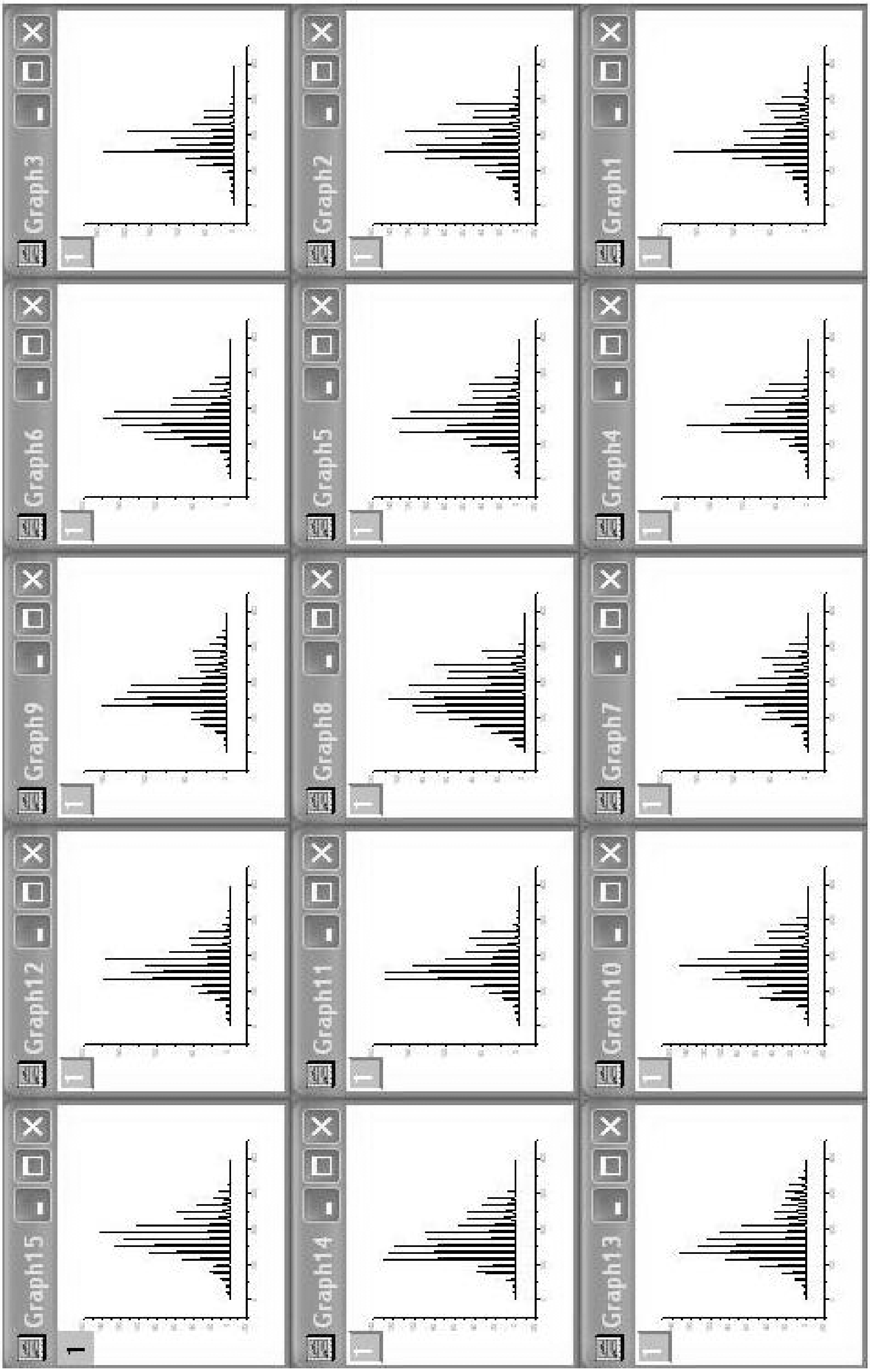


**Fig. 5.7(e):** Reconstructed Phase Space Distribution Parameter for 15 repeated utterances (Vowel ഉ / **u** / )

## 5.8 Conclusion

Inspecting the above distribution plots, it is evident that the recognition application based on these parameters will give improved accuracy. The method has a strong theoretical justification provided by the nonlinear dynamics literature. Since it is a time domain based approach, this method represents a fundamental philosophical shift from the frequency domain to the time domain. Here we are presenting an entirely different way of viewing the speech processing problem, and offering an opportunity to capture the nonlinear characteristics of the acoustic structure. The usefulness of the above parameter for efficient speech recognition application is studied and established in the later chapters.

# Chapter 6

# Nonlinear Phase Space Features for Robust Pitch Detection

## 6.1 Introduction

Pitch detection, also referred to as Fundamental frequency ($f_0$) estimation has been a popular research topic for many years, and is still being investigated today. Fundamental frequency, $f_0$ is the lowest frequency component, in the signal, which relates well to most of the other frequency components. In a periodic waveform, most of the components are harmonically related, meaning that the frequency of most of the components are related to the lowest component by a small whole-number ratio. The frequency of this lowest frequency component is ‘$f_0$’ of the waveform.

Isolated word recognition engines often discard the pitch information as irrelevant to the recognition task. While it is true that individual phonemes are recognizable regardless of the pitch of the driving function, or even in the absence of pitch as in whispered speech, this does not imply that pitch information is not useful. Much semantic information is passed on through pitch that is above the phonetic and lexical levels. In tonal languages, the relative pitch motion of an utterance contributes to the lexical information in a word. In this case, speech recognition algorithms must attend to the pitch or the context of the utterance to avoid ambiguity in continuous speech recognition applications. Pitch information can also be used for speaker identification.

Accurate pitch estimation plays a very important role in speech compression and speech synthesis, as well as in the musical world. A large number of *Pitch Determination Algorithms* (PDA) have been developed to date. Most of them can be loosely classified as *time-domain* or *short-term analysis* PDAs [Hess.W.J, 1992]. The most popular and reliable techniques in use today (for example, those based on correlation, spectrum or cepstrum) are short-term methods operating on short segments of a signal. $f_0$ estimators developed for a particular application, such as musical note detection or speech analysis, are well understood, but depend on the domain of the data, that is, a detector designed for one domain is less accurate when applied to a different domain [Rabiner.L.R and Schafer.R.W, 1978 ].

Most algorithms for pitch detection involve one or more of the following components. (i) preprocessing to enhance the periodicity of the waveform (eg. low pass filtering, center clipping), (ii) short time analysis of speech to obtain initial estimates of $f_0$ and (iii) post processing to correct isolated errors and produce smooth contours (eg. median filtering, dynamic programming). There are basically two kinds of determination methods (i) methods that operate in the time domain as the famous autocorrelation method [Rabiner.L.R and Schafer.R.W, 1978], [Hess.W.J, 1983] and (ii) methods that operate in the frequency domain typically rely on FFTs, detecting $f_0$ from peaks of the magnitude cepstrum or harmonic power spectrum [Hess.W.J, 1992, Bagshaw.P.C, *et al*, 1993].

However most of the conventional techniques are not fully satisfactory on real speech. One of the reasons for such deficiency is the linear nature of signal processing employed by many conventional methods. As explained in chapter 3, human speech production is a complex nonlinear and non-stationary process. Its complete and most accurate description can only be achieved in terms of nonlinear fluid dynamics. Traditionally it has been described using techniques like source-filter model and spectral analysis. These techniques work very well for many aspects of speech analysis, but they are inherently limited in their ability to describe the true dynamics of speech production. Consequently, to study such nonlinear aspects of speech production as excitation function, it is advantageous to dismiss traditional linear techniques and to use more general nonlinear approach. Without making too many simplifying assumptions one can state that voiced speech is generated by a relatively low-dimensional nonlinear dynamical system, while an unvoiced speech is generated by a high dimensional dynamical system.

In chapter 4 we have discussed that one of the profound results established in chaos theory is the celebrated Taken's embedding theorem [Taken.F, 1981], which states that it is possible to reconstruct a state space topologically equivalent to the original state space of a system from a single observable. Some attempts have already been made to apply nonlinear and chaotic signal analysis methods to pitch detection algorithms. In particular, it was previously noted that pitch period can be measured in state space by using Poincaré sections [Kubin.G, 1995]. However, a truly reliable and accurate method for pitch detection using state-space embedding of a signal has not been proposed to date.

In this chapter we introduce a general method for pitch estimation using reconstructed Phase Space in two dimensions. Theoretical matters are discussed first followed by implementation details and experimental results.

This chapter is organized in three sessions. In first session a general discussion about fundamental frequency or pitch in a signal and the methods pitch tracking and pitch marking is presented. Following this, conventional pitch detection algorithms based on time domain and frequency domain methods are discussed. Finally a new method for robust pitch detection based on Reconstructed Phase Space features is introduced.

### 6.2 Measuring Fundamental Frequency

Pitch is a perceptual quantity related to $f_0$ of a periodic or pseudo-periodic waveform. Note that in this thesis, the terms ‘pitch’ and its primary acoustical correlate ‘fundamental frequency’ are used inter changeably. To find the frequency of oscillation, it is sufficient to determine the period of such oscillation, the inverse of which is the frequency. The problem comes when the waveform consists of more than a simple sinusoid. As harmonic components are added to a sinusoidal waveform, the appearance of pitch of the waveform becomes less clear and the concept of “fundamental frequency” or $f_0$ must be considered. The goal of a $f_0$ estimator is to find $f_0$ in the midst of the other harmonically related components of the sound.

The difficulty of finding the $f_0$ of a waveform depends on the waveform itself. If the waveform has few higher harmonics or the power of the higher harmonics is small, the $f_0$ is easier to detect, as in Figures 6.1 and 6.2. If the harmonics have more power than the $f_0$, then the period is harder to detect, as in Figures 6.3 and 6.4.

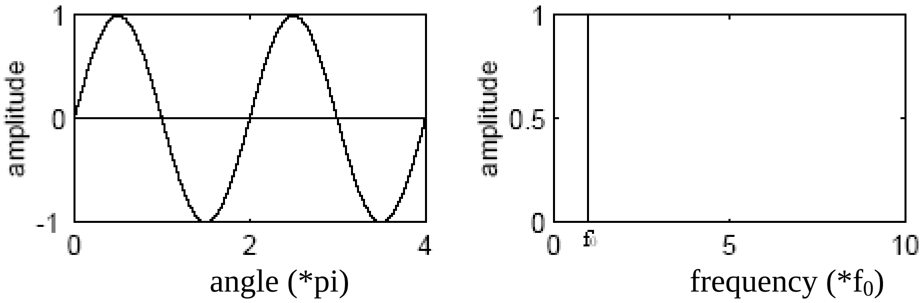


**Fig.6.1** Waveform with no upper harmonics

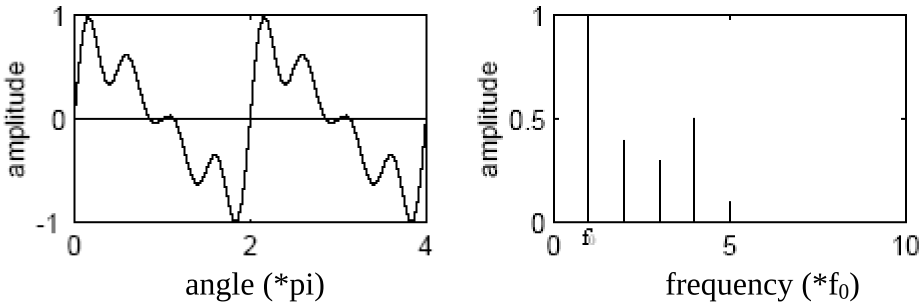


**Fig.6.2** Waveform with lower power upper harmonics

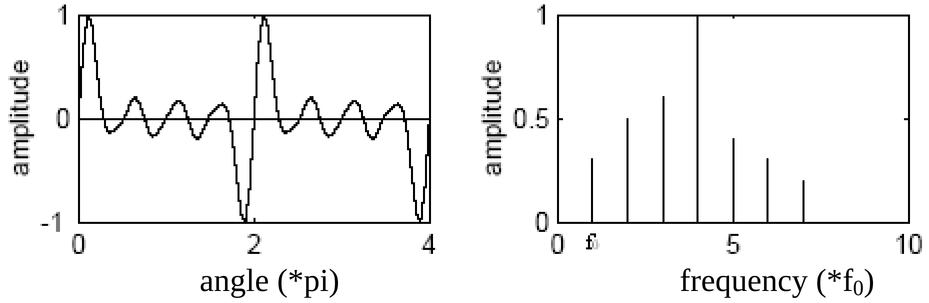


**Fig.6.3** Waveform with higher power upper harmonics

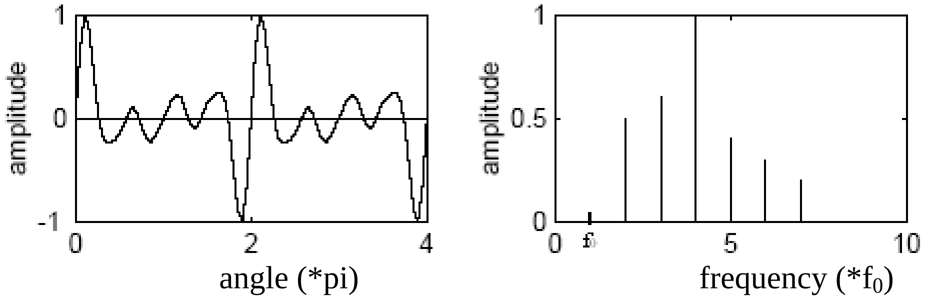


**Fig.6.4** Waveform where power of upper harmonics is very high

### 6.3 Pitch tracking and pitch marking

Pitch tracking consists of classifying speech into voiced and unvoiced regions, and for voiced regions determining the fundamental frequency of the vibrations of the vocal chords. Generally pitch tracking involves determining pitch ($f_0$) over a large interval of speech as shown in Figure 6.5. The pitch track is represented by the block dots overlaid on the spectrogram. Notice that each dot represents a value of $f_0$. The dots in the voiced region (0.05-0.3 seconds in the figure) have a non-zero value whereas dots in unvoiced region represent a value of zero frequency implying an absence of periodic component ($f_0$).

Pitch tracking algorithms generally do not identify the temporal location of each vibration of the vocal chords but rather are based on average time spacing between a numbers of vibrations. The averaging, typically due to the use of an autocorrelation type of calculation in the pitch tracking, is used to improve the accuracy of the tracking.

Pitch marking (PM), on the other hand, attempts to locate every vibration of the vocal chords. That is, the beginning and end of each pitch cycle is to be located by timing markers. PM does not involve classifying speech into voiced or unvoiced regions but rather may use such pre-existing knowledge for locating pitch cycle markers. Figure 6.6 shows the markers by vertical dotted lines, identified in pitch marking process for an actual speech signal.

Since the speech signal is not really periodic and also highly non-stationary, pitch tracking and pitch marking turn out to be extremely difficult problems to solve accurately. That is, even over short time intervals of the order of 50 ms, the speech signal is often changing in $f_0$, in amplitude and in overall spectral characteristics.

For voiced speech, typical $f_0$ ranges are of the order of 50-250 Hz for male speakers, 120-400 Hz for female speakers and around 150-450 Hz for child speakers, these ranges differ for different speaker conditions [Baken.R, 1987]. However, there are wide variations from one individual to another. Even during the normal speech of a single speaker, there can be pitch variations spanning from one to four octaves. These wide variations, and other factors, make it very difficult to detect pitch with 100% accuracy with any set of parameters.

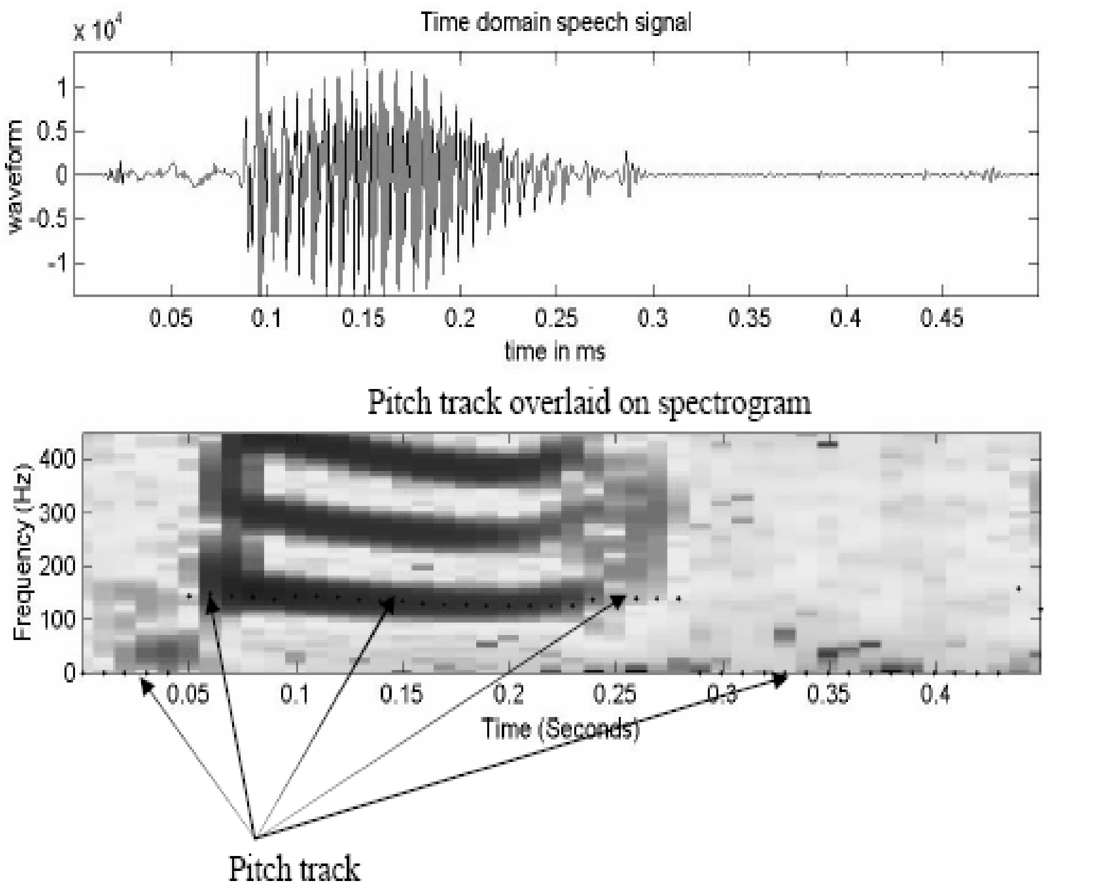


**Fig.6.5:** Illustration of an example of pitch track of a speech signal. The top panel shows the time domain speech signal and bottom panel shows the pitch track overlaid on a spectrogram of the speech signal.

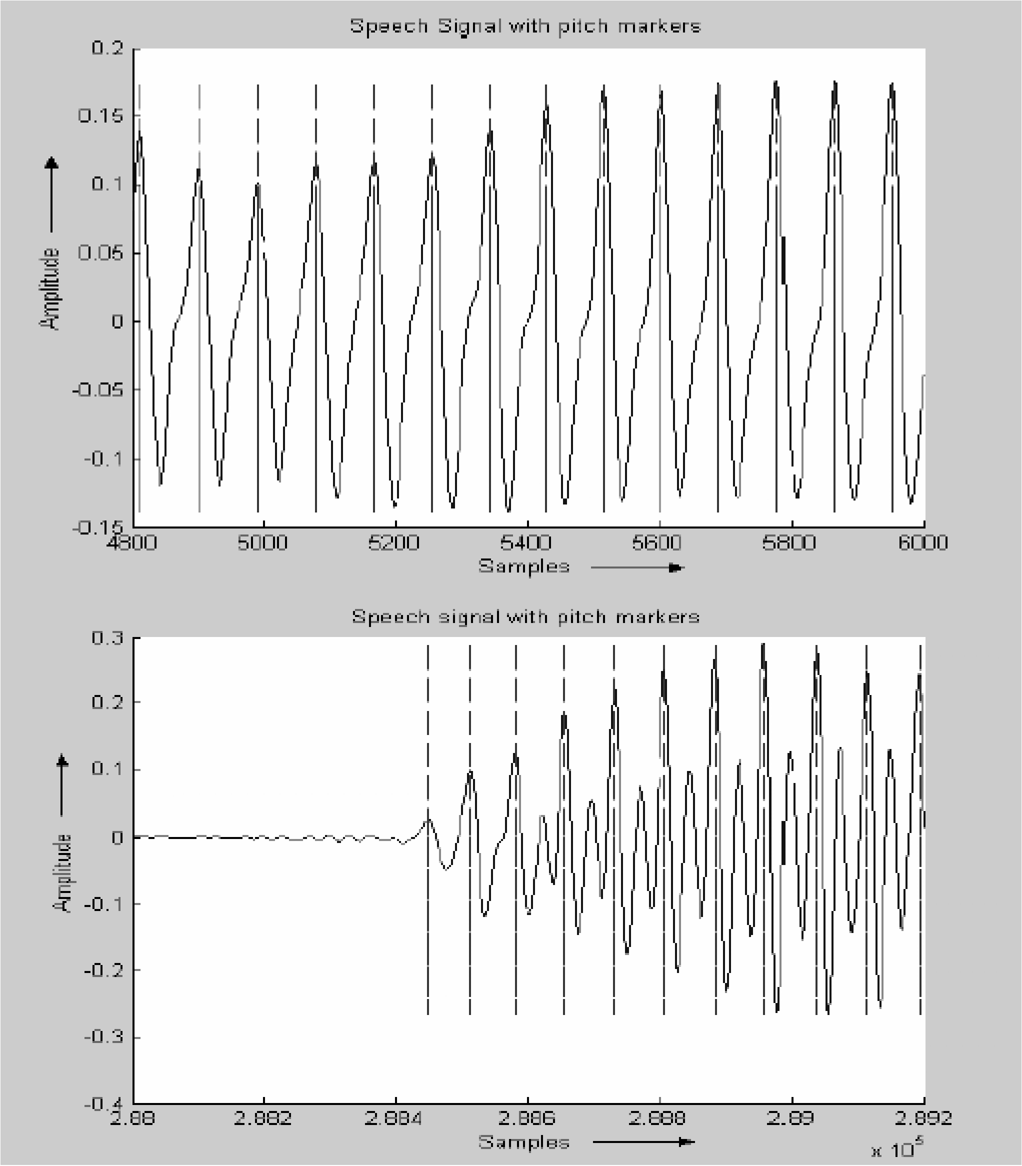


**Fig.6.6:** Illustration of pitch markers identified in speech signal for different regions of speech. The top panel shows markers in speech signal that

is purely voiced whereas the bottom panel shows markers in a section of speech signal that has unvoiced and voiced region.

### 6.4 Time-Domain Methods for Pitch Detection

Pitch tracking algorithms can be broadly classified into the following categories: time domain based and frequency domain based pitch tracking There is a family of related time-domain $f_0$ estimation methods which seek to discover how often the waveform fully repeats itself. The theory behind these methods is that if a waveform is periodic, then there are extractable time-repeating events that can be counted, and the number of these events that happen in a second is directly related to the frequency. Each of these methods is useful for particular kinds of waveforms. If there is a specific time-event that is known to exist once per period in the waveform, such as a discontinuity in slope or amplitude, it may be identified and counted. There are many popular methods in time domain, such as Zero-crossing rate (ZCR), Peak rate, Slope event rate, Auto correlation etc, where the last one is the most reliable method.

### 6.4.1 Autocorrelation Method for Pitch Detection

We briefly discuss the autocorrelation in this section. The auto-correlation approach is the most widely used method for estimating the pitch of a periodic signal. The correlation between two waveforms is a measure of their similarity. The waveforms are compared at different time intervals, and their “sameness” is calculated at each interval. The result of a correlation is a measure of similarity as a function of time lag between the beginnings of the two waveforms. The autocorrelation function is the correlation of a waveform with itself. One would expect exact similarity at a time lag of zero, with increasing dissimilarity as the time lag increases. In the next chapter we have explained it in detail during the calculation of proper time delay.

Periodic waveforms exhibit an interesting autocorrelation characteristic: the autocorrelation function itself is periodic. As the time lag increases to half of the period of the waveform, the correlation decreases to a minimum. This is because the waveform is out of phase with its time-delayed copy. As the time lag increases again to the length of one period, the autocorrelation again increases back to a maximum, because the waveform and its time-delayed copy are in phase. For periodic signals, the autocorrelation function attains a maximum at sample lags of 0, ±P, ±2P, etc., where P is the period of the signal.

A major limitation of the *auto-correlation* function is that it may contain many other peaks other than those due to basic periodic components. For speech signals, the numerous peaks present in the *auto-correlation* function are due to the damped oscillations of the vocal tract response [Talkin.D, 1995]. It is difficult for any simple peak picking process to discriminate those peaks due to periodicity from these “extraneous” peaks. The peak picking is more robust if a relatively large time window is used, but has the disadvantage that the rapid changes in pitch cannot be tracked properly.

Some of the shortcomings in the *auto-correlation* method are overcome using the *cross-correlation* function The *Normalized Cross Correlation* function is very similar to the auto-correlation function, but is better able to follow the rapid changes in pitch and amplitude. The major disadvantage is an increase in the computational complexity

## 6.5 Frequency-Domain Methods for Pitch Detection

There is much information in the frequency domain that can be related to the $f_0$ of the signal. Pitched signals tend to be composed of a series of harmonically related partials, which can be identified and used to extract the $f_0$. Many attempts have been made to extract and follow the $f_0$ of a signal in this manner. There are many popular methods based on frequency domain, such as Component Frequency Ratios, Filter-Based Methods, Cepstrum Analysis etc, in which the last one is more efficient.

### 6.5.1 Cepstrum Analysis Method for Pitch Detection

As mentioned above, another method for pitch tracking technique is computation of the C*epstrum,* followed by peak picking over a suitable range. The Cepstrum is defined as the inverse Fourier Transform of the short time log magnitude spectrum.

$$C(t) = F^{-1}( \log | F(x(t)) |^{2})$$

Where, x(t) = the input signal under consideration.

For voiced speech, the *Cepstrum* tends to have local maxima at times, kT, corresponding to integer multiples of the glottal periods. The “log” in the

*Cepstrum* equation tends to flatten the harmonic peaks in the spectrum and thus leads to more distinct peaks in the *Cepstrum* function, as compared to the peaks in the autocorrelation function.

The cepstrum method assumes that the signal has regularly-spaced frequency partials. If this is not the case, such as with the inharmonic spectrum the method will provide erroneous results. As with most other $f_0$ estimation methods, this method is well suited to specific types of signals. It was originally developed for use with speech signals, which are spectrally rich and have evenly spaced partials. Thus, as for *autocorrelation* methods, in regions where there are rapid changes in $f_0$, the method does not perform well.

The computational difficulties in the above time domain and frequency domain approaches in the detection of pitch can be considerably reduced by developing a pitch detection method in the Phase Space domain. The following session describes the detailed formulation of a Pitch Detection Algorithm based on the Phase Space Analysis.

## 6.6 Phase Space and Frequency

The phase space signal representation is a way of observing the short-time history of a waveform in a way that makes repetitive cycles clear. As explained in the earlier chapters the basic phase space representation is to plot the value of the waveform at time t versus the slope of the waveform at the same point. Phase space of a dynamical system can be reconstructed through the use of time-delayed versions of the original scalar measurements. We know that according to Takens’ embedding theorem [Taken.F 1981], it is possible to reconstruct a Phase space, topologically equivalent to the original state space of a system from a single observable. Any periodic signal forms a closed cycle in phase space, and the shape of the cyclic path depends on the harmonic composition of the signal. The $f_0$ of a signal is related to the speed with which the path completes the cycle in phase space. The task then becomes detecting the difference between new values in phase space crossing the old path, and new values intersecting and re-tracing the old path. The simplest solution would be to compare distances between points in phase space, and detect when the distance becomes minimal. An initial point would be selected, and the distance from that point would be traced as a function of time. When this distance became zero (or a minimum value) the waveform may have repeated.

A bigger problem with phase space $f_0$ estimation is how to deal with pseudo-periodic signals. In a phase space representation, the path of a pseudo-periodic signal will never re-trace itself, although it will follow a closely parallel path.

A Poincar´e section of a phase space plot is a lower-dimensional orthogonal slice through the plot which produces a cross-section of a path being considered. A Poincar´e section of a periodic signal will be one or more discrete points, indicating the locations that the path intersects the section. Figure 6.7 shows the Poincar´e section for a periodic signal in 3 dimensional Phase Space.

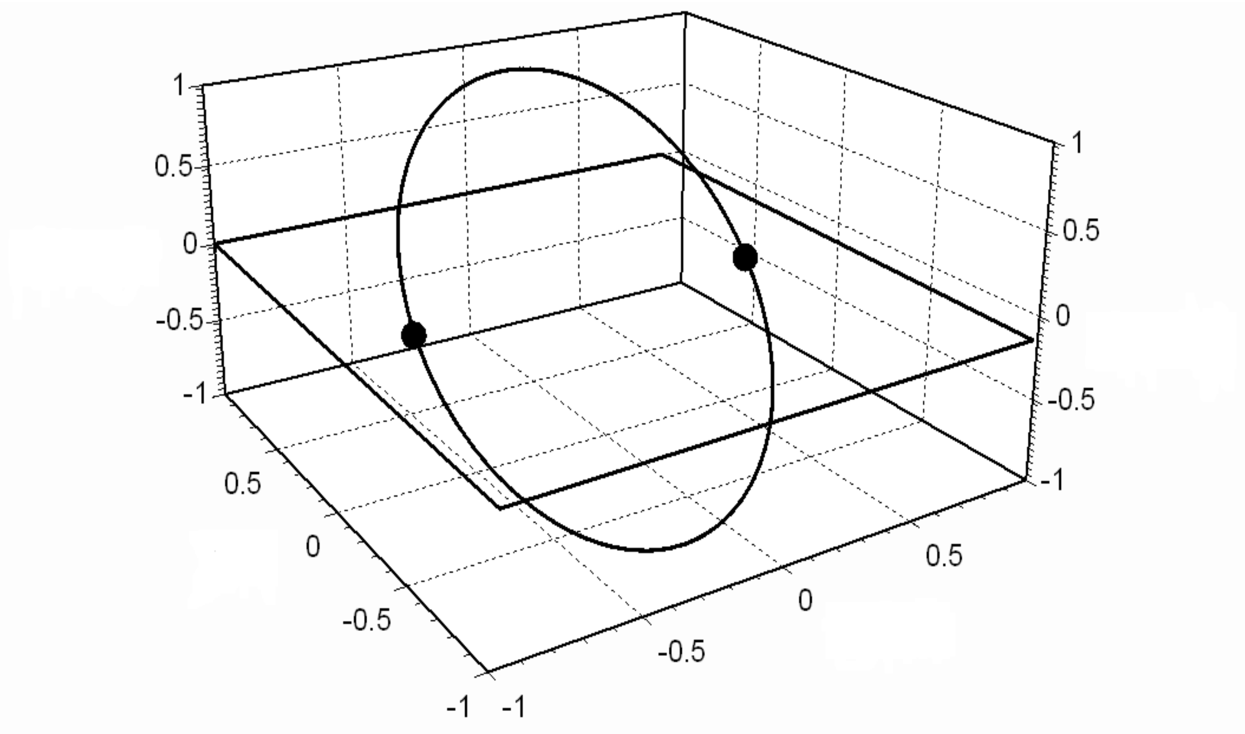


**Fig.6.7:** Poincare's section for a periodic signal in a 3 dimensional Phase Space

A pseudo-periodic signal will generate a cloud of points in a Poincar´e section, localized in one or more clusters. If these clusters are separate, the mean location of each cluster can be treated as the intersection point for that cluster, and the period can be calculated by the time lag between successive points in the same cluster.

A problem arises when two clusters of points are close together, such that for some points it is not clear which cluster they should belong to. In this case, higher-dimension phase-space representations should be employed until the clusters are shown to be disjointed. There are many potential problems with this suggested method such as the dimensionality of Phase space etc., but it may provide another alternative to the many $f_0$ estimation algorithms that are currently available.

In our proposed method the algorithm is not depending on the dimension of Phase space and not on the type of data being investigated.

### 6.7 Reconstructed Phase Space for a Periodic signal

A periodic signal forms closed loops in Phase space. For a periodic complex signal the displacements at two points with a phase difference of 2π would have same values. This implies that in the two dimensional Phase Space diagram the points representing such pairs would be lying on a straight line with a slope of π/4 to the axes (Phase Space Diagonal), and this is the Reconstructed Phase Space with time lag corresponds to time period of the signal 'T'. Figure 6.8 shows the Reconstructed Phase Space for a sinusoidal signal with a time lag T.

It may be seen that as the phase lag decreases the points are scattered over a broad region. The scattered region reaches a maximum width for a phase difference of π/2, or time lag corresponds to 'T/4'. With this time delay, a pure sinusoidal signal forms circular orbits in the Phase space. It collapses into a straight line when the phase difference becomes 2π, which corresponds to the time period 'T'. Figure 6.9 gives the Reconstructed Phase Space for a sinusoidal signal with time lag T/4, and Figure 6.10 represents the RPS diagram with time lag T/1.5.

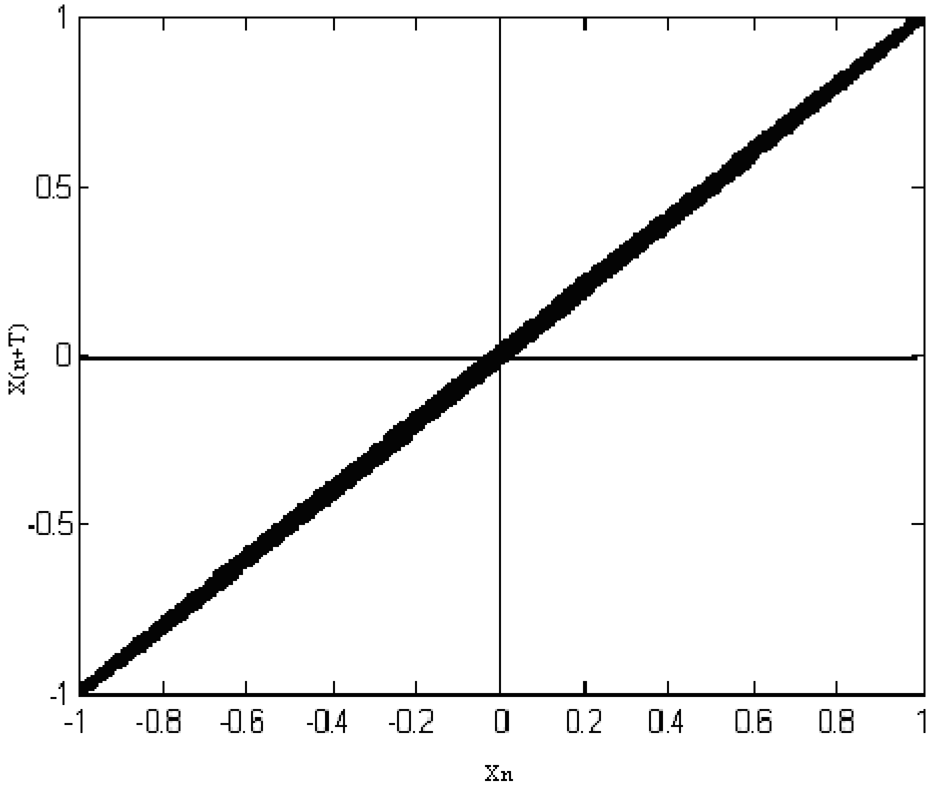


**Fig. 6.8:** Reconstructed Phase Space for sinusoidal signal with time lag T

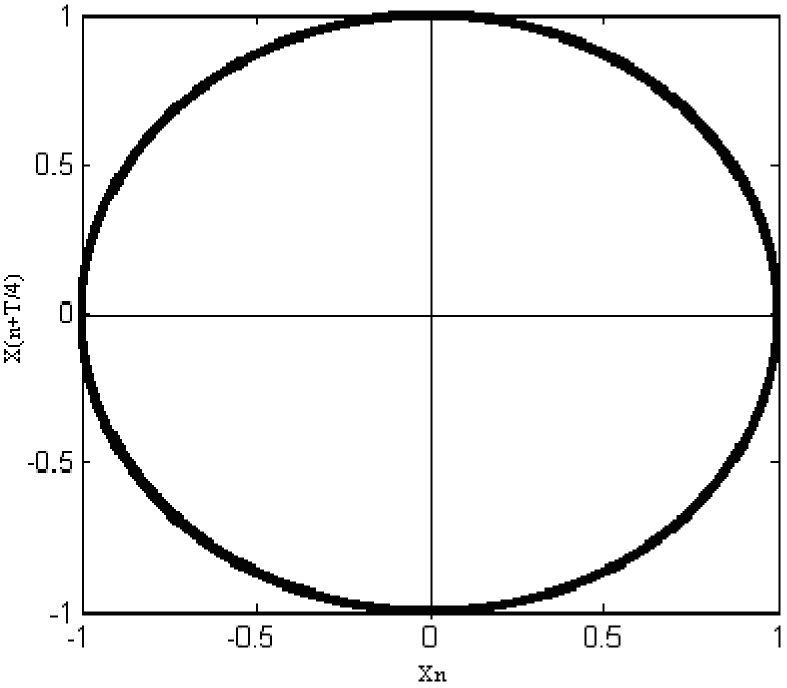


**Fig. 6.9:** Reconstructed Phase Space for sinusoidal signal with time lag T/4

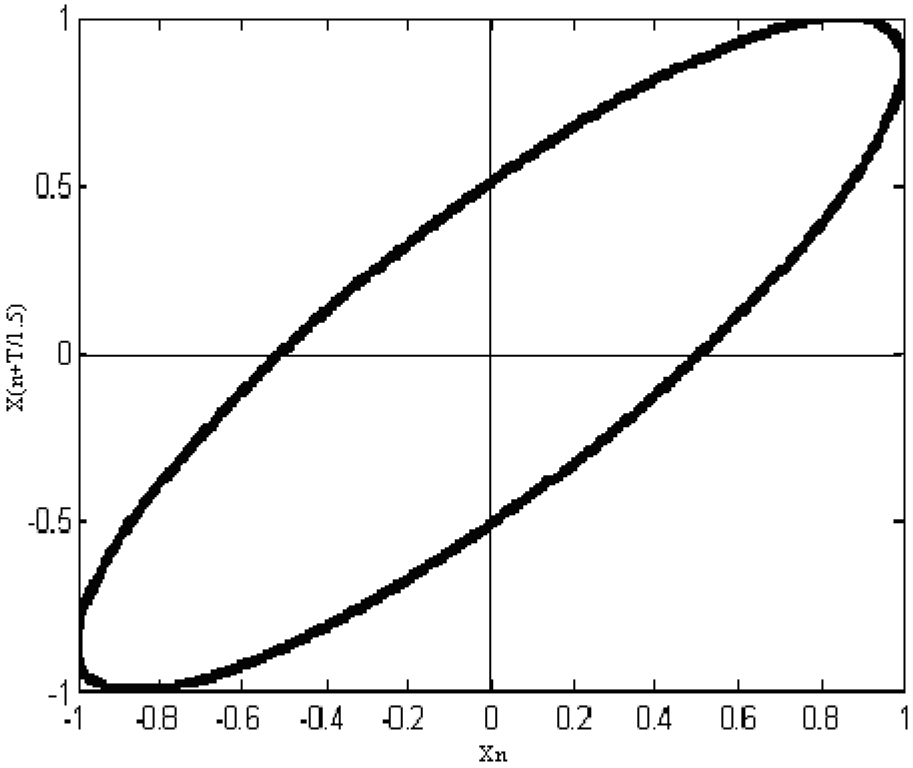


**Fig. 6.10:**Reconstructed Phase Space for sinusoidal signal with time lag T/1.5

Therefore we can say that in the Reconstructed Phase Space of a complex quasi periodic signal, if the time lag introduced τ corresponds to the pitch period, the spread of Phase Space points from the Phase Space Diagonal with slope π/4 will be minimum. The spread can be measured in terms of perpendicular Euclidean distance of the Phase Space points from the diagonal.

### 6.8 Euclidian Distance Measure from the Phase Space Diagonal

Figure 6.11 shows the two dimensional Phase Space Reconstruction of Malayalam vowel ഔ/Λ/. The spread of Phase Space points from the Phase Space diagonal can be quantified in terms of Euclidean Distance Measure from the diagonal as shown in the figure 6.11.

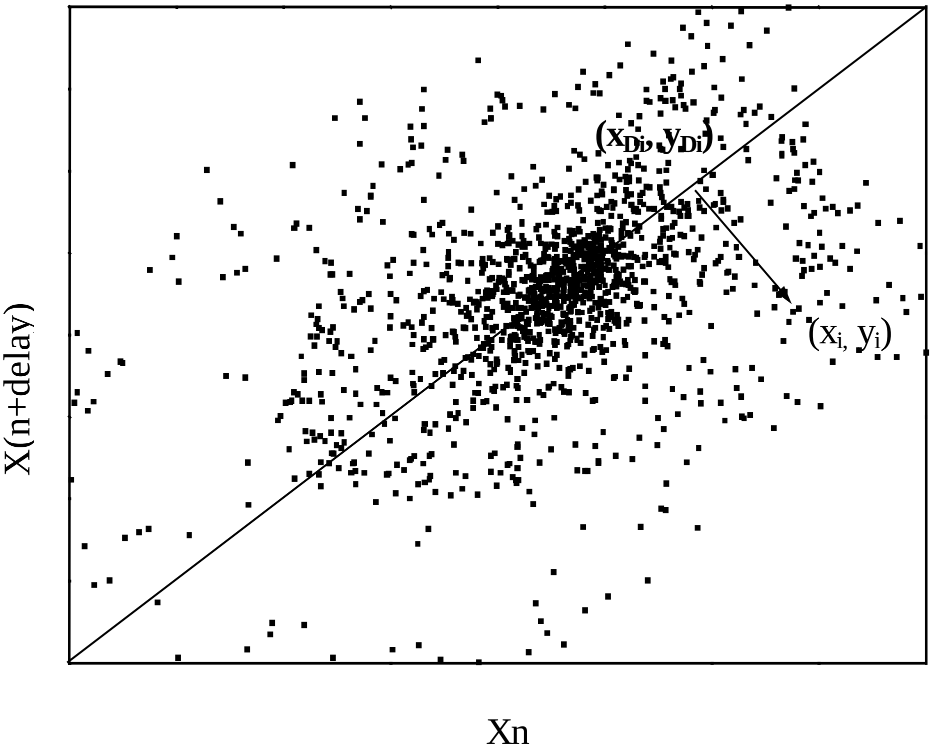
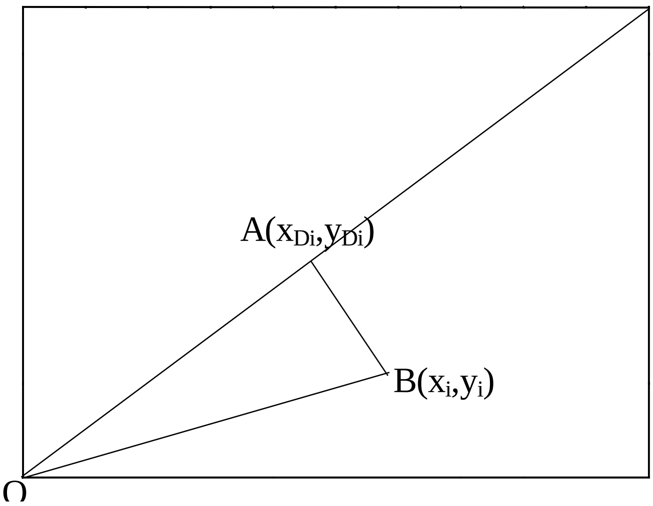


**Fig.6.11** Euclidean Distance Measure from Phase Space Diagonal

The Euclidean distance measure of the i<sup>th</sup> Phase Space point with co-ordinates $x_i$ and $y_i$ is given by

$$d_i = \sqrt{(x_i - x_{Di})^2 + (y_i - y_{Di})^2} \qquad \dots\dots\dots\dots\dots\dots\dots\dots (6.1)$$

Here $x_{Di}$ and $y_{Di}$ represent the co-ordinates of the point where the perpendicular from the i<sup>th</sup> Phase Space point meets the Phase Space Diagonal. Since the points $x_{Di}$ and $y_{Di}$ are on the diagonal, we have $x_{Di} = y_{Di}$



Also from triangle OAB in the above figure, we have $(OB)^2 = (OA)^2 + (AB)^2$

$$\text{Or} \quad (x_i^2 + y_i^2) = (x_{Di}^2 + y_{Di}^2) + (x_{Di} - x_i)^2 + (y_{Di} - y_i)^2$$

$$x_i^2 + y_i^2 = x_{Di}^2 + y_{Di}^2 + x_{Di}^2 + x_i^2 - 2x_{Di}\, x_i + y_{Di}^2 + y_i^2 - 2y_{Di}\, y_i$$

$$\text{ie} \quad 0 = 2\, x_{Di}^2 + 2\, y_{Di}^2 - 2x_{Di}\, x_i - 2y_{Di}\, y_i$$

Since $x_{Di} = y_{Di}$, $0 = 4\, x_{Di}^2 - 2x_{Di}\, x_i - 2\, x_{Di}\, y_i$

$$\text{Or} \quad x_{Di}\, (\, 2\, x_{Di}^2 - x_i - y_i\,) = 0$$

Since $x_{Di} \neq 0$, we have $(\, 2\, x_{Di}^2 - x_i - y_i\,) = 0$

$$2\, x_{Di}^2 = (x_i + y_i\,)$$

$$x_{Di} = \frac{x_i + y_i}{2}$$

Knowing $x_{Di}$ and $y_{Di}$ in terms of $x_i$ and $y_i$ we can calculate the Euclidean distance measure $d_i$.

Then the total Euclidean Distance Measure

$$D = \sum_{i=1}^{N-1} d_i \qquad \text{……………………..(6.2)}$$

The Distance Measure is calculated using equation 6.2 for the speech segment with a time delay embedding. This process is repeated by increasing the time delay. It can be seen that at a particular delay, the Distance Measure becomes minimum and that delay is the pitch candidate for the speech segment.

Repeating the procedure and increasing the delay up to the last sample in the frame can select the entire pitch candidates. The selected pitch candidates are shown in the Delay versus Distance Measure graph (figures 6.12(a-e)). Once the pitch candidates are selected in terms of time delays, pitch period or fundamental frequency can be computed by knowing the sampling frequency.

### 6.9 Computational Efficiency

The proposed method requires finding the least distance among N-1 points in the 2 dimensional space, where N is proportional to sampling rate. The only expensive part of the procedure is the N-1 Euclidean distance measure. Therefore Compared to conventional  Pitch Detection Algorithms the present method is computationally efficient.

### 6.10 Experimental Results

Usually the implementation of short term Pitch Determination Algorithm includes three usual stages [Talkin.D, 1995] (a) signal pre-processing, (b) generation of pitch period candidates and (c) post-processing. The proposed method works well on raw speech waveforms and does not explicitly require any signal pre-processing.

In this work 5 Malayalam vowels, namely അ/Λ/, **ഇ/I/**, **എ/ae/**, **ഒ/o/**, and **ഉ/u/**. are taken for the experiment. For each vowel, Minimum Distance Measure and the corresponding time delays are calculated. Pitch candidates obtained as described above usually include only a true pitch period and its integer multiples. Selecting the lowest multiple can give a reliable local pitch estimate for the vowel.

Delay versus Distance Measure graph for the vowels are shown in figures 6.12(a-e). Let $\tau_m$ be the delay corresponding to first minimum of distance measure. Then the pitch period T is calculated as

$$T = \tau_m \times T_s$$

Where $T_s$ is the sampling period ( 125μ sec. if sampling frequency is 8 kHz.). Then the fundamental frequency $f_0$ = 1 / T. The experiment is repeated for different samples of vowel sounds. Minimum pitch, maximum pitch and average pitch is tabulated in table 6.1

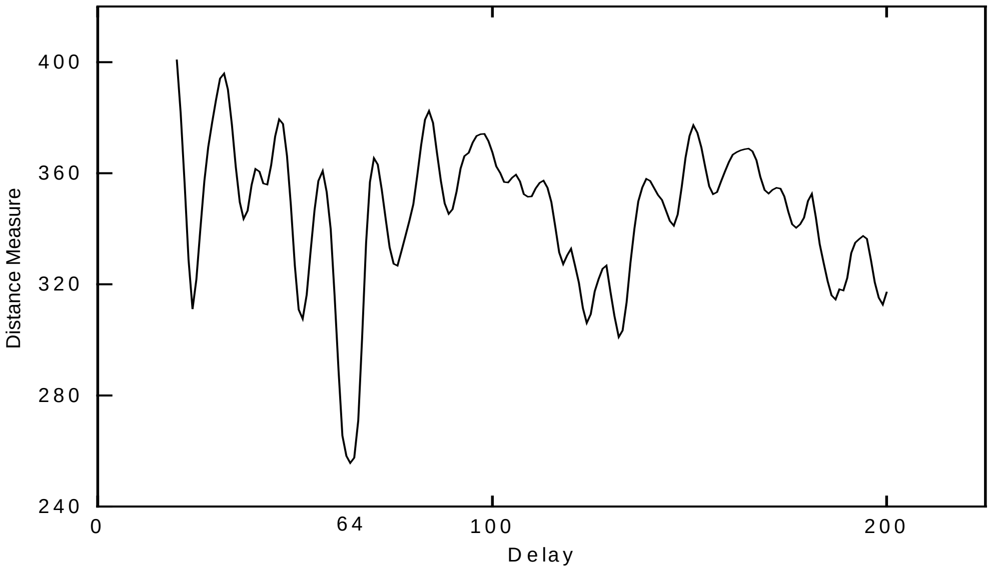


**Fig. 6.12(a):** Delay versus Distance Measure graph for vowel അ/Λ/

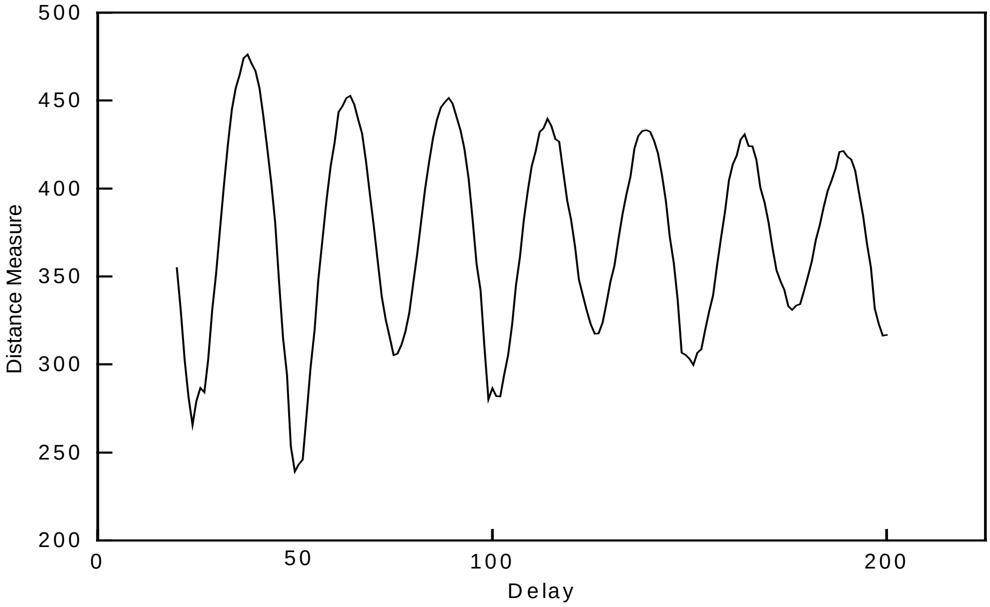


**Fig. 6.12(b):** Delay versus Distance Measure graph for vowel **ഇ/I/**

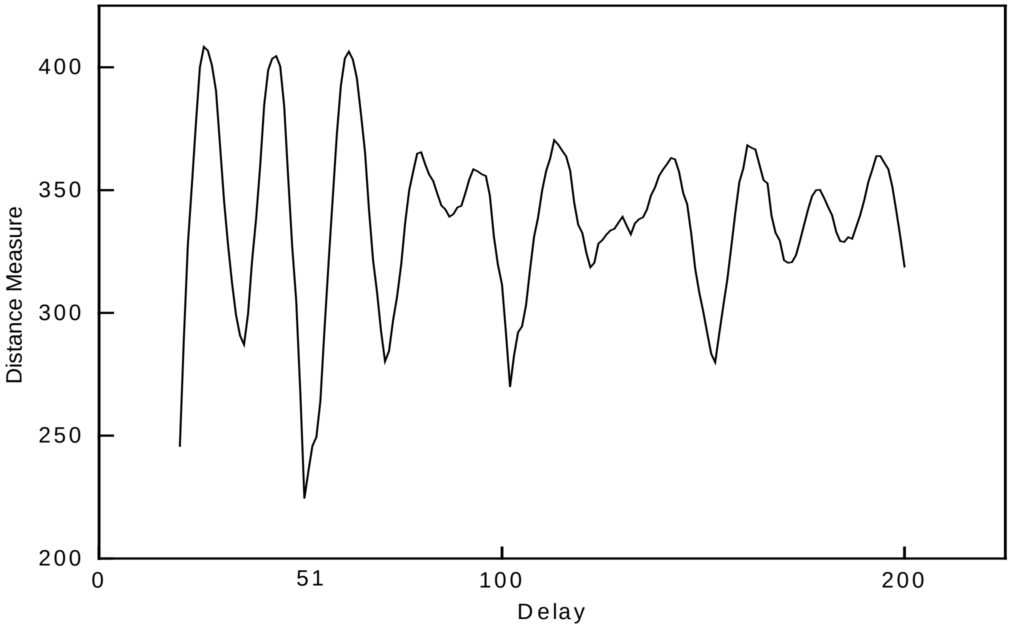


**Fig. 6.12(c):** Delay versus Distance Measure graph for vowel ഐ**/ae/**

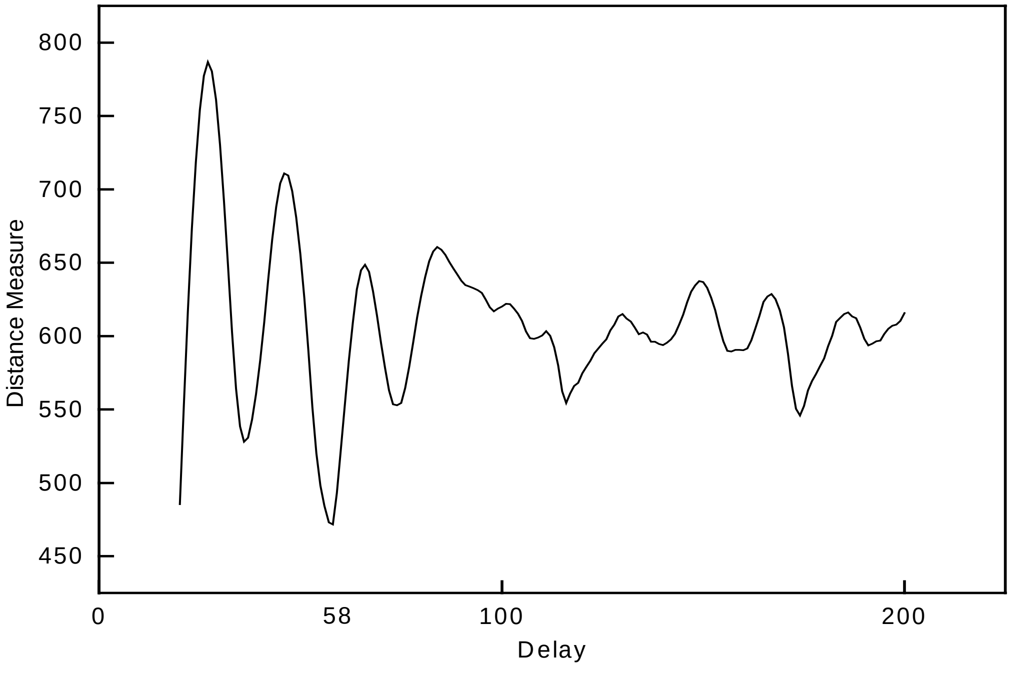


**Fig. 6.12(d):** Delay versus Distance Measure graph for vowel ഒ**/o/**

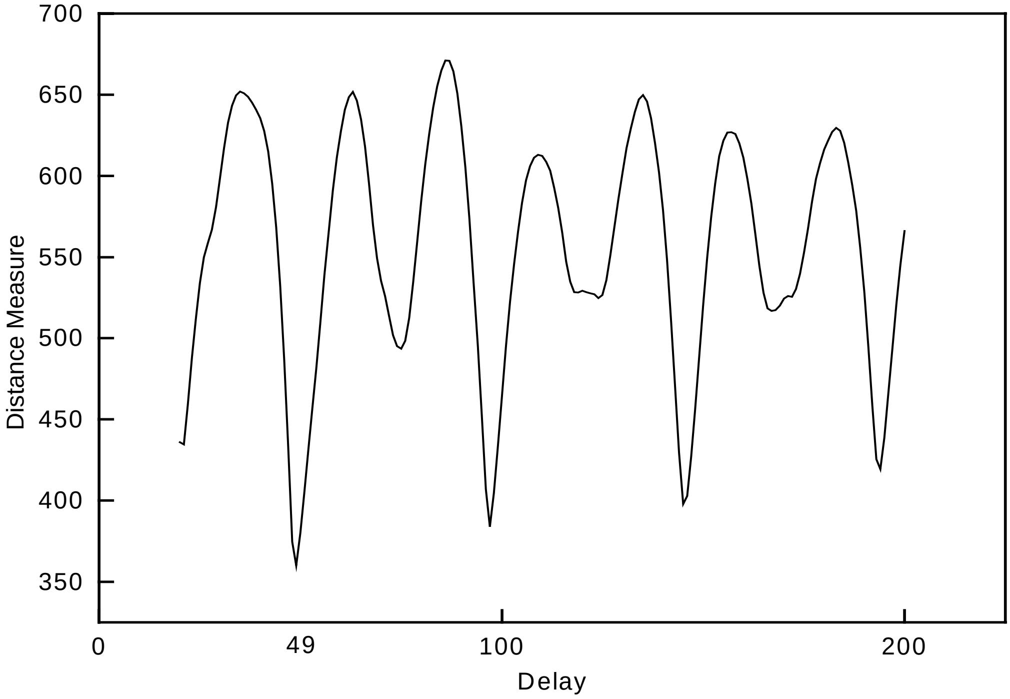


**Fig. 6.12(e):** Delay versus Distance Measure graph for vowel **ഉ/u/**

| **Vowel Unit** | **Minimum Pitch (Hz)** | **Maximum Pitch (Hz)** | **Average Pitch (Hz)** |
|---|---|---|---|
| അ/Λ/ | 119.40 | 135.59 | 125.00 |
| **ഇ/I/** | 114.28 | 166.67 | 160.00 |
| **എ/ae/** | 106.66 | 163.26 | 156.86 |
| **ഒ/o/** | 105.26 | 148.15 | 137.93 |
| **ഉ/u/** | 106.66 | 170.21 | 163.26 |

**Table. 6.1:** Experimental Results

The pitch value is also estimated using conventional methods autocorrelation and cepstrum methods. A comparison of average value of pitch obtained by these methods and the pitch value obtained by the newly proposed Reconstructed Phase Space based method are shown in table 6.2.

| **Vowel Unit** | **Average Pitch in Hertz** | |
|---|---|---|
| | **Conventional Method (Hz)** | **RPS Method (Hz)** |
| അ/Λ/ | 126.21 | 125.00 |
| ഇ/**I**/ | 158.47 | 160.00 |
| എ/**ae**/ | 153.98 | 156.86 |
| ഒ/**o**/ | 131.77 | 137.93 |
| ഉ/**u**/ | 165.48 | 163.26 |

**Table. 6.2:** Comparison of results with conventional methods

### 6.11 Conclusion

Methodologies originally developed for analyzing chaotic time series have been successfully applied to pitch determination problem. The proposed new method does not suffer from the limitations of other short-term pitch estimation techniques. The algorithm is very straightforward and flexible. The preliminary evaluation results show its robust performance on real speech. The experimental results show that the pitch estimated using Reconstructed Phase Space features agrees with that obtained using conventional Pitch Detection Algorithms.

## Chapter 7

# Optimum Embedding Parameters for Phase Space Reconstruction

### 7.1 Introduction

Phase space reconstruction is usually the first step in the analysis of dynamical systems. Typically, an experimenter obtains a scalar time series from one observable of a multidimensional system. State-space reconstruction is then needed for the indirect measurement of the system’s invariant parameters like, dimension, Lyapunov exponent etc. Several techniques for attractor reconstruction are currently employed, such as derivative coordinates [Packard.N.H, Shaw.R.S, *et.al.,* 1980], [Taken.F, 1981] and principal components (or singular value decomposition) [Broomhead.D.S. and King. G.P, 1986], the method of delays [Packard.N.H., Shaw.R.S, *et al.* 1980], [Taken.F, 1981] etc.

Takens' theorem [Taken.F, 1980] demonstrates that in the absence of noise a multidimensional state space can be reconstructed from a scalar time series, using the method of delays. It is the most widespread approach because it is the most straightforward and the noise level is constant for each delay component [Casdagli.M, Eubank.S, *et. al.*, 1991]. The method of delays reconstructs the attractor dynamics by using *delay coordinates* to form multiple state-space vectors, $\mathbf{X}_n$. The reconstructed state of the system at each discrete time *n* is

$$\mathbf{X}_n = [x_n \;\; x_{n+\tau} \;\; \ldots \; x_{n+(d-1)\,\tau} \,],$$

Where τ is the reconstruction time delay and d is the embedding dimension. Taken's theorem gives little guidance, about practical considerations for reconstructing a good state space. It is silent on the choice of time delay (τ) to use in constructing d-dimensional data vectors. Indeed, it allows any time delay as long as one has an infinite amount of infinitely accurate data. However, for reconstructing state spaces from real-world, finite, noisy data, it gives no direction [Casdagli.M, Eubank.S, *et. al.*, 1991].

Experiments show that the quality of reconstruction depends on the value chosen for 'τ' and experimenters and theorists note that there are no criteria for choosing 'τ' in the literature. The following facts are to be remembered while choosing proper time delay [Abarbanel.H.D.I, 1996.]

i) It must be some multiple of the sampling time $T_s$, since we only have data at those times.

ii) If time delay is too short, the coordinates, which we wish to use in our reconstructed data vector, will not be independent enough. That is, not enough time will have evolved for the system, for its phase space to produce new information about that phase space. If 'τ' is too small each coordinate is almost the same, and the trajectories of the reconstructed space are squeezed along the identity line; this phenomenon is known as *redundance.*

iii) If $\tau$ is too large, in the presence of chaos and noise, the dynamics at one time become effectively causally disconnected from the dynamics at a later time, so that even simple geometric objects look extremely complicated; this phenomenon is known as *irrelevance* [Casdagli.M, Eubank.S, *et. al.*, 1991].

For the embedding dimension ‘d’, the theorem states the sufficient (but not necessary) condition is $d \geq 2D$, where D is the fractal dimensions of the underlying attractor. Normally, one has no a priori knowledge regarding the topological dimension and it is unclear what values of d will satisfy the condition.

Most of the research on the state space reconstruction problem has centered on the problems of choosing ‘$\tau$’ and dimension ‘d’ for delay coordinates. In practice, the determination of the time lag can be difficult, because there is no theoretically well-founded method to ascertain it, where an obvious criterion function can be formulated. Despite this limitation, two heuristics have been developed in the literature for establishing a time lag [Kantz.H and Schreiber.T, 2003]. 1) The first zero of the autocorrelation function and 2) the first minimum of the auto-mutual information curve [Fraser.A.M and Swinney.H.L, 1986]. These heuristics are premised on the principle that it is desirable to have as little information redundancy between the lagged versions of the time series as possible.

There have been many discussions on how to determine the optimal embedding dimension from a scalar time series based on Taken’s theorem or its extensions [Sauer.T, Yorke.J.A., and Casdagli.M, 1991]. The standard way to find the minimum embedding dimension ‘d’ is to use some criterion, which the geometry of the attractor must meet and check for which embedding dimension this is fulfilled. Among different geometrical criteria, the most popular seems to be the method of False Nearest Neighbors [Kennel.M.B, Brown.R, and Abarbanel.H.D.I, 1992]. This criterion concerns the fundamental condition of no self-intersections of the reconstructed attractor.

In this chapter we are discussing the selection of proper time delay in terms of first zero of autocorrelation function and first minimum of auto mutual information function along with a new method in terms of Maximum Euclidean Distance Measure. For the embedding dimension we have used Cao’s algorithm in which the method overcomes the shortcomings of the conventional techniques for finding False Nearest Neighbors. With these optimum embedding parameters, Reconstructed Phase Space Distribution Parameter (RPSDP) is modified.

In the first session of this chapter, selection of optimum time delay for Phase Space Reconstruction based on first zero of Autocorrelation function is described. Following the limitations of this method, selection of ‘τ’ based on first minimum of average mutual information function is presented. Then a new technique for the selection of proper time delay is introduced. In the next session, minimum embedding dimension for Phase Space Reconstruction is determined using Cao's method. Finally modification of RPSDP with the optimum embedding parameters is presented.

### 7.2 Proper Time Delay based on Autocorrelation Function

Autocorrelation is a mathematical tool used frequently in signal processing for analysing functions or series of values, such as time diomain signals. Informally, it is a measure of how well a signal matches a time-shifted version of itself, as a function of the amount of time shift. Autocorrelation is useful for finding repeating patterns in a signal, such as determining the presence of a periodic signal which has been buried under noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies.

The Autocorrelation function (ACF) shows the value of the autocorrelation coefficient for different time lags τ [Abarbanel.H.D.I, 1996] :

$$C(\tau) = \frac{E[(X_n - \mu)(X_{n+\tau} - \mu)]}{\sigma^2} \qquad \text{….(7.1)}$$

Where E is the expectation value, μ is the mean and σ the standard deviation. The estimation of the autocorrelations from a time series is straightforward as long as the lag τ is small compared to the total length of the time series and C(τ) quantifies how these points are distributed. If they spread out evenly over the plane, or if they are uncorrelated, then C(τ) =0. If they tend to crowd along the diagonal $X_n = X_{n+\tau}$, then C(τ) > 0, and if they are closer to the line $X_n = - X_{n+\tau}$ we have C(τ) < 0. The latter two cases reflect some tendency of $X_n$ and $X_{n+\tau}$ to be proportional to each other. $C(\tau)$ is maximized when the delay '$\tau$' is zero.

The autocorrelation function is expected to provide a reasonable measure of the transition from redundance to irrelevance as a function of delay. A common choice for delay '$\tau$' is the time at which the autocorrelation function has its first zero, which makes the coordinates linearly uncorrelated.

Figure 7.1(a) shows the autocorrelation function for vowel അ/Λ/. Here $C(\tau)$ crosses first zero when the delay is 4. Figure 7.2(a) gives the corresponding reconstructed phase space for vowel അ/Λ/ with dimension 3.

Figures 7.1(b-e) show the autocorrelation functions for vowels ഇ**/I/,** എ**/ae/,** ഒ/**o**/, ഉ/**u**/ and figures 7.2(b-e) show the corresponding Phase portraits.

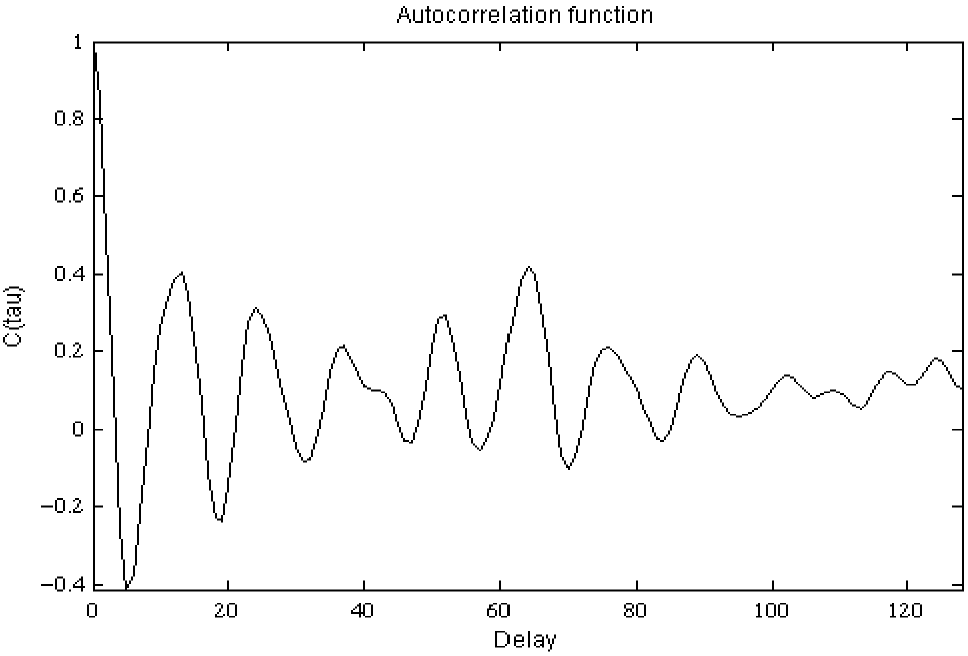


**Fig.7.1(a):** Delay Vs. Autocorrelation function for Malayalam vowel അ/Λ/

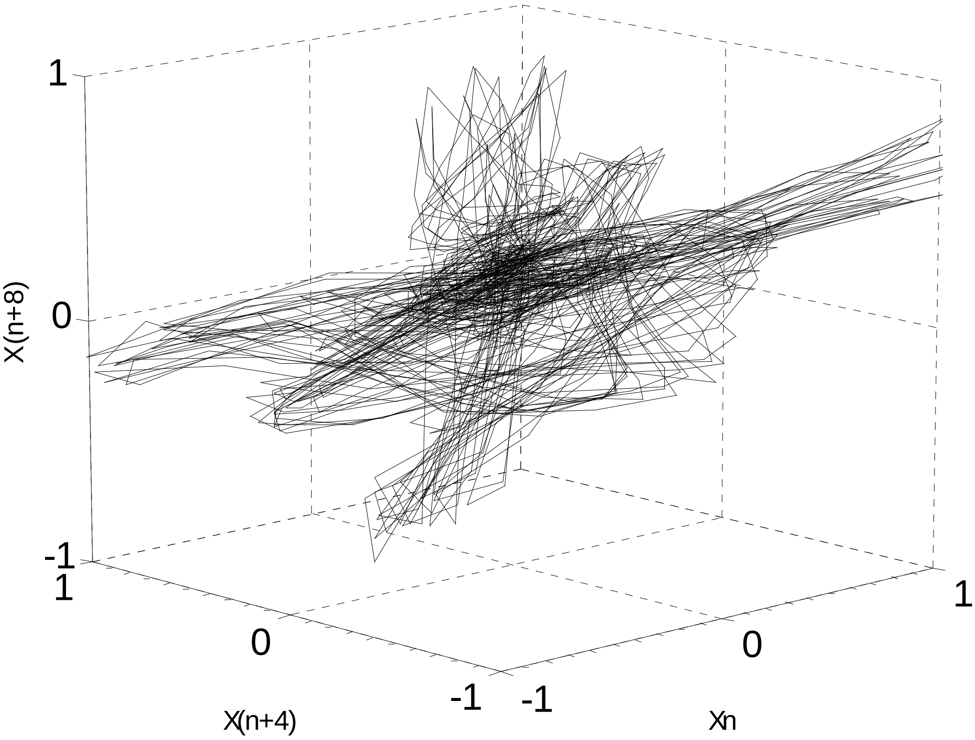


**Fig.7.2(a)**Reconstructed Phase space for vowel അ/Λ/ with delay 4 and dimension 3

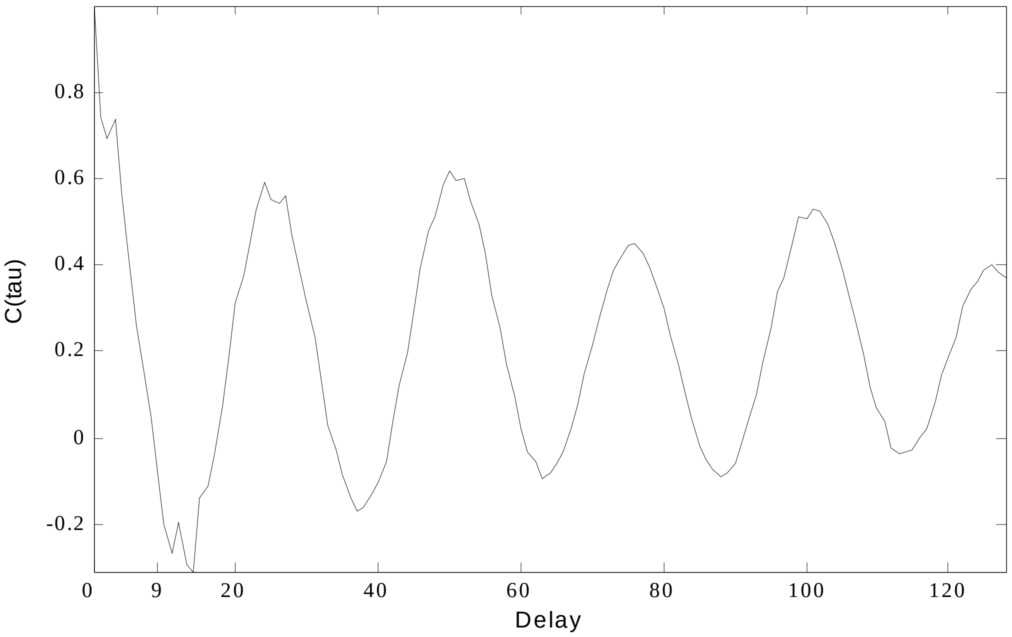


**Fig.7.1(b):** Delay Vs. Autocorrelation function for Malayalam vowel ഈ**/I/**

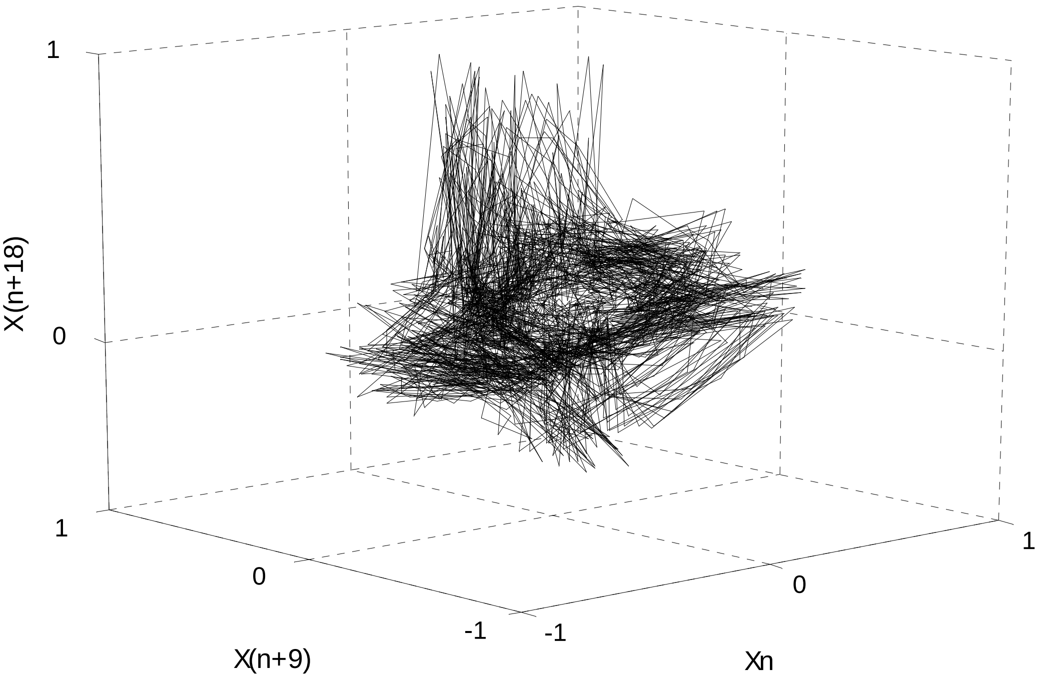


**Fig. 7.2(b)** Reconstructed Phase space for vowel ഈ**/I/**with delay 9 and dimension 3

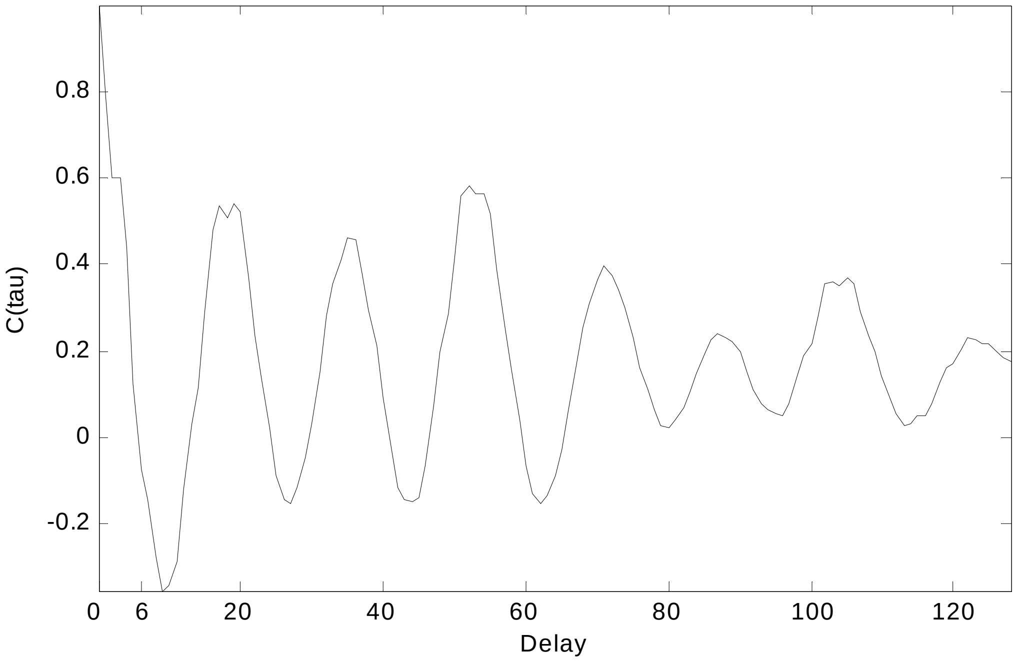


**Fig.7.1(c):** Delay Vs. Autocorrelation function for Malayalam vowel ഏ/**ae/**

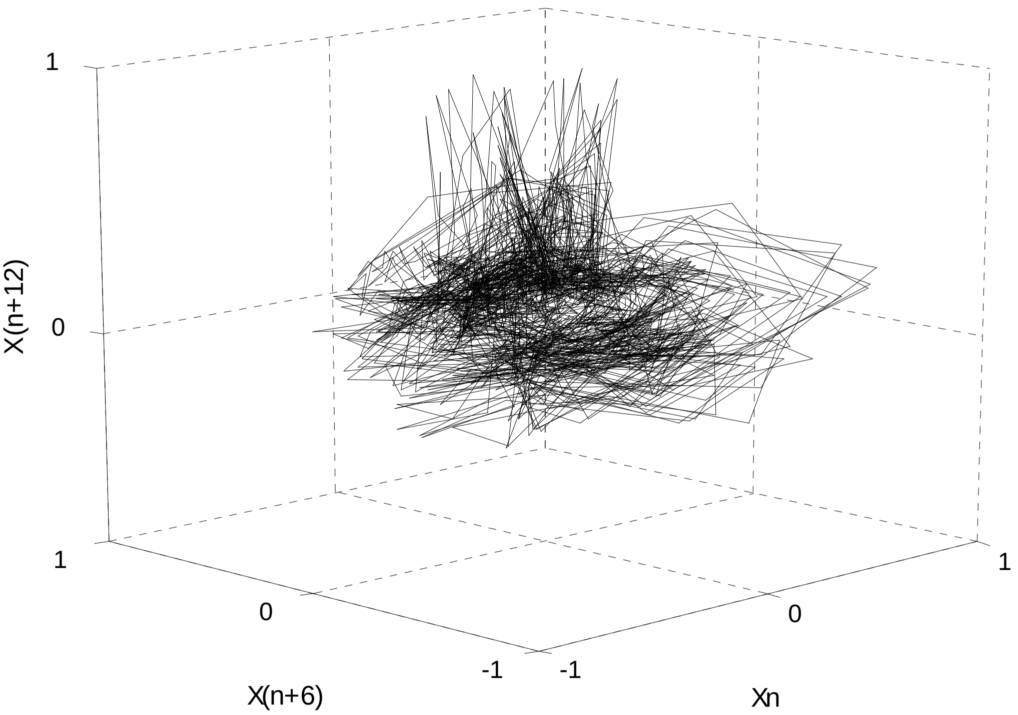


**Fig.7.2(c)**Reconstructed Phase space for vowel ഏ/ae/ with delay 6 and dimension 3

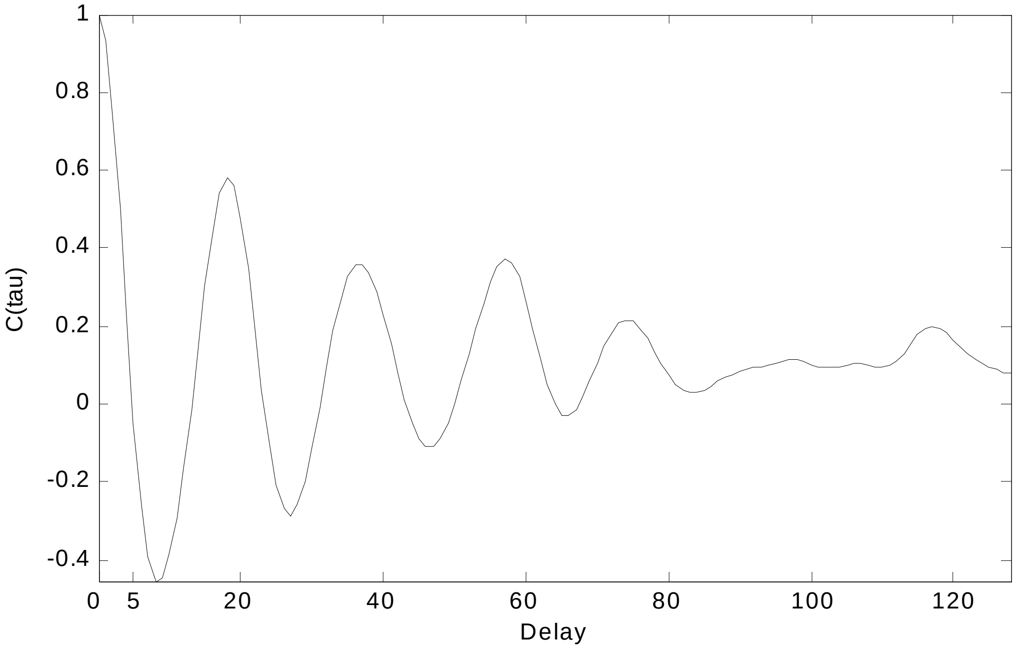


**Fig.7.1(d):** Delay Vs. Autocorrelation function for Malayalam vowel ഓ/**o**/

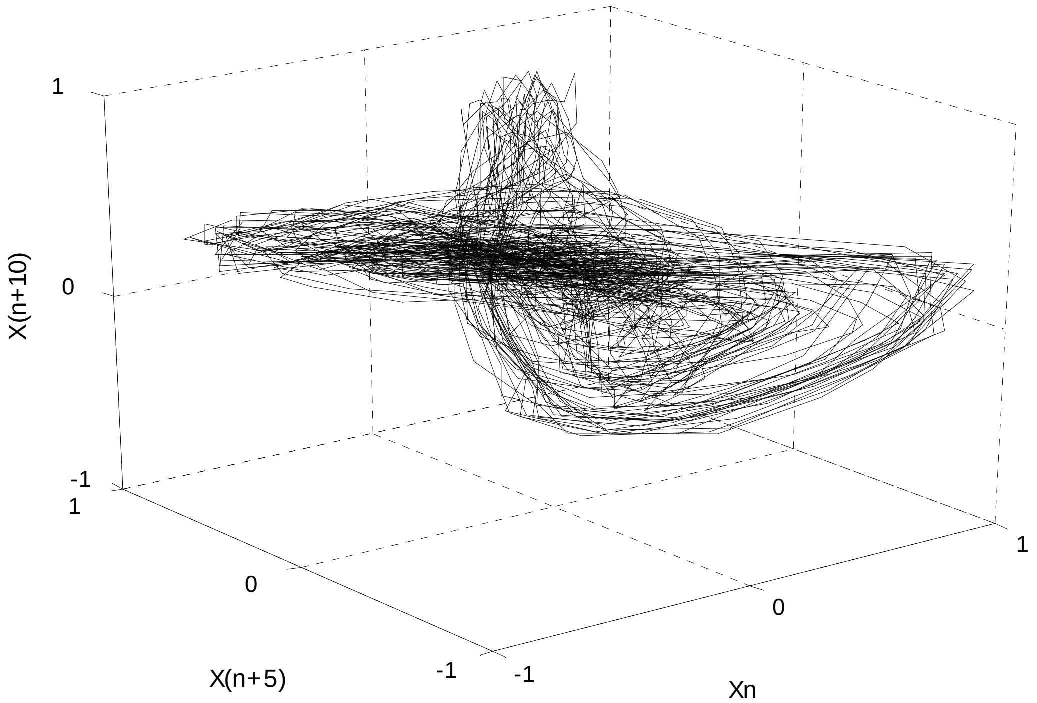


**Fig. 7.2(d)** Reconstructed Phase space for vowel ഓ/**o**/with delay 5 and dimension 3

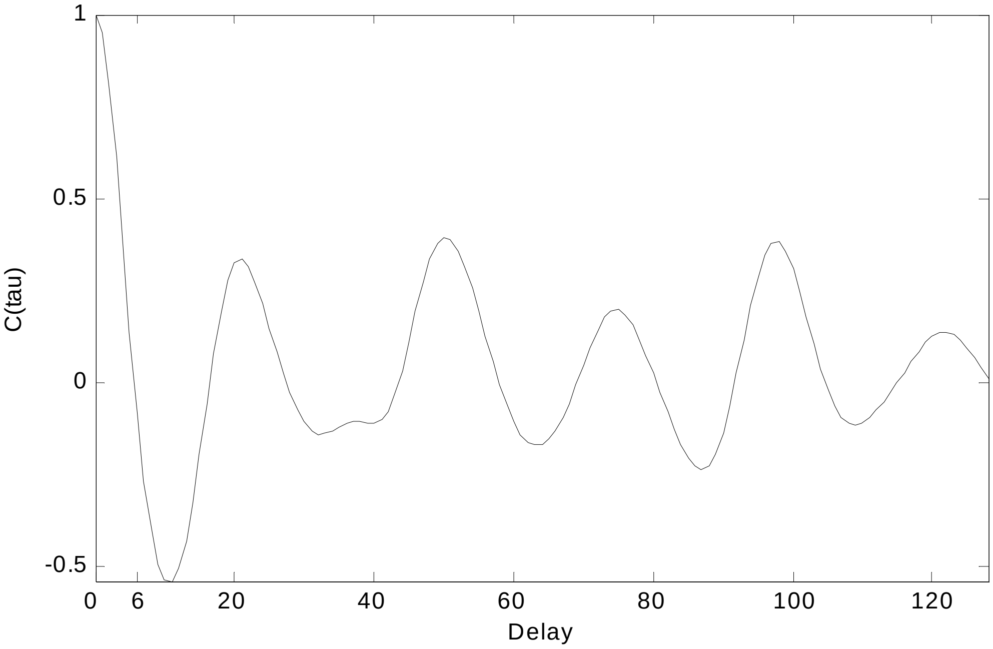


**Fig.7.1(e):** Delay Vs. Autocorrelation function for Malayalam vowel ഉ/**u**/

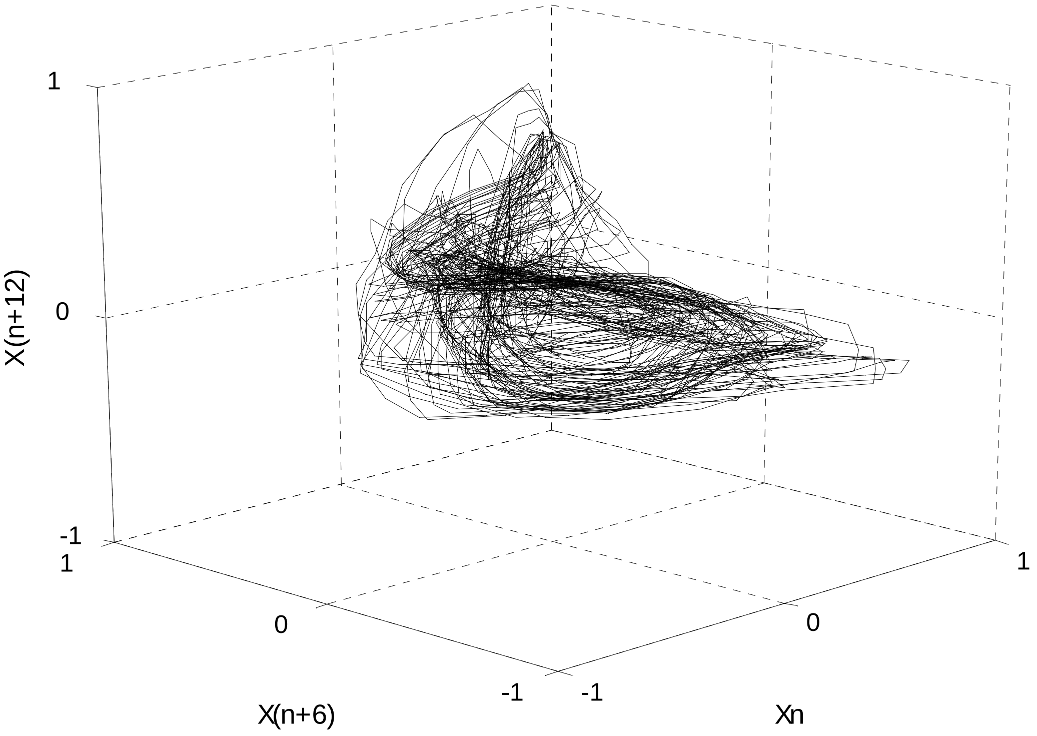


**Fig. 7.2(e)** Reconstructed Phase space for vowel ഉ/**u**/with delay 6 and dimension 3

The autocorrelation based methods have the advantage of short computation times when calculated via the fast Fourier transform (FFT) algorithm. A quite reasonable objection of this procedure is that it is based on linear statistics and only measures linear dependence. It is not taking into account nonlinear dynamical correlations. Since the relationship between the spatial distribution of a reconstructed attractor and the temporal autocorrelation of a single time series is an ill-defined one, the method tends to be inconsistent [Fraser.A.M. and Swinney.H.L, 1986], [Albano.A.M, Rapp.P.E *et. al.,* 1988], [Martinerie.J.M, Rapp.P.E *et. al.*, 1992]. For this reason, Fraser and Swinney [Fraser.A.M. and Swinney.H.L, 1986] suggested a spatial measure based on mutual information.

### 7.3 Proper Time Delay based on Mutual Information Function

In contrast to the *linear* dependence measured by autocorrelation, mutual information, I(τ), supplies a measure of *general* dependence [Fraser.A.M. and Swinney.H.L, 1986]. Therefore, I(τ) is expected to provide a better measure of the shift from redundance to irrelevance with nonlinear systems. Mutual information answers the following question : Given the observation of $x$(t), how accurately can one predict $x$(t + τ)? Thus, successive delay coordinates are interpreted as relatively independent when the mutual information is small. According to Fraser and Swinney greatest independence, or the lowest I(τ), can be associated with the least redundance and, therefore is the best for attractor reconstruction. Hence the proper time delay 'τ' can be selected as the lag that produces a local minimum of I(τ). In a system with positive metric entropy any measurement stripe will eventually spread back into the invariant measure. This is accomplished by the well known stretching and folding effect. To avoid this type of spreading, the first local minimum of I(τ) is preferred to later minima [Fraser.A.M. and Swinney.H.L, 1986], [Liebert.W and Schuster.H.G, 1989].

### 7.3.1 Average Mutual Information

In probability theory and information theory, the mutual information of two random variables is a quantity that measures the mutual dependence of the two variables, or it is the amount of information that is shared between two data sets. Shannon’s information theory provides a formalism for quantifying the concept of information among measurements.

Consider a set of possible measurements $s_1$, $s_2$, ……… $s_n$ of a system ‘S’. Let $P_s(s_1)$, $P_s(s_2)$, ………..$P_s(s_n)$ be the associated probabilities of the measurements. Then $P_s$ maps the measurements to probabilities. The amount of information gained from a measurement that specifies s is the entropy ‘H’ of the system,

$$H(S) = -\sum_i P_s(s_i) \log P_s(s_i) \qquad \text{……….. ( 7.2)}$$

If the log is taken to the base two, H is in units of bits.

Now we consider a general coupled system ‘S’ and ‘Q’. We have given that ‘s’ has been measured and found to be $s_i$, then the uncertainty in a measurement of ‘q’ is given by,

$$H(Q|s_i) = -\sum_j P_{q|s}(q_j|s_i)\log[P_{q|s}(q_j|s_i)] \qquad \text{……(7.3)}$$

where $P_{q|s}(q_j|s_i)$ is the probability that a measurement of q will yield q$_j$, given that the measured value of s is s$_i$. But the probability associated with the combined system S and Q, that a measurement of q will yield q$_j$, and the measurement of s yield s$_i$, $P_{sq}(s_i,q_j)$ is given by :

$$P_{sq}(s_i,q_j) = P_{q|s}(q_j|s_i) \,.\, P_s(s_i)$$

then, $$H(Q|s_i) = -\sum_j [P_{sq}(s_i,q_j)/P_s(s_i)]\log[P_{sq}(s_i,q_j)/P_s(s_i)] \qquad \text{….(7.4)}$$

Then for a time series x(t), we can measure how dependent the values of x(t+τ) on the values of x(t), by making the assignment [s,q] = [x(t), x(t+τ)]. Given that x has been measured at time t, then the average uncertainty in a measurement of x at time t+τ is given by averaging $H(Q|s_i)$ over s$_i$, which yields

$$H(Q|S) = \sum_i P_s(s_i)H(Q|s_i) \qquad \text{………..(7.5)}$$

$$= -\sum_{i,j} P_{sq}(s_i,q_j)\log[P_{sq}(s_i,q_j)/P_s(s_i)]$$

$$= -\sum_{i,j} P_{sq}(s_i,q_j)\{\log[P_{sq}(s_i,q_j)] - \log[P_s(s_i)]\}$$

$$= -\sum_{i,j} P_{sq}(s_i,q_j)\log[P_{sq}(s_i,q_j)] + \sum_{i,j} P_{sq}(s_i,q_j)\log[P_s(s_i)]$$

But $-\sum_{i,j} P_{sq}(s_i, q_j)\log[P_{sq}(s_i, q_j)] = H(S,Q)$ …….(7.6)

and $\sum_{i,j} P_{sq}(s_i, q_j)\log\left[P_s(s_i)\right] = \sum_{i,j} P_{q|s}(q_j|s_i).P_s(s_i)\log\left[P_s(s_i)\right]$

$= \sum_{i} P_s(s_i)\log\left[P_s(s_i)\right]$ since, $\sum_{j} P_{q|s}(q_j|s_i) = 1$

so $\sum_{i,j} P_{sq}(s_i, q_j)\log\left[P_s(s_i)\right] = - H(S)$ ……..(7.7)

Hence $H(Q|S) = H(S,Q) - H(S)$

H(Q) is the uncertainty of q in isolation, and H(Q | S) is the uncertainty of q given a measurement of s. So the amount that a measurement of s reduces the uncertainty of q is

$$I(Q,S) = H(Q) - H(Q|S)$$

$$= H(Q) + H(S) - H(S,Q)$$

This is the mutual information. It is the answer to the question, ‘Given a measurement of s, how many bits on the average can be predicted about q?’

From equation 7.6, we have $H(S,Q) = -\sum_{i,j} P_{sq}(s_i, q_j)\log[P_{sq}(s_i, q_j)]$

and from equation 7.7, $H(S) = -\sum_{i,j} P_{sq}(s_i, q_j)\log\left[P_s(s_i)\right]$

similarly, $H(Q) = -\sum_{i,j} P_{sq}(s_i, q_j)\log\left[P_q(q_j)\right]$

Hence,

I(Q,S) = $\sum_{i,j} P_{sq}(s_i, q_j)\log[P_{sq}(s_i, q_j)] - \sum_{i,j} P_{sq}(s_i, q_j)\log\left[P_s(s_i)\right] - \sum_{i,j} P_{sq}(s_i, q_j)\log\left[P_q(q_j)\right]$

$$I(Q,S) = \sum_{i,j} P_{sq}(s_i, q_j)\log[P_{sq}(s_i, q_j) / P_s(s_i)P_q(q_j)]$$

If the measurement of a value from Q resulting in $q_j$ is completely independent of the measurement of a value from S resulting in $s_i$, then $P_{sq}(s_i,q_j)$ factorizes :

$P_{sq}(s_i,q_j) = P_s(s_i)\ P_q(q_j)$ and the amount of information between the measurements, average mutual information is zero.

For time series x(t), the average mutual information between x(t) and x(t+τ) is given by

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

If the measurements x(t) and x(t+τ) are independent, then I(τ) will tend to zero. It was the suggestion of Fraser that one can use the function I(τ) as a kind of nonlinear autocorrelation function to determine, when the values x(t) and x(t+τ) are independent enough of each other to be useful as coordinates in a time delay vector but not so independent as to have no connection with each other at all. The actual prescription suggested is to take the 'τ' where the first minimum of the average mutual information I(τ) occurs as that value to use in time delay reconstruction of phase space.

The algorithm suggested by Fraser and Swinney is implemented using Mathlab and the proper delay for Malayalam vowels are calculated.

Figures 7.3(a-e) show the average mutual information function for Malayalam vowels and figures 7.4(a-e) give the corresponding reconstructed phase spaces with dimension 3. In figure 7.3(a) first minimum of I(τ) is when delay is 3

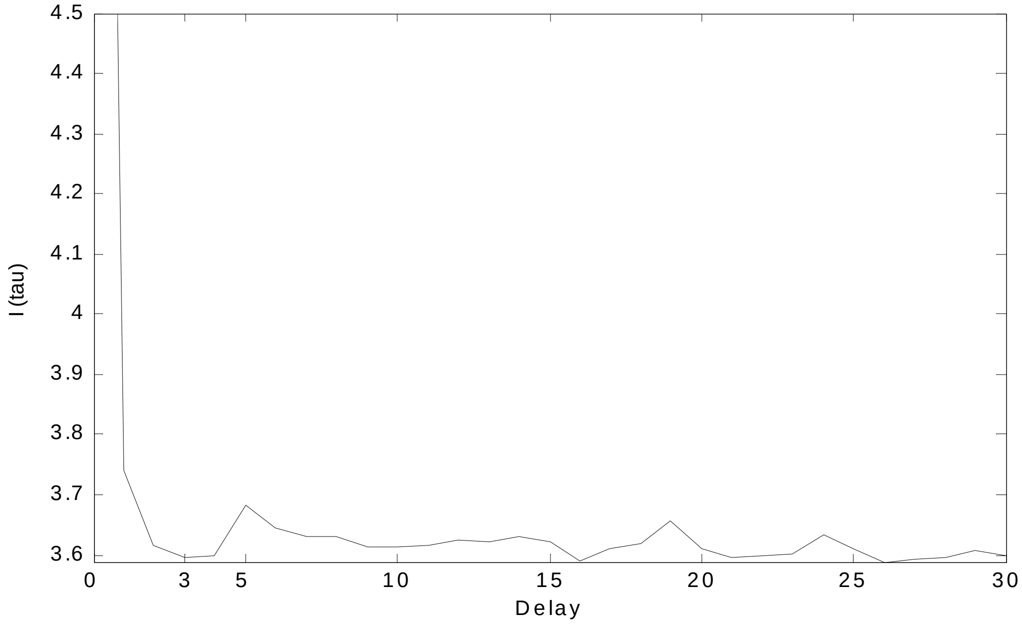


**Fig.7.3(a):** Delay Vs. Average Mutual Information function (vowel അ/Λ/)

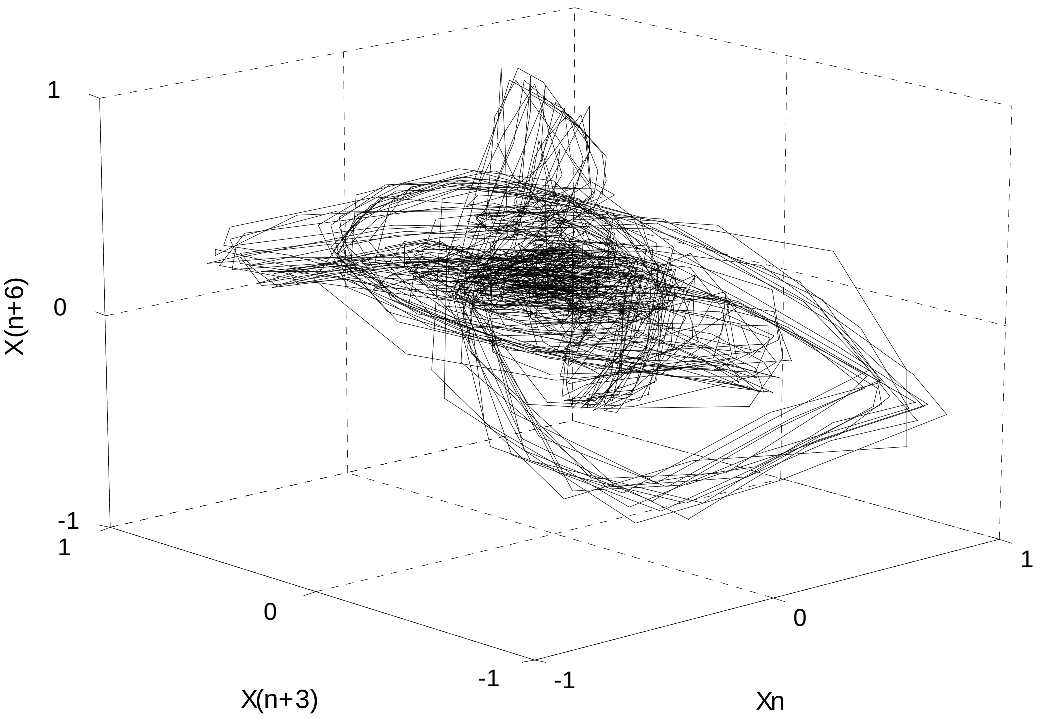


**Fig7.4(a):**Reconstructed Phase space for vowel അ/Λ/with delay 3 and dimension 3

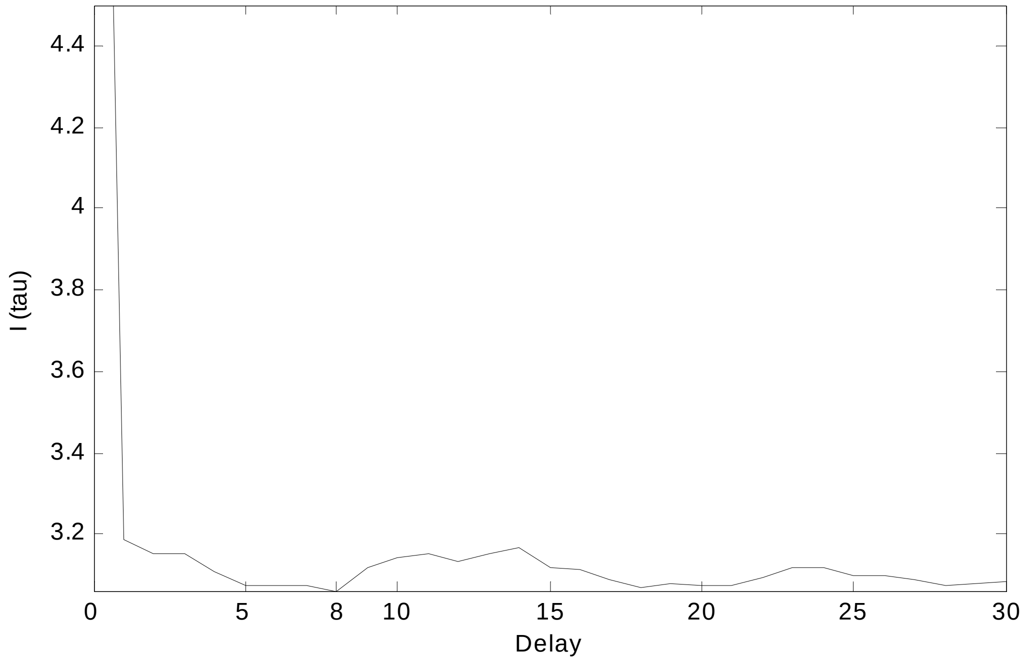


**Fig.7.3(b):** Delay Vs. Average Mutual Information function (vowel ഇ/I/)

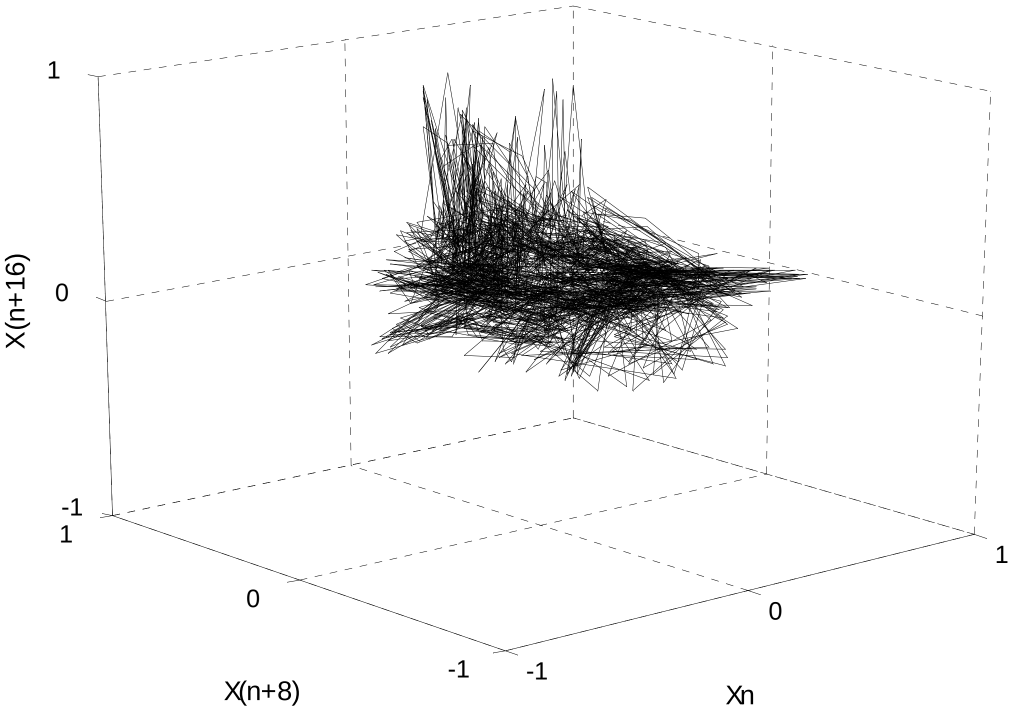


**Fig.7.4(b):** Reconstructed Phase space for vowel ഇ/I/with delay 8 and dimension 3

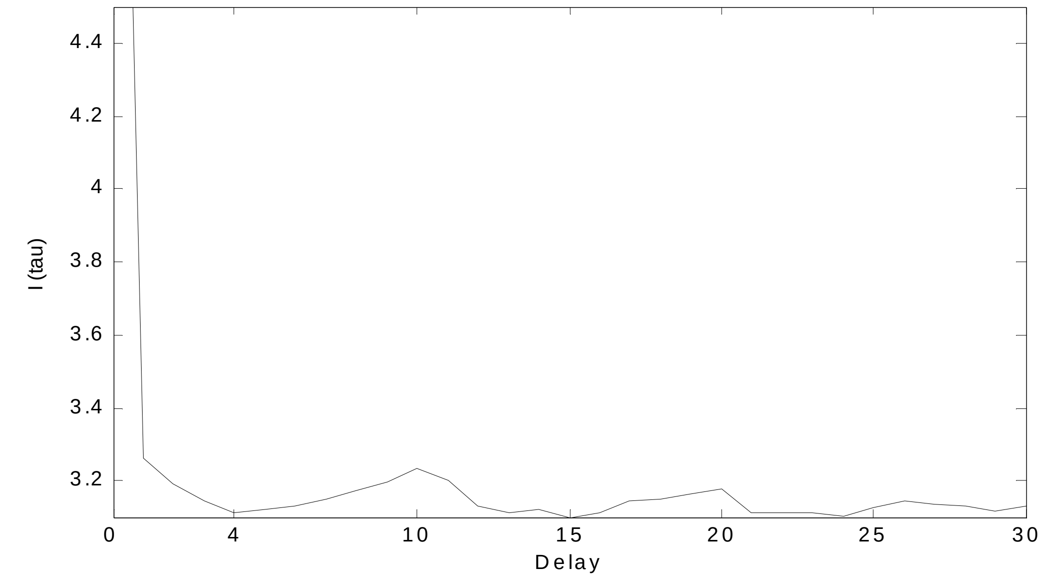


**Fig.7.3(c):** Delay Vs. Average Mutual Information function (vowel ഏ/ae/)

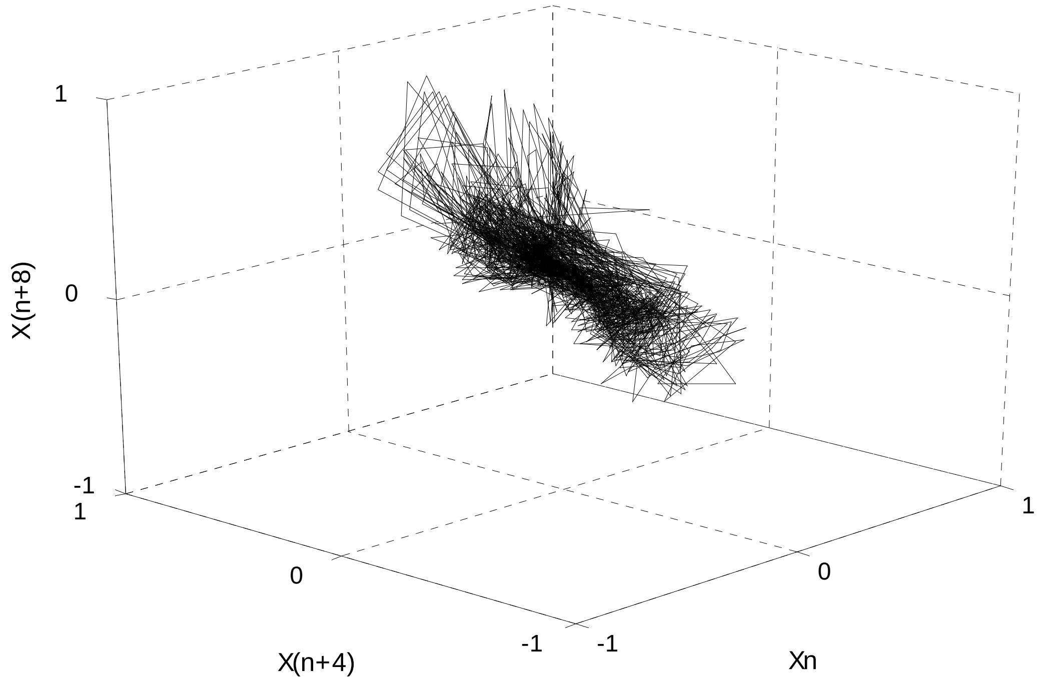


**Fig7.4(c):**Reconstructed Phasespace for vowel ഏ/ae/with delay 4 and dimension 3

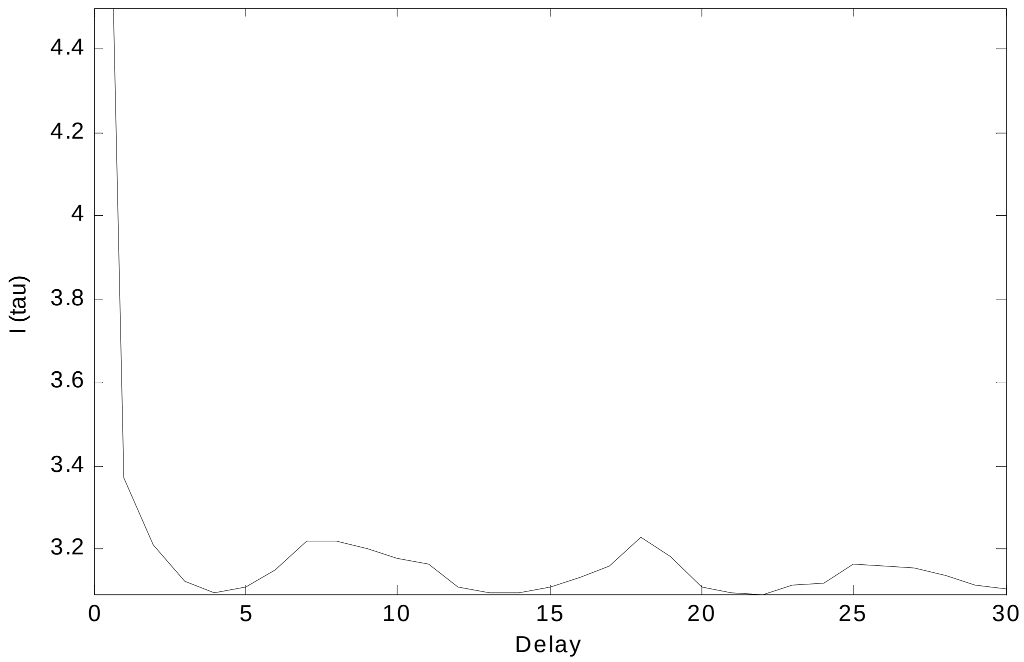


**Fig.7.3(d):** Delay Vs. Average Mutual Information function (vowel ഓ/o/)

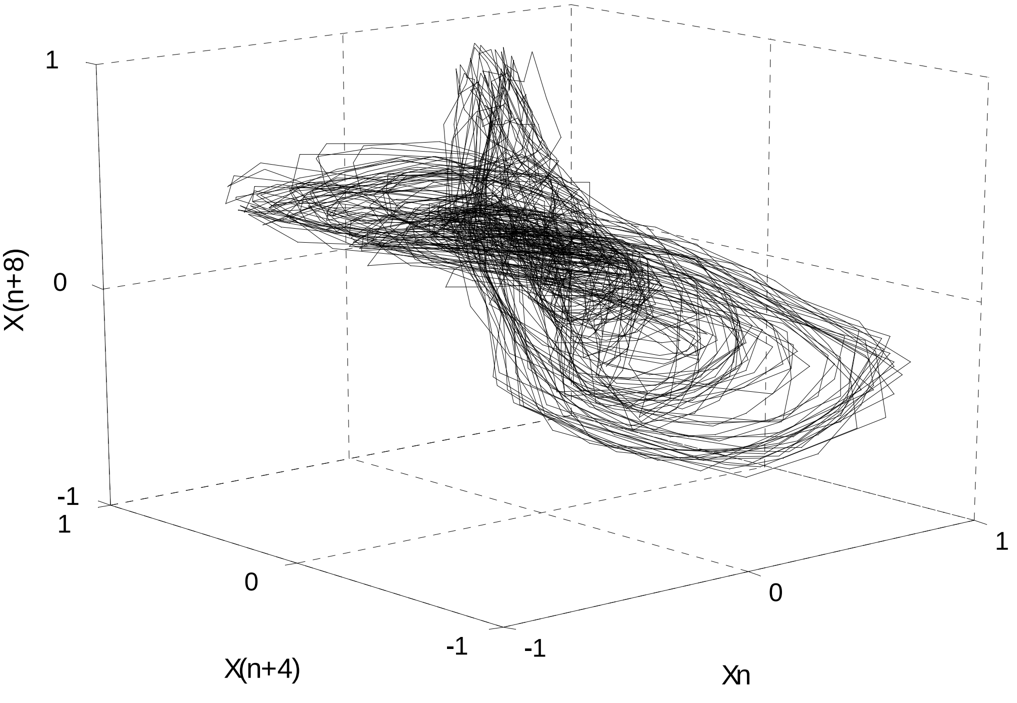


**Fig.7.4(d):** Reconstructed Phase space for vowel ഓ/o/with delay 4 and dimension 3

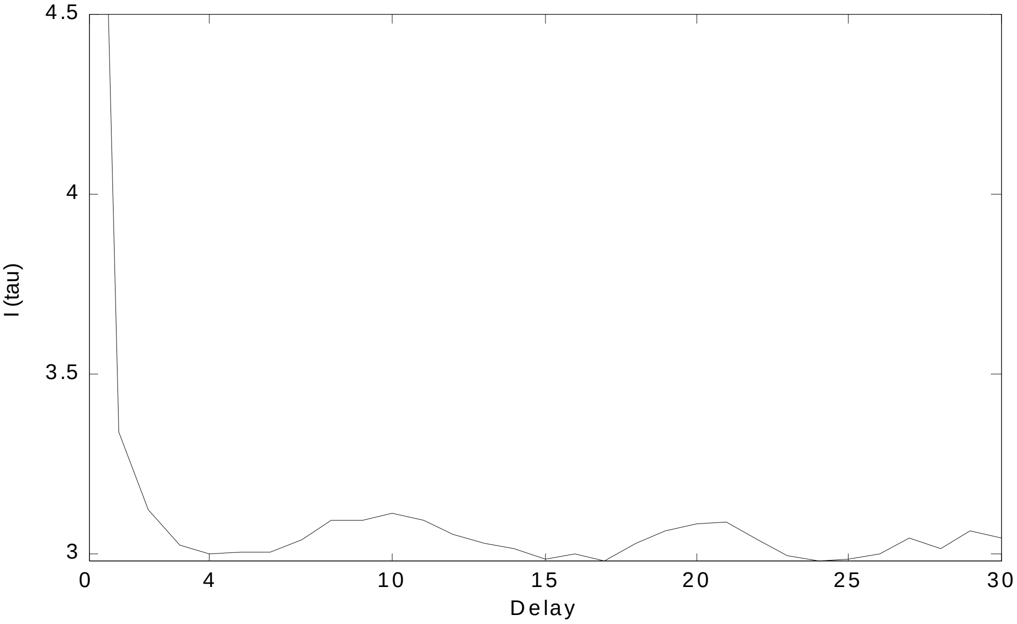


**Fig.7.3(e):** Delay Vs. Average Mutual Information function (vowel ഉ/u/)

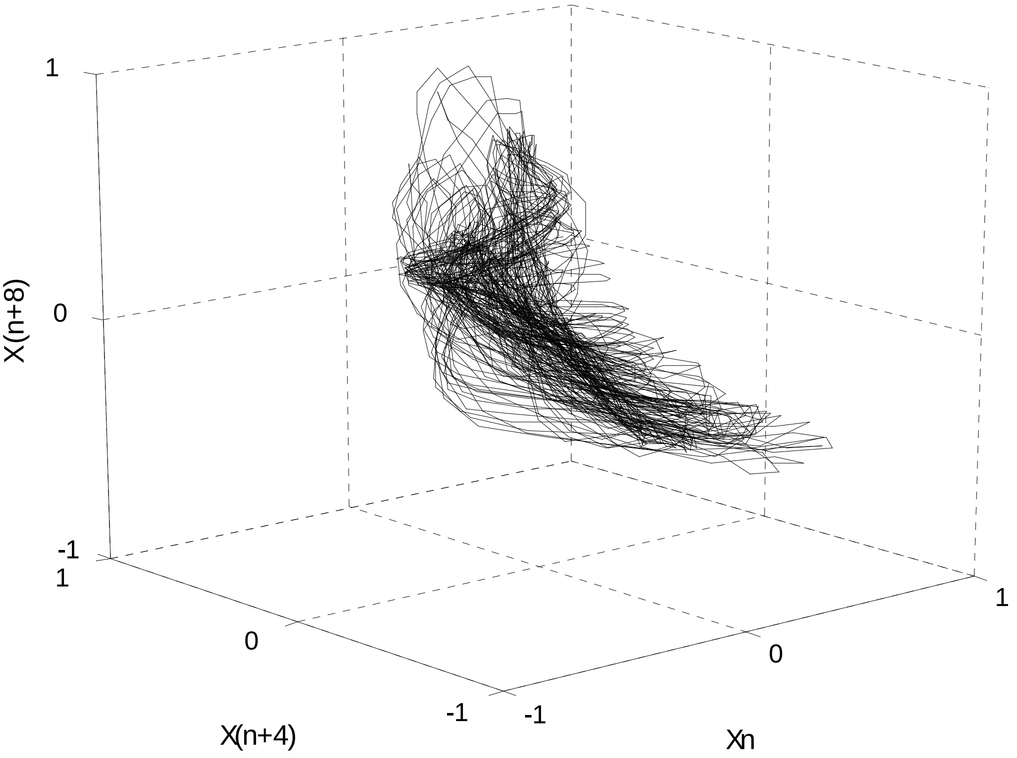


**Fig.7.4(e):** Reconstructed Phase space for vowel ഉ/u/with delay 4 and dimension 3

Average mutual information I(τ) reveals quite a different insight about the nonlinear characteristics of an observed time series than the more familiar autocorrelation function, where the latter is tied to linear properties of the source. The primary drawback of this approach is the enormous computational costs. Martinerie *et. al.* showed that mutual information is also inconsistent in identifying the optimal value of 'τ' [Martinerie.J.M *et. al.*, 1992].

The literature contains many more suggestions on methods of how to determine an optimal time lag. Some of them have a nice heuristic justification. For a nonlinear time series analysis, it is our impression that it might be more useful to optimize the time lag with respect to a particular source and application. Hence for speech signal analysis, we introduced a novel method for determining the proper time lag and is explained in the next section.

### 7.4 Geometry based method for Proper Time Delay

Geometry-based methods for determining τ may be interpreted as various attempts to answer the following question: What value for the delay results in the most space filling reconstruction? Or for what value of delay the attractor in the reconstructed phase space is most dilated? In the previous chapter we have utilized the reconstructed phase space for the determination of pitch period of the speech signal. The idea behind this approach is that, if the time delay corresponds to pitch period of the signal (T), the Euclidean distance measure of phase space points from the phase space diagonal (the identity line) is the minimum. If the Euclidean Distance Measure from the Phase Space Diagonal is maximum, the Phase Space points are widely spread in the reconstructed space. The delay corresponding to the Maximum Euclidean Distance Measure can be interpreted as the delay, which results in the most space filling reconstruction.

The expansion from the main diagonal can be best quantified by measuring the Euclidean distance of the phase space points, as explained in the previous chapter. The delay corresponds to Maximum Euclidian Distance Measure (MEDM) can be used for the reconstruction of phase space, as it results in most space filling reconstruction and provides most dilated attractor in the Phase Space.

Figures 7.5(a-e) show the variation of Euclidean Distance Measure of Phase Space points from the Phase Space diagonal with delay for Malayalam vowels and figures 7.6(a-e) give the corresponding reconstructed phase spaces with dimension 3. In figure 7.5(a), Maximum Distance Measure from the phase space diagonal corresponds to time delay 6.

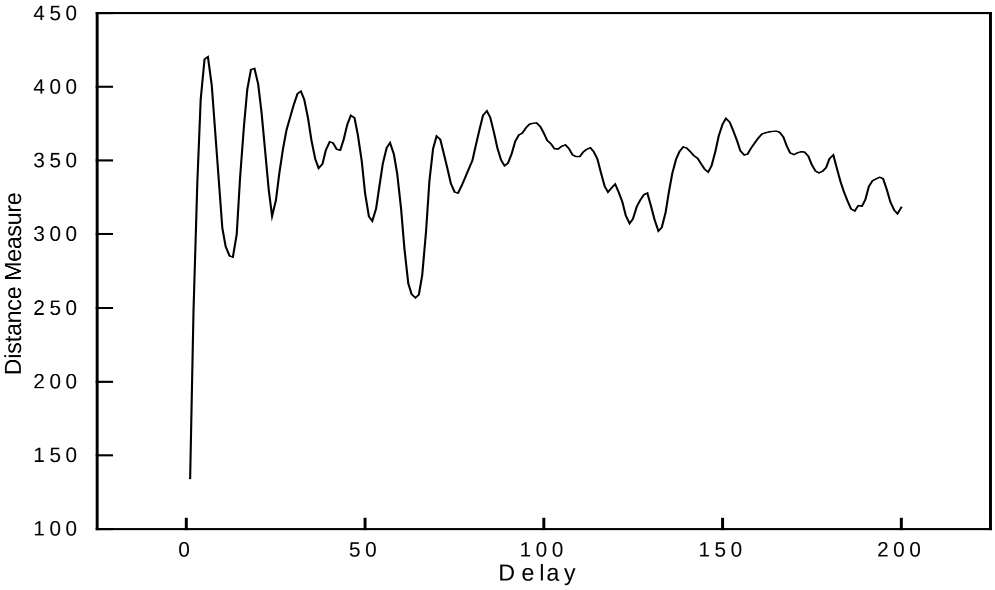


**Fig.7.5(a):**Delay Vs. Distance Measure from the phase space diagonal (vowel അ/Λ/)

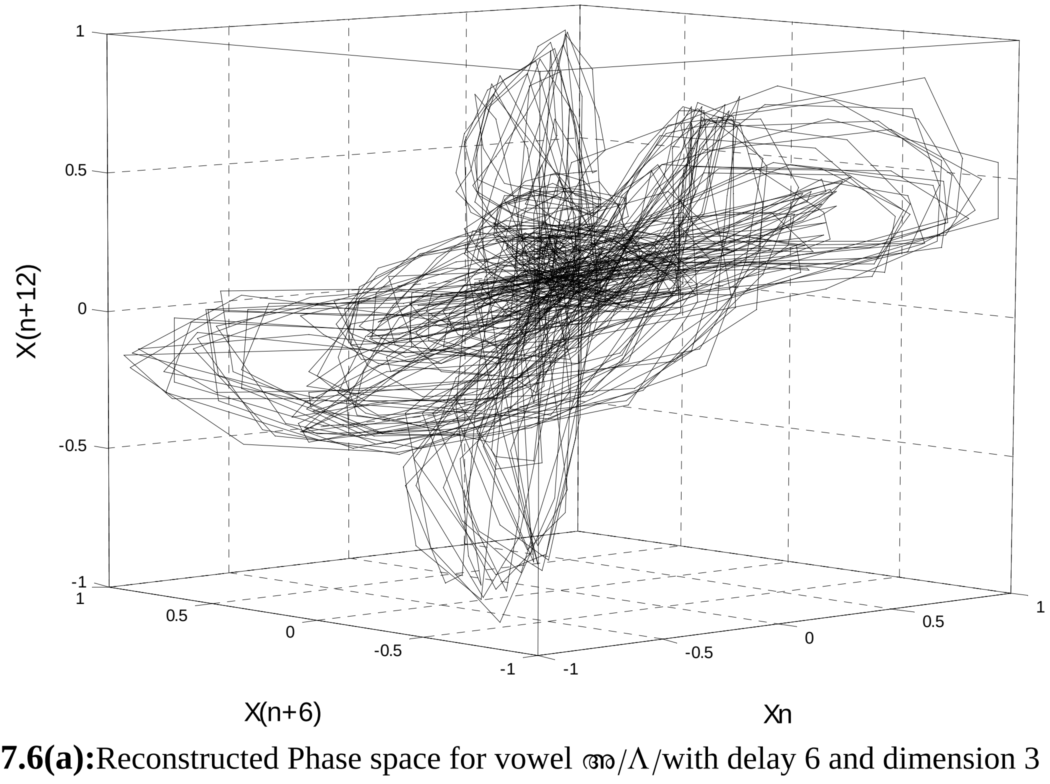


**Fig7.6(a):**Reconstructed Phase space for vowel അ/Λ/with delay 6 and dimension 3

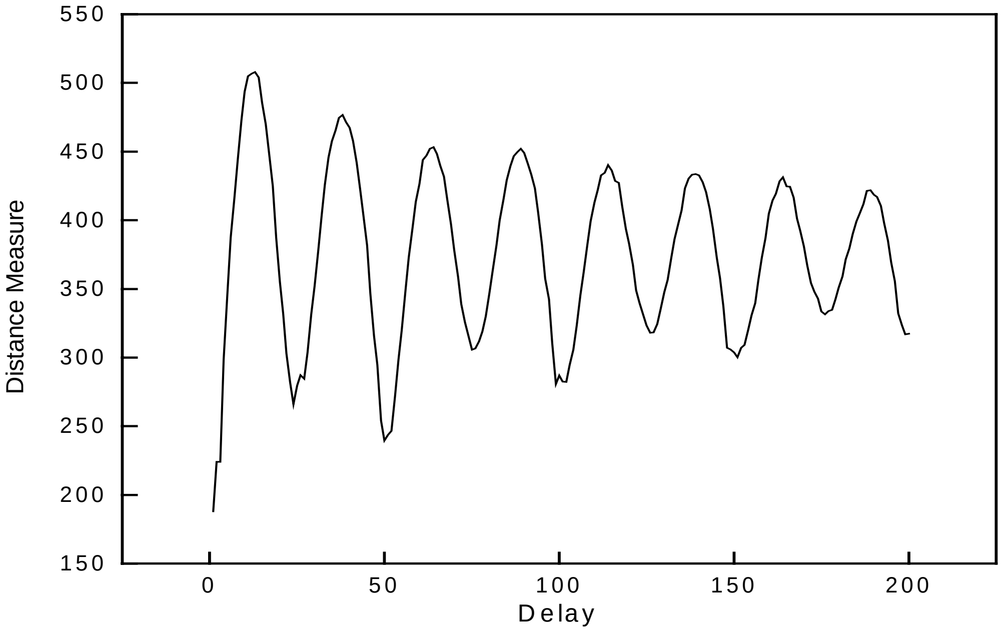


**Fig.7.5(b):**Delay Vs.Distance Measure from the phase space diagonal (vowel ഇ/I/)

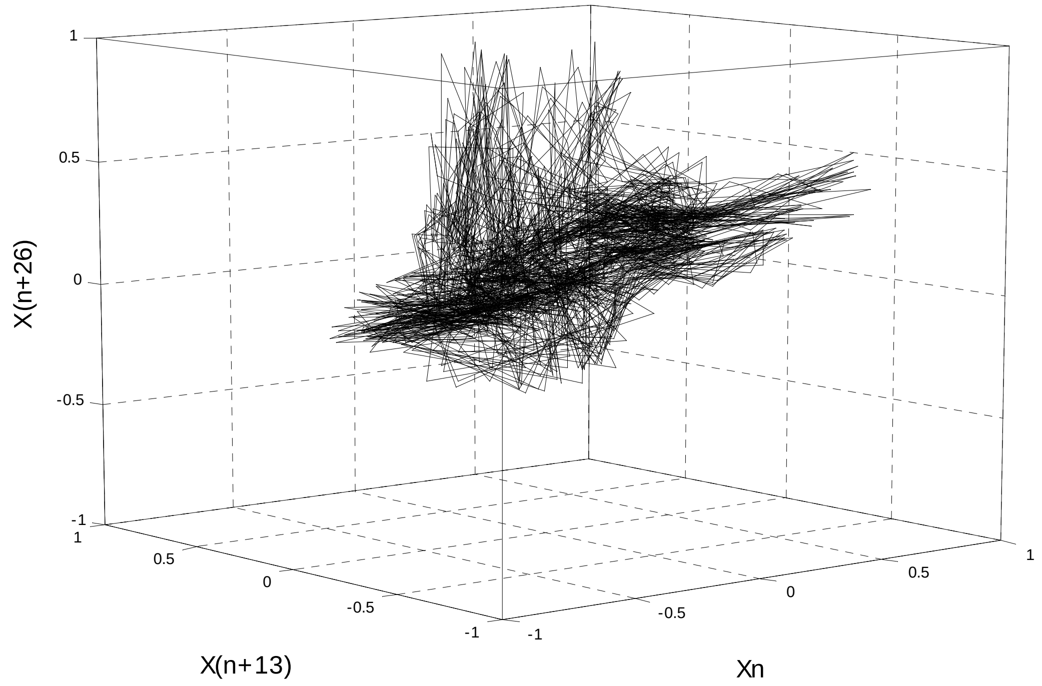


**Fig7.6(b):**Reconstructed Phase space for vowel ഇ/I/with delay 13 and dimension 3

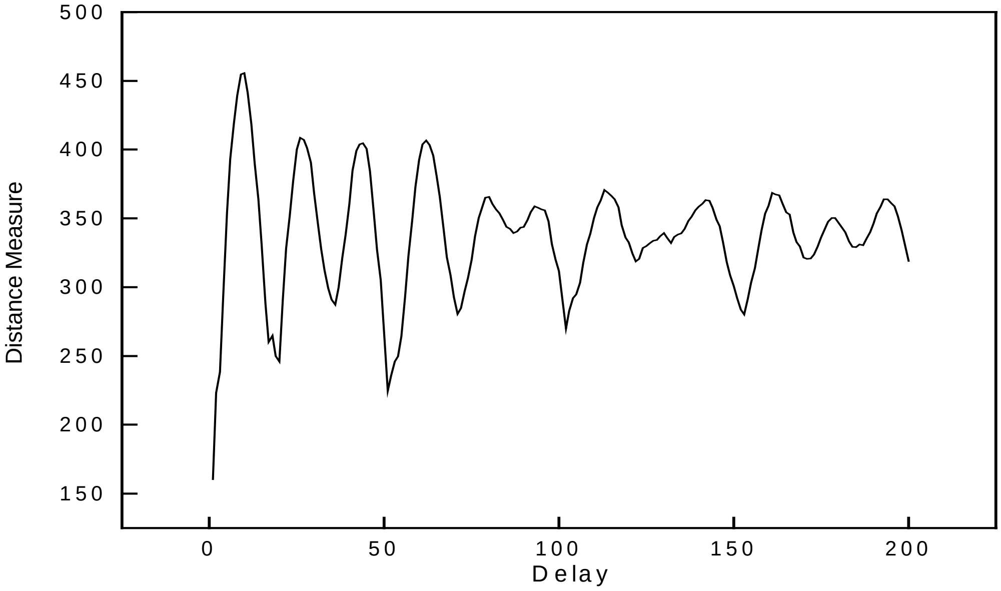


**Fig.7.5(c):**Delay Vs.Distance Measure from the phase space diagonal (vowel ഐ/ae/)

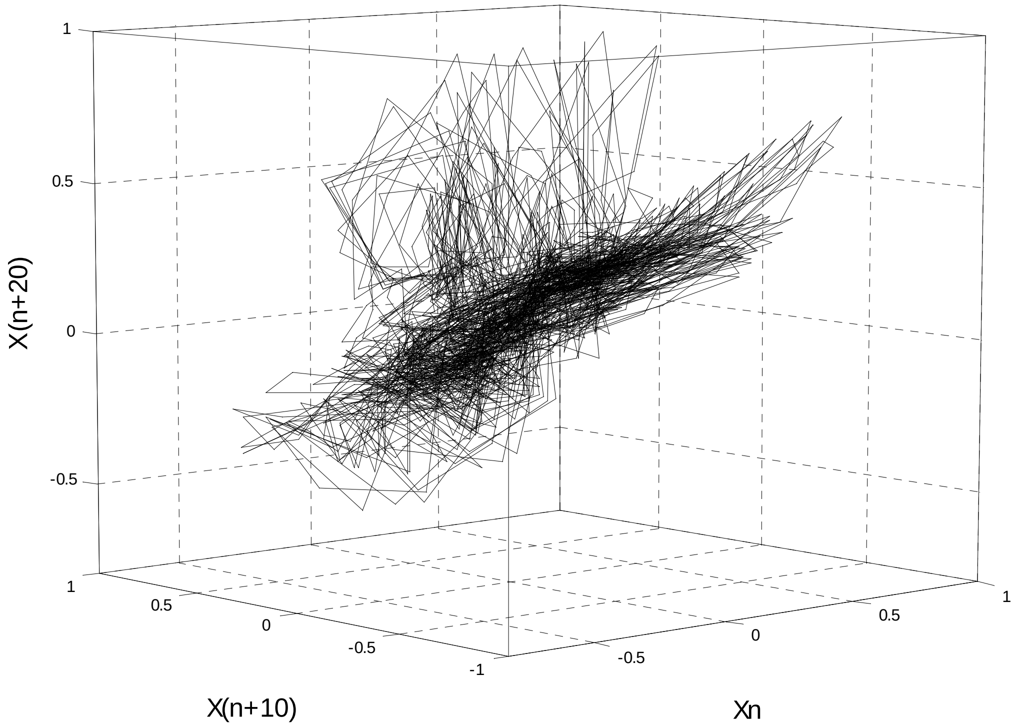


**Fig7.6(c):**Reconstructed Phase space for ഐ/ae/with delay 10 and dimension 3

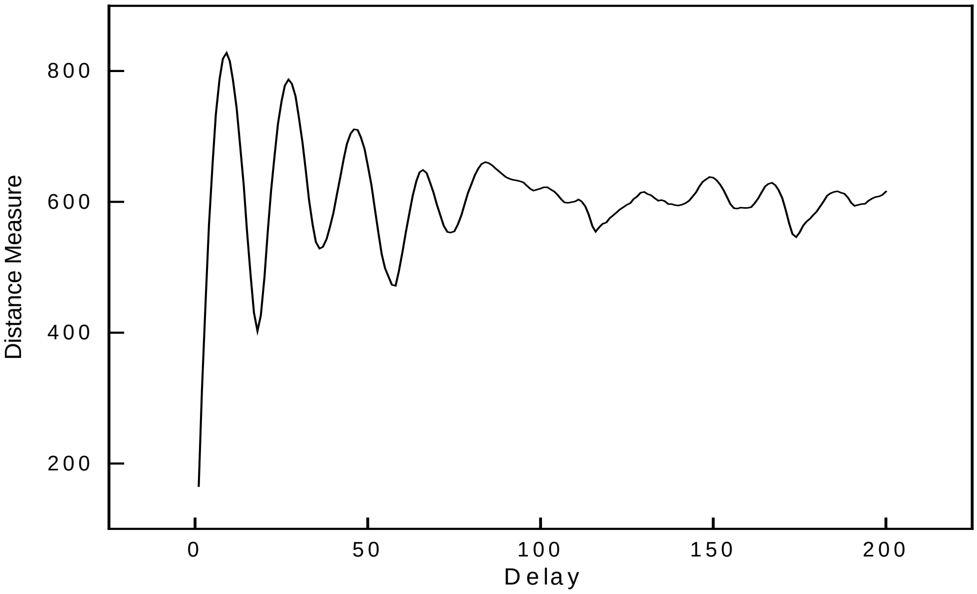


**Fig.7.5(d):**Delay Vs.Distance Measure from the phase space diagonal (vowel ഒ/o/)

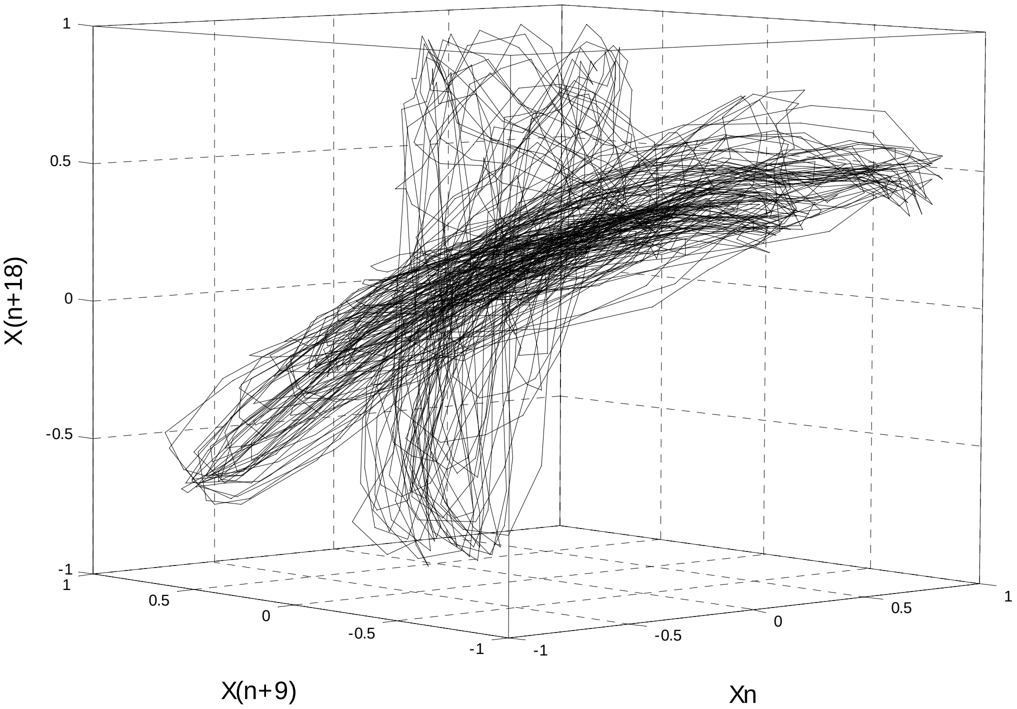


**Fig.7.6(d):** Reconstructed Phase space for vowel ഒ/o/with delay 9 and dimension 3

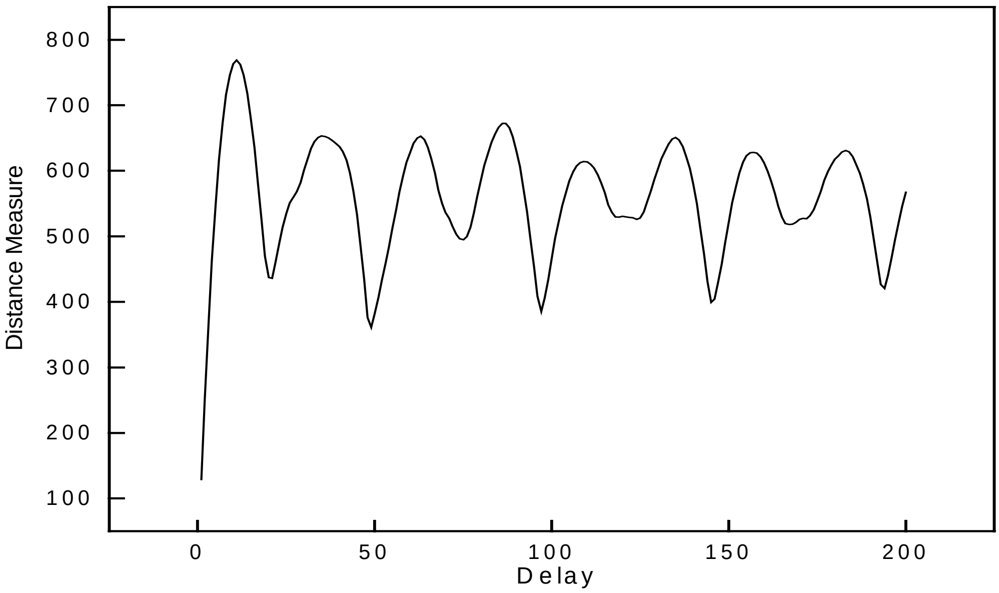


**Fig.7.5(e):**Delay Vs.Distance Measure from the phase space diagonal (vowel ഉ/u/)

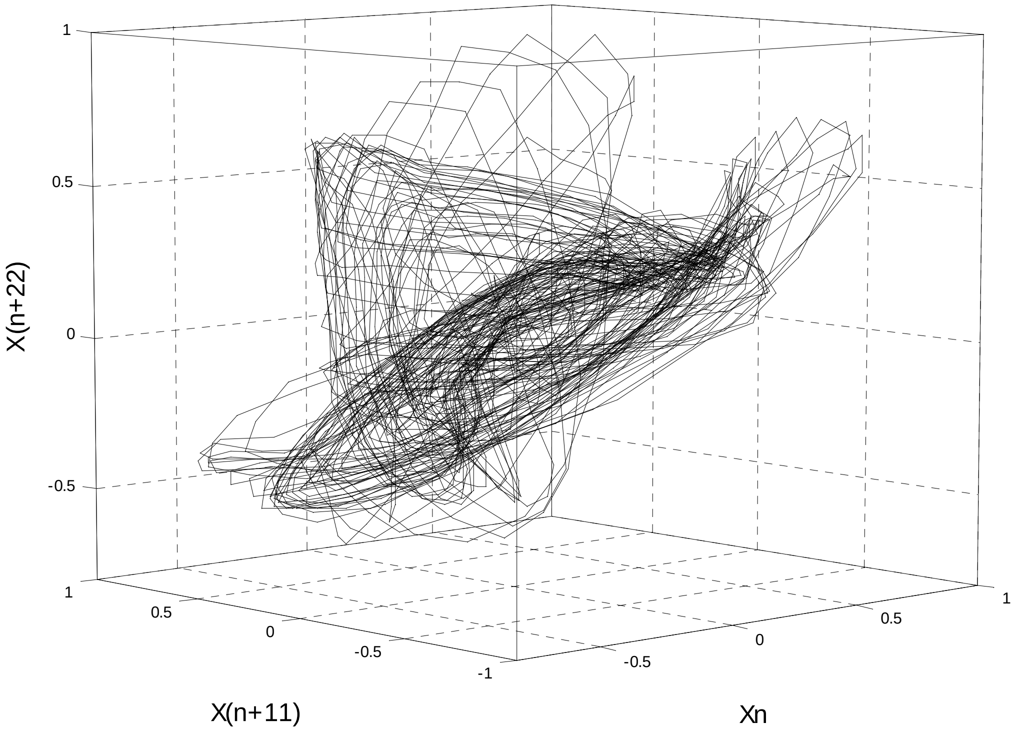


**Fig.7.6(e):**Reconstructed Phase space for vowel ഉ/u/ with delay 11and dimension 3

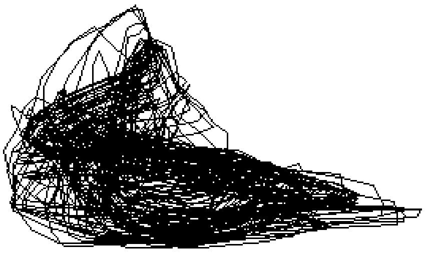


**Fig.7.7 :** Phase portrait of Malayalam vowel ഉ/u/ with first zero of Autocorrelation function as time delay

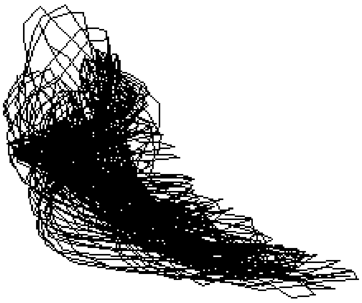


**Fig. 7.8 :** Phase portrait of Malayalam vowel ഉ/u/ with first minimum of Mutual Information function as time delay

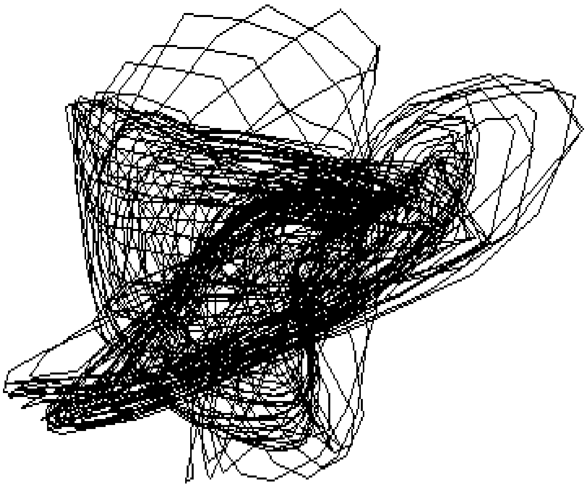


**Fig. 7.9 :** Phase portrait of Malayalam vowel ഉ/u/ with maximum distance measure from the phase space diagonal as time delay

The criteria chosen for the selection time delay in the reconstruction of phase space are compared in figures 7.7, 7.8 and 7.9. The phase portrait shown in Figure 7.7 is constructed for time delay corresponding to the first zero in the autocorrelation function and it occurs for delay 6. In Figure 7.8 the phase portrait corresponds to first minimum of mutual information function and here the delay is 4. In these figures it is clear that the trajectories are indistinguishable and the attractor is not dilated much. Therefore it is difficult to deduce quantitative information about the dynamics from these phase portraits. On the other hand in the phase portrait shown in figure 7.9 the criteria used for the selection of time delay is the delay corresponds to maximum distance measure from the main diagonal, and the delay selected is 11. Since it is the most space filling reconstruction compared to the other two methods, in applications like speech recognition, parameters can be quantitatively deduced from this portrait. Phase space distribution parameter extracted from these portraits can be effectively utilized for the improvement of recognition accuracy as explained in chapter 8

## 7.5 Minimum Embedding dimensions for Phase Space Reconstruction

The embedding theorem tells us that if the dimension of the attractor defined by the orbits is D, then we will certainly unfold the attractor in an integer dimensional space of dimension d, where d ≥ 2D. This is not the necessary dimension for unfolding, but is sufficient and certainly tells us when to stop adding components to the time delay vector. The box counting dimension of the strange attractor for the Lorenz model is D ≈ 2.06, which would lead us to anticipate d = 5 to unfold the Lorenz attractor. But it is shown that d = 3 will do well for this system [Abarbanel.H.D.I, 1996]. Therefore we are certain from the theorem that some dimension ≤ 2D will do for d, and in the following session we will discuss which dimension is to be selected.

There have been many discussions in the literature on selecting optimum embedding dimension from a time series data [Martinerie.J.M., AlbanoA.M..*et al,* 1992], [Kennel. M. B., Brown. R., and Abarbanel. H. D. I,1992]. Following are the three basic methods which are usually used to choose the minimum embedding dimension:

(1) computing some invariant on the attractor[ Grassberger.P and Procaccia.I, 1983]. By increasing the embedding dimension used for the computation one notes when the value of the invariant stop changing. The typical problem with this approach is that it is often very data intensive, certainly subjective, and time-consuming for computation. (2) singular value decomposition and (3) the method of false neighbors [Kennel.M.B, Brown.R, *et.al.*, 1992]. It was developed based on the fact that choosing too low an embedding dimension results in points that are far apart in the original phase space being moved closer together in the reconstruction space.

### 7.5.1 The method of False Nearest Neighbors

The concept called false nearest neighbors, was introduced by Kennel, Brown & Abarbanel (1992). The basic idea is to search for points in the data set which are neighbors in embedding space. Imagine that the correct embedding dimension for some data set is d. Now study the same data in a lower embedding dimension, $d_0$ (ie $d_0 < d$). The transition from d to $d_0$ is a projection, eliminating certain axes from the coordinate system. Hence points whose coordinates are eliminated by the projection can become 'false neighbors' in the $d_0$ dimensional space.

If $d_0$ is qualified as an embedding dimension by the embedding theorems, then any two points which stay close in the $d_0$ dimensional reconstructed space will be still close in the $d_0$ +1 dimensional reconstructed space. Such pair of points are called true neighbors, otherwise they are called false neighbors. Perfect embedding means that no false neighbors exist. This is the idea of the false neighbor method by the authors Kennel, Brown & Abarbanel [Kennel.M.B, Brown.R, *et.al.*, 1992].

For each point of the time series, take its closest neighbor in $d_0$ dimensions. Then compute the ratio of distances between these two points in $d_0$+1 dimensions and $d_0$ dimensions. If this ratio is larger than a threshold 'r', the neighbor was false. The process is repeated by increasing the dimension and the percentage of false nearest neighbors will drop from 100% to zero when the proper dimension 'd' is reached. Further it will remain zero from then onwards, since once the attractor is unfolded, it is unfolded.

But the criterion in this approach is subjective in some sense that, different values of parameters may lead to different results [Cao.L, 1997)]. For realistic time series data, different optimal embedding dimensions are obtained if we use different values of the threshold value. Also with noisy data this method gives spurious results. [Kantz.H and Schreiber.T, 1997].

### 7.5.2 Minimum Embedding Dimension using Cao's Method

Consider a time series $x_1, x_2, \ldots\ldots x_N$ The time-delay vectors can be reconstructed as follows:

$$y_i(d) = (x_i \;\; x_{i+\tau} \;\; \ldots \; x_{i+(d-1)\,\tau}), \;\; i = 1,2,\ldots\ldots\ldots N-(d\text{-}1)\,\tau$$

where d is the embedding dimension and $\tau$ is the time-delay. Note that $y_i(d)$ means the i$^{th}$ reconstructed vector with embedding dimension d. Similar to the idea of the false neighbor method [Kennel. M. B., Brown. R., and Abarbanel. H. D. I,1992], we define

$$a(i,\, d) \;=\; \frac{\left\| y_i(d+1) - y_{n(i,d)}(d+1) \right\|}{\left\| y_i(d) - y_{n(i,d)}(d) \right\|}, \quad i = 1,2,\ldots\ldots\ldots N\text{-}d\tau$$

where $\| \; \|$ is measurement of Euclidean distance, $y_i(d+1)$ is the i$^{th}$ reconstructed vector with embedding dimension d+1, i.e., $y_i\,(d+1) = (x_i \;\; x_{i+\tau} \;\ldots\;\ldots. x_{i+d\tau})$ and n(i,d), ($1 \le n(i,d) \le N\text{-}d\tau$) is an integer such that $y_{n(i,d)}(d)$ is the nearest neighbor of $y_i(d)$ in the m dimensional reconstructed phase space in the sense of Euclidean distance $\| \; \|$.

In false nearest neighbors method, the authors diagnosed a false neighbor by seeing whether the a(i, d) is larger than some given threshold value. The problem is how to choose this threshold value. It is very difficult and even impossible to give an appropriate and reasonable threshold value, which is independent of the dimension d and each trajectory's point, as well as the considered time series data.

To avoid the above problem, in Cao’s method, the mean of all a(i,d)s is defined by,

$$E(d) = \frac{1}{N - d\tau}\sum_{i=1}^{N-d\tau} a(i,d)$$

E(d) is dependent on only the dimension d and the lag $\tau$. To investigate its variation from d to d+1, a quantity is defined by

$$E1(d) = E(d+1)/ E(d)$$

It is found that E1(d) stops changing when d is greater than some value $d_0$. Then $d = d_0+1$ is the minimum embedding dimension we are looking for.

The parameter $\tau$ is a necessary parameter, which must be given before the minimum embedding dimension is determined numerically. Although in principle the embedding dimension is independent of the time delay $\tau$, the minimum embedding dimension is dependent on $\tau$ in practice. Different values of $\tau$ may lead to different minimum embedding dimensions. We will use the method of Maximum Euclidian Distance Measure (MEDM) from the phase space diagonal to choose the parameter $\tau$ in this work, on account of reasons already explained. The algorithm is implemented using Matlab and the experimental results are shown in the figures 7.10(a-e).

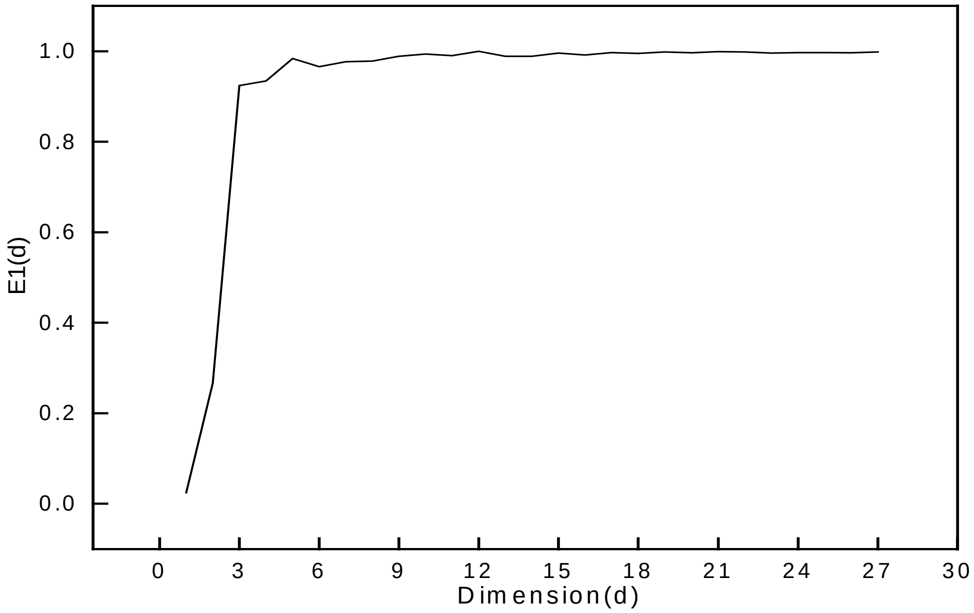


**Fig.7.10(a):** variation of E1(d) with dimension d for Malayalam vowel അ/Λ/

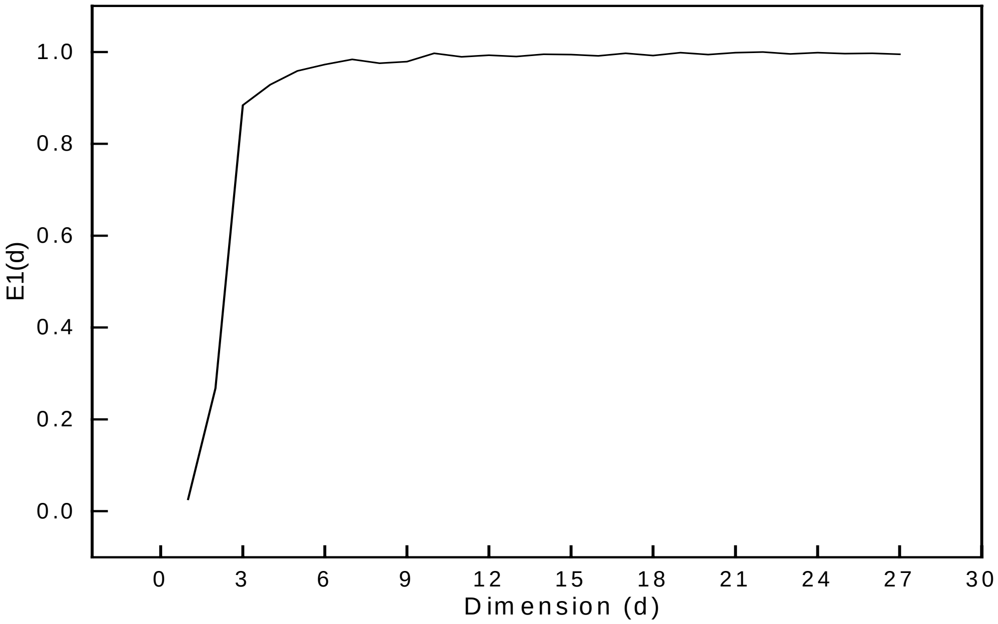


**Fig.7.10(b):** variation of E1(d) with dimension d for Malayalam vowel ഇ/I/

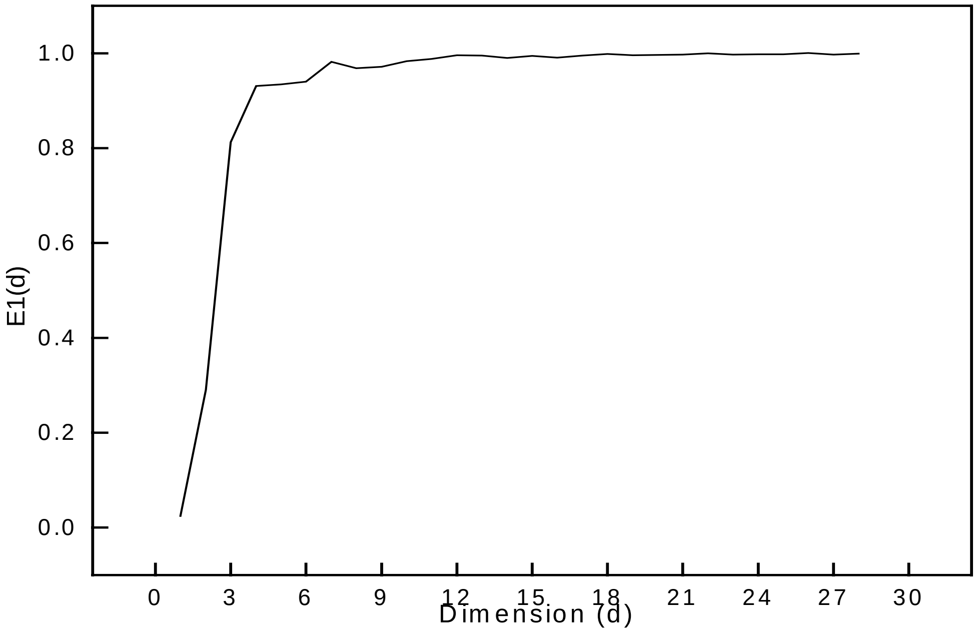


**Fig.7.10(c):** variation of E1(d) with dimension d for Malayalam vowel ഏ/ae/

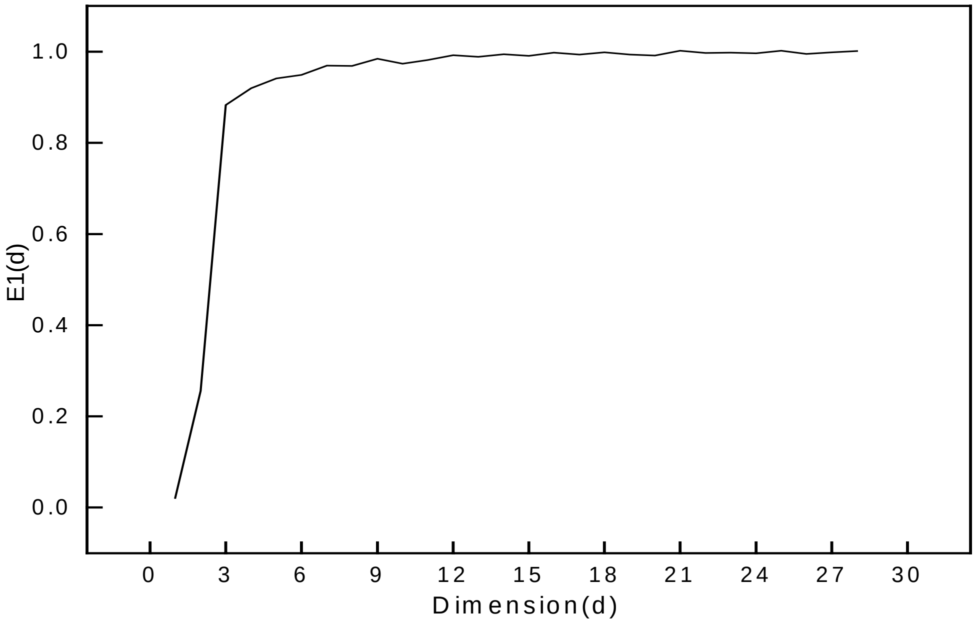


**Fig.7.10(d):** variation of E1(d) with dimension d for Malayalam vowel ഒ/o/

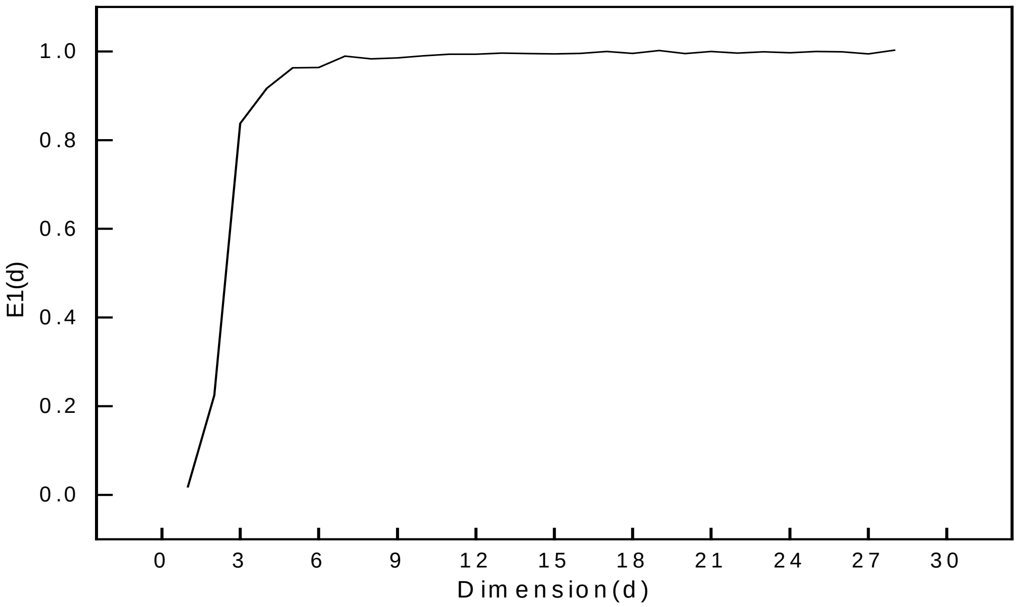


**Fig.7.10(e):** variation of E1(d) with dimension d for Malayalam vowel ഉ/u/

From these figures it may be noted that E1(d) almost remains a constant when the embedding dimension is greater than 2, when time lag is selected by the method of MEDM. That is the minimum embedding dimension according to this result is 3. Therefore in the further analysis we have used the minimum embedding dimension for the reconstruction of phase space as 3.

### 7.6 RPS Features with Optimum Embedding Parameters

The concepts of minimum embedding dimension ‘d’ and the optimum time lag ‘τ’ play a significant role in both the theoretical and practical aspects of Reconstructed Phase Spaces. In the previous session, we found that as far as speech signal is concerned, the optimum time lag corresponding to the Maximum Euclidian Distance Measure (MEDM) from the phase space diagonal is a good choice. The heuristic procedures of false nearest neighbors for determining minimum embedding dimension contain subjective parameters. Cao's algorithm is a solution to avoid such factors. Therefore we have optimized the parameters for reconstructing the Phase Space based on the above mentioned algorithms. In chapter 5 we have described an algorithm for extracting useful parameters from the Reconstructed Phase Space (RPSDP). In the next session modified Reconstructed Phase Space features are extracted with the optimum embedding parameters.

For each vowel, optimum time delay is determined using MEDM from the phase space diagonal. From Cao's algorithm, we have got the minimum embedding dimension as 3.Hence phase space is reconstructed for each vowel with dimension 3 and optimum time delay. RPS distribution plot (scatter graph) for five Malayalam vowels are shown in figures 7.11(a-e).

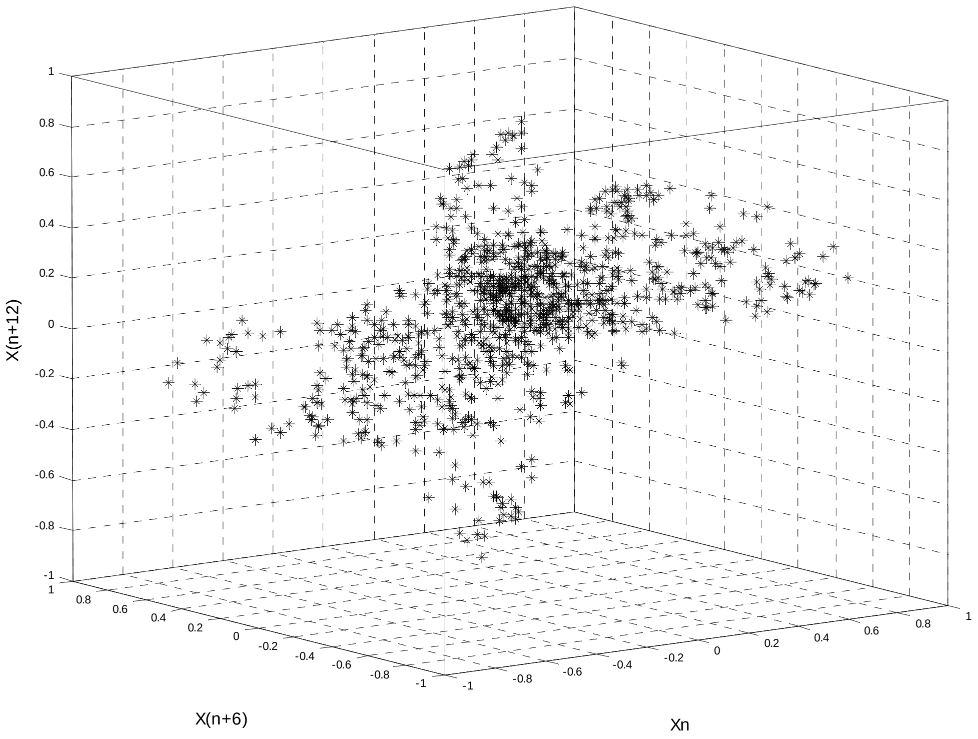


**Fig.7.11(a):** Reconstructed Phase Space Distribution for vowel അ/Λ/

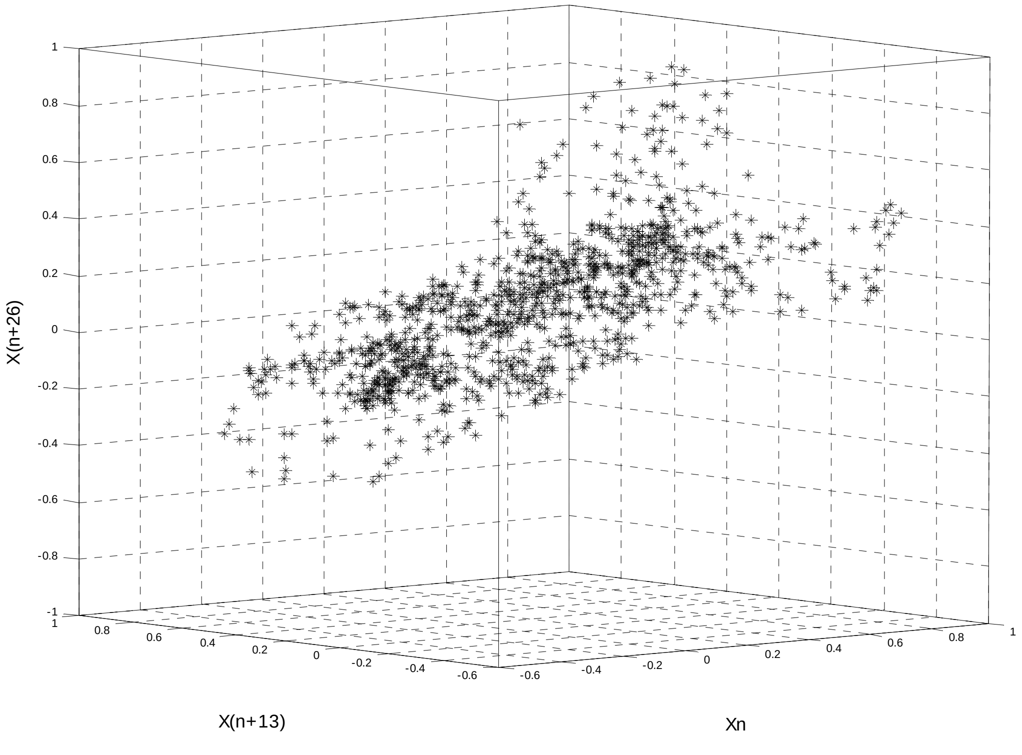


**Fig.7.11(b):** Reconstructed Phase Space Distribution for vowel ഇ/I/

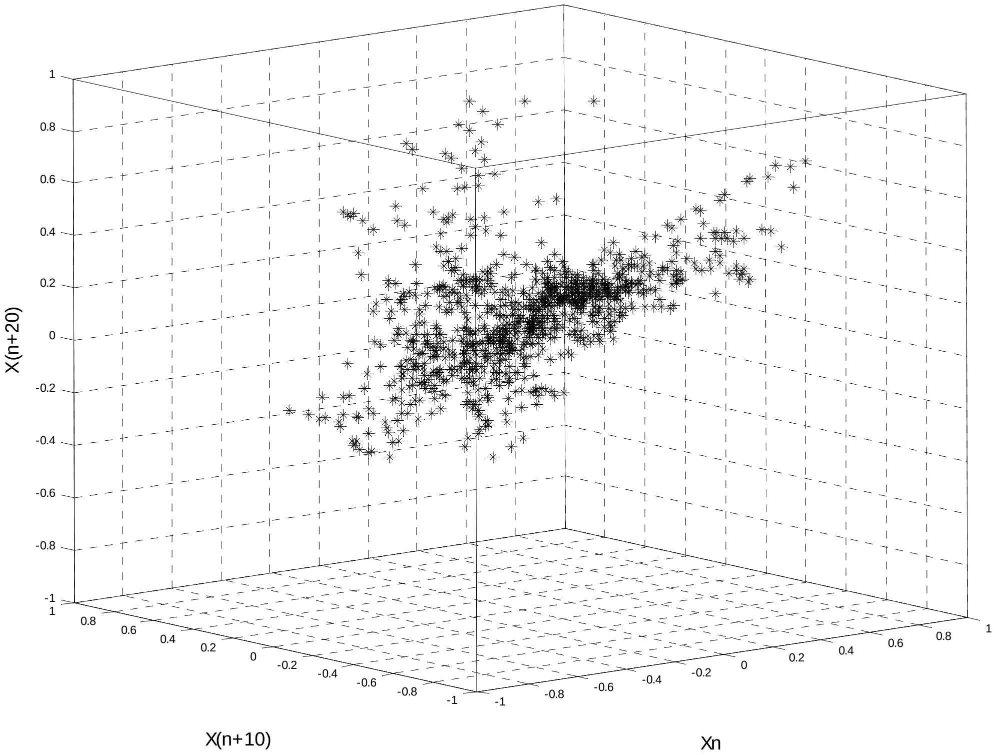


**Fig.7.11(c):** Reconstructed Phase Space Distribution for vowel ഏ/ae/

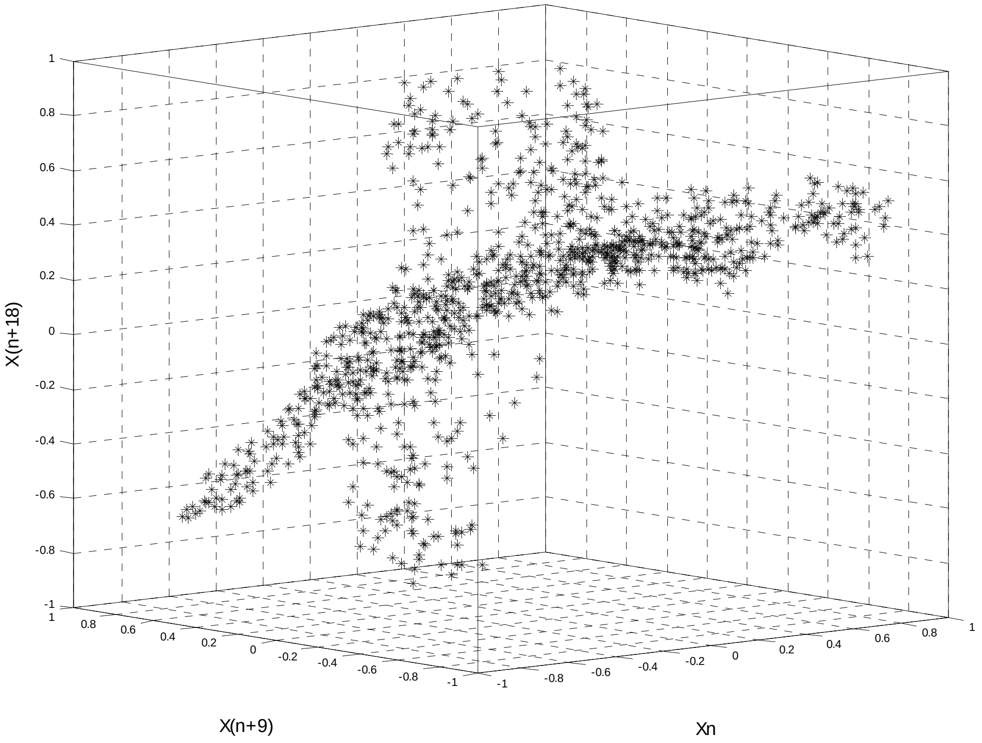


**Fig.7.11(d):** Reconstructed Phase Space Distribution for vowel ഒ/o/

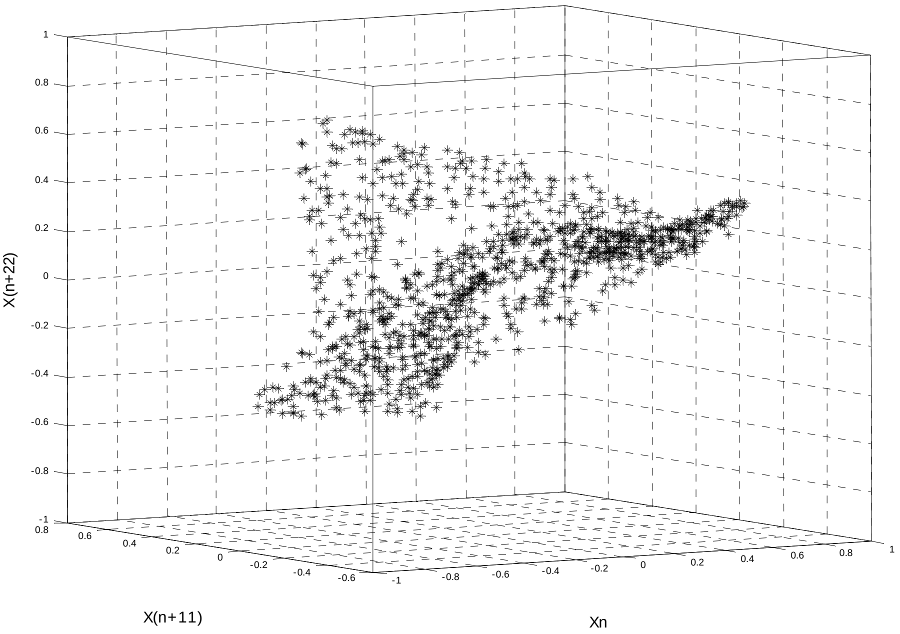


**Fig.7.11(e):** Reconstructed Phase Space Distribution for vowel ഉ/u/

### 7.6.1 Modified RPS Distribution Parameter

To extract the distribution parameter from the three dimensional Reconstructed Phase Space, the entire three dimensional space is divided into 1000 locations. The number of Phase Space Points distributed in each location is calculated as follows.

RPS is divided into grids with 10 x 10 x 10 boxes. The box defined by co-ordinates (-1, .8, -.8) , (-.8, 1, -1) is taken as location 1. Box just right side to it is taken as location 2 and it is extended towards X direction, with the last box in the row (.8, .8, -.8), (1, 1, -1) as location 10. This is repeated for the next row, taking the starting box as location 11 and repeated for all other rows. Then the entire steps are repeated in positive Z direction. The Reconstructed Phase Space Distribution Parameter (RPSDP) is calculated by estimating the number of Phase Space points distributed in each location. Figures 7.12(a-e) show the Modified Reconstructed Phase Space Distribution Parameter versus locations for the vowels അ/Λ/, **ഇ/I/**, **എ/ae/**, **ഒ/o/**, and **ഉ/u/**.

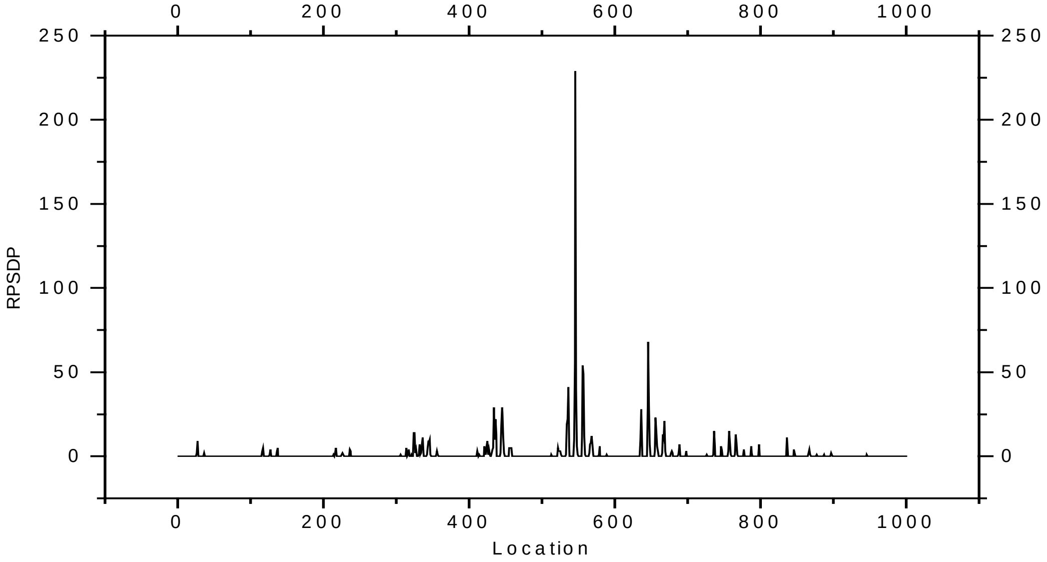


**Fig. 7.12(a)** : Modified RPS Distribution Parameter (Vowel അ/Λ/)

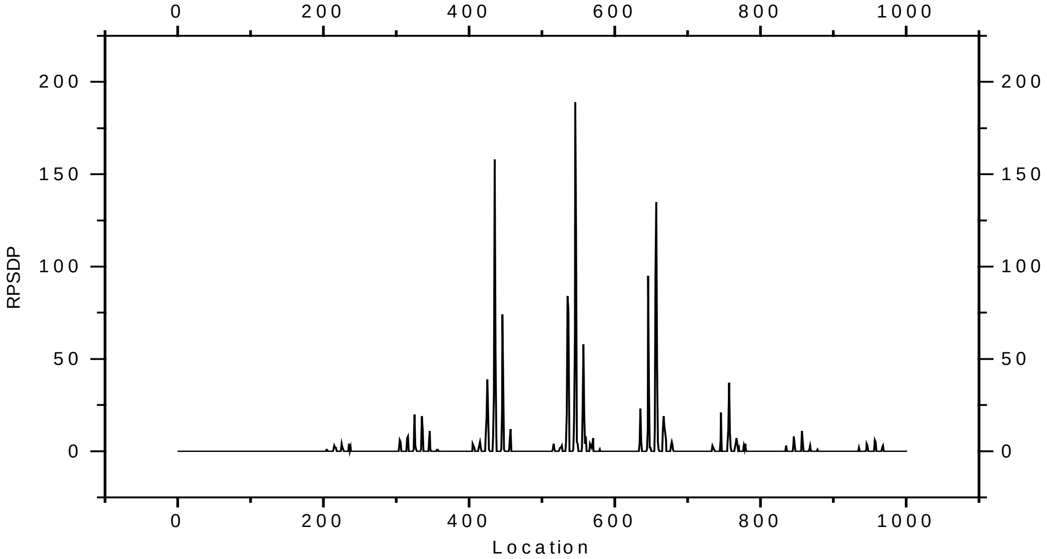


**Fig. 7.12(b)** : Modified RPS Distribution Parameter (Vowel **ഇ/I/**)

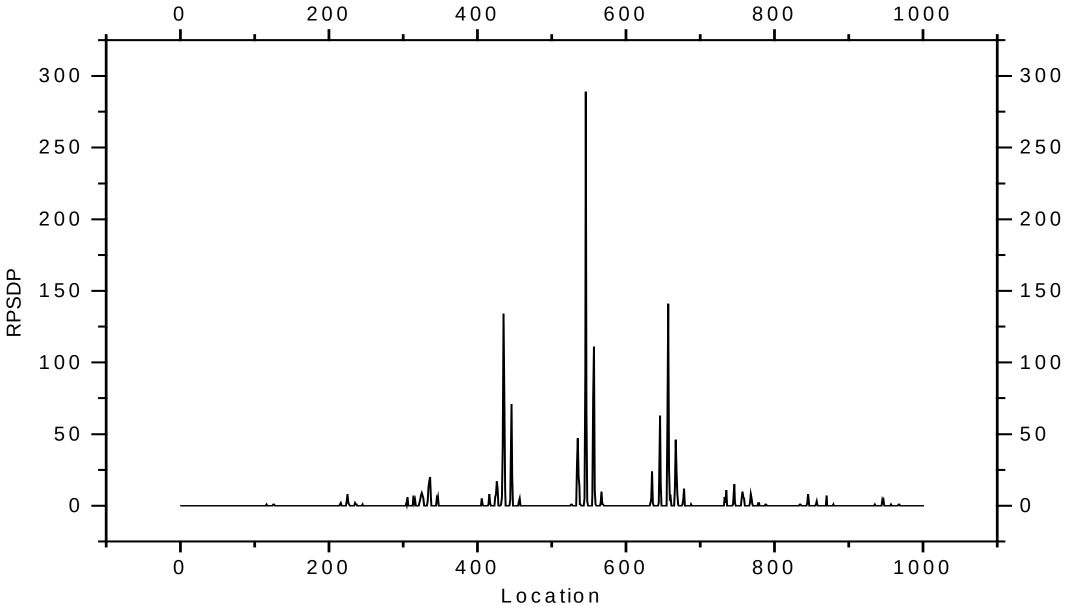


**Fig. 7.12(c)** : Modified RPS Distribution Parameter (Vowel ഏ/**ae**/)

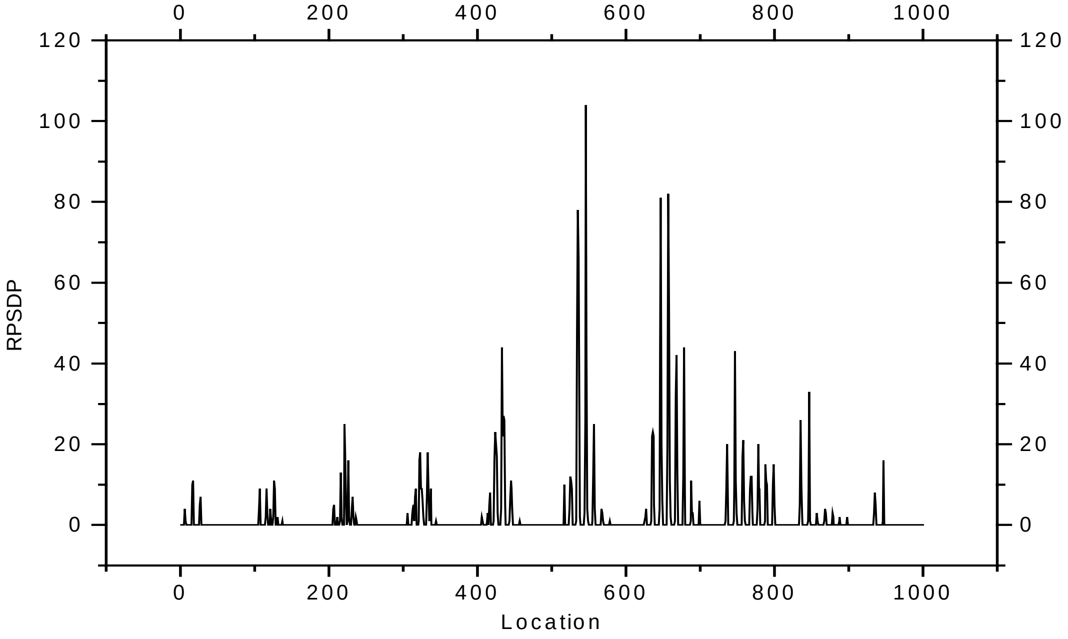


**Fig. 7.12(d)** : Modified RPS Distribution Parameter (Vowel ഒ/**o**/)

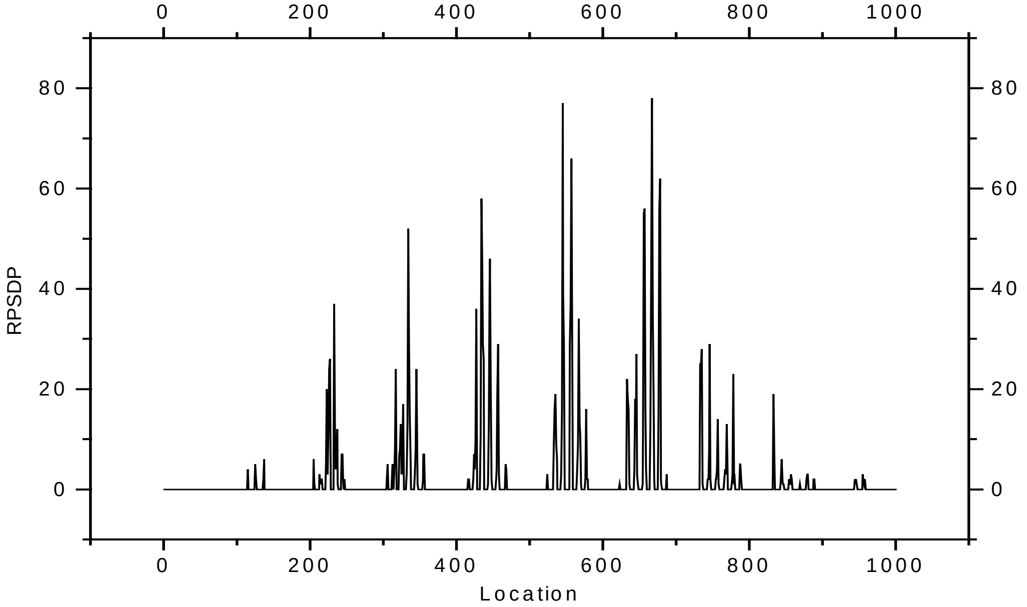


**Fig. 7.12(e)** : Modified RPS Distribution Parameter (Vowel **ഉ/u/**)

This operation is repeated for the same vowel uttered at different occasions. Figures 7.13(a-e) show the Modified Reconstructed phase space distribution parameters for each vowel uttered at different occasions. The graph thus plotted for different vowels shows the identity for each vowel as regard to pattern. Therefore this technique can be effectively utilized for speech recognition applications.

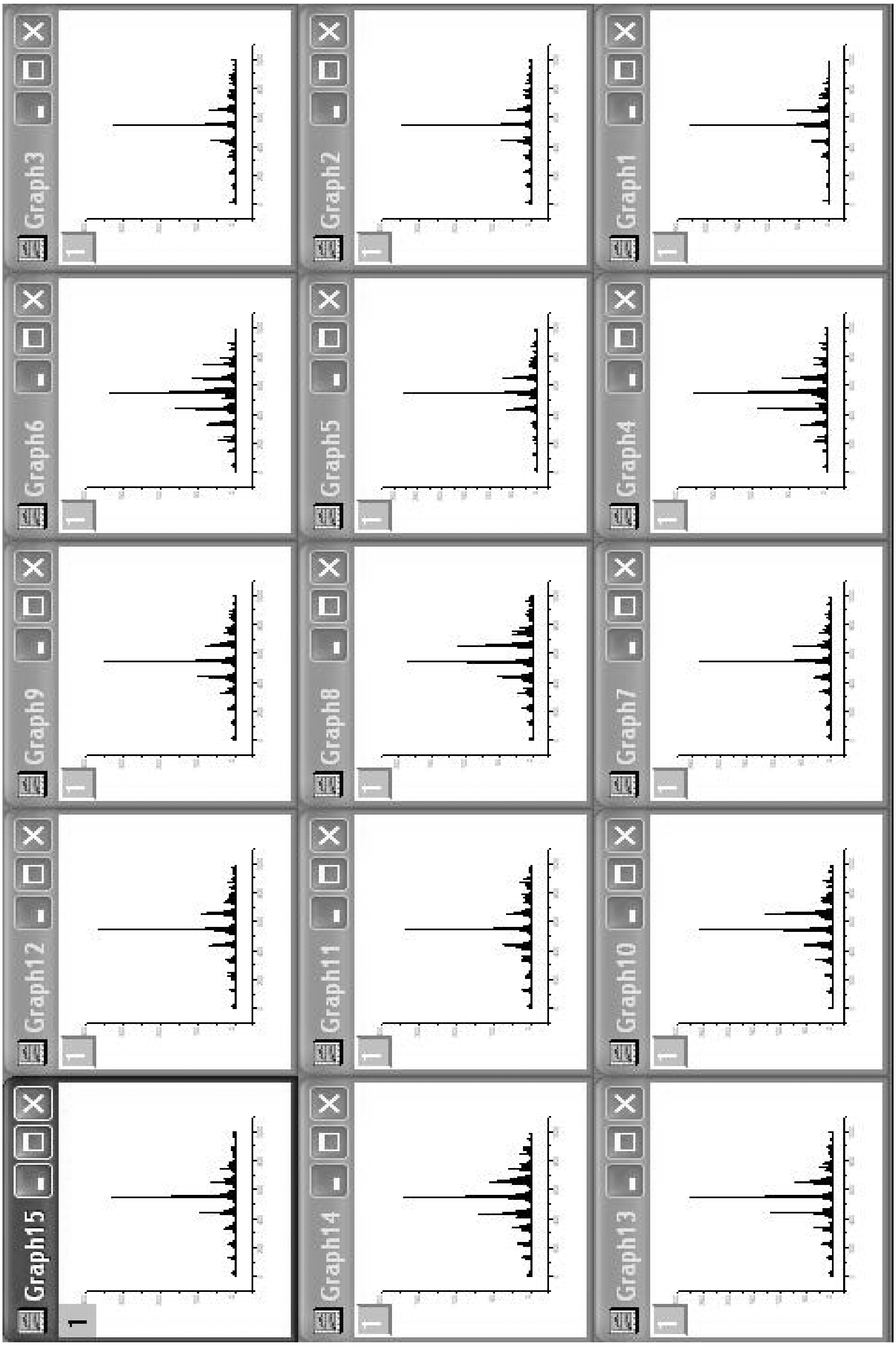


**Fig.7.13(a)** :Modified RPS Distribution Parameter for 15 repeated utterances (Vowel അ/Λ/)

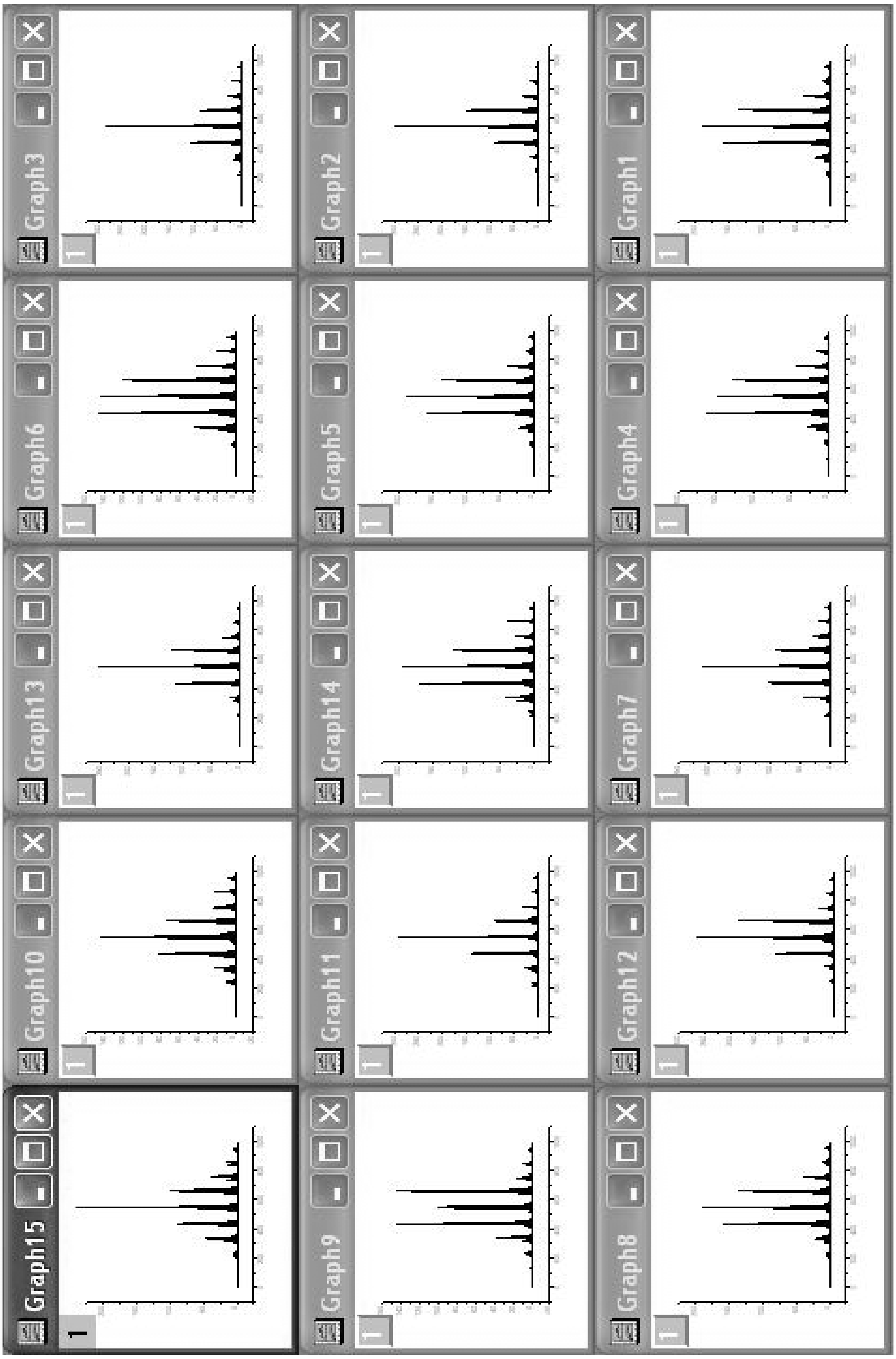


**Fig.7.13(b)** :Modified RPS Distribution Parameter for 15 repeated utterances (Vowel ഇ / **I** /)

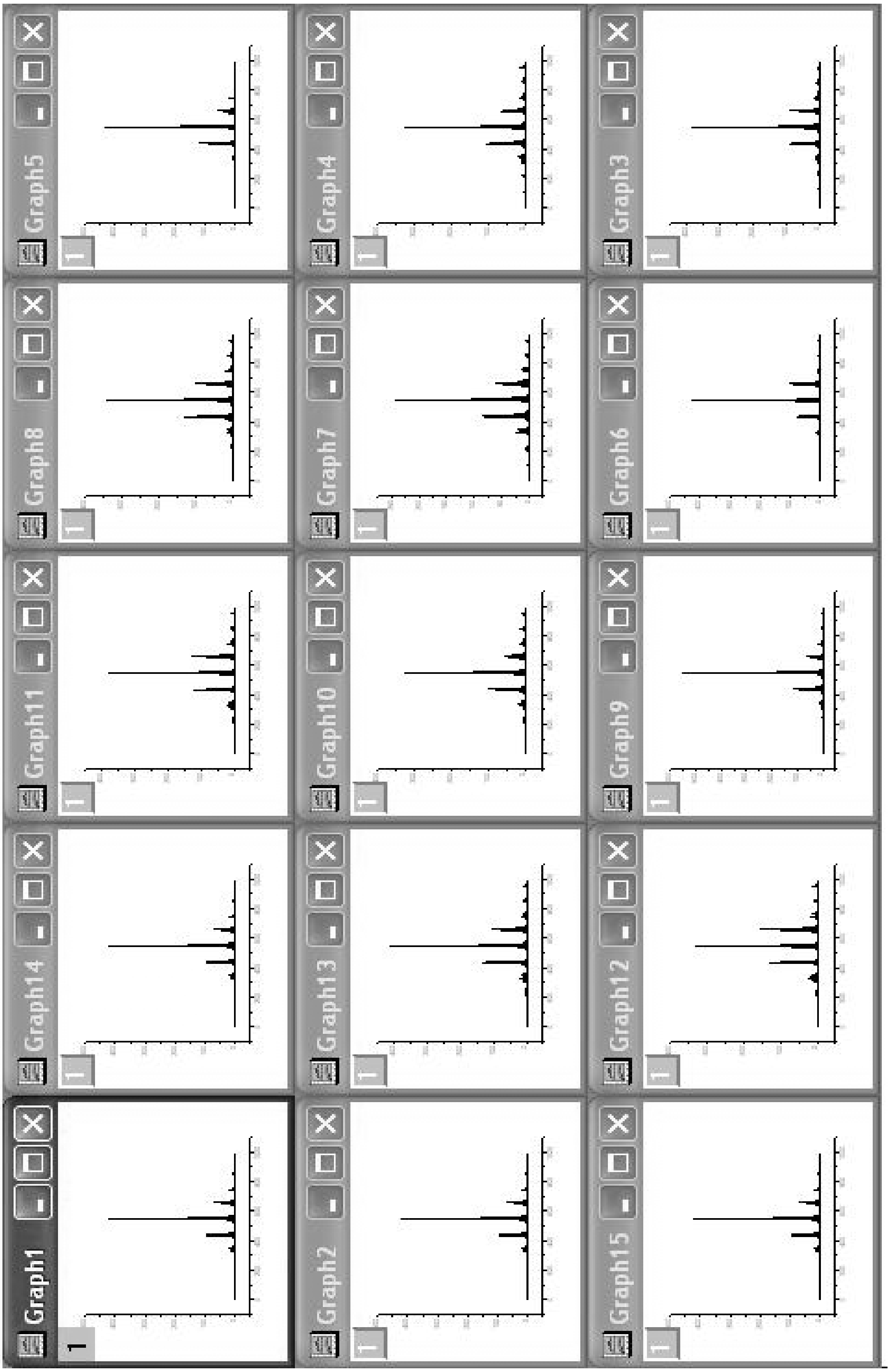


**Fig.7.13(c)** :Modified RPS Distribution Parameter for 15 repeated utterances (Vowel ഐ /**ae**/)

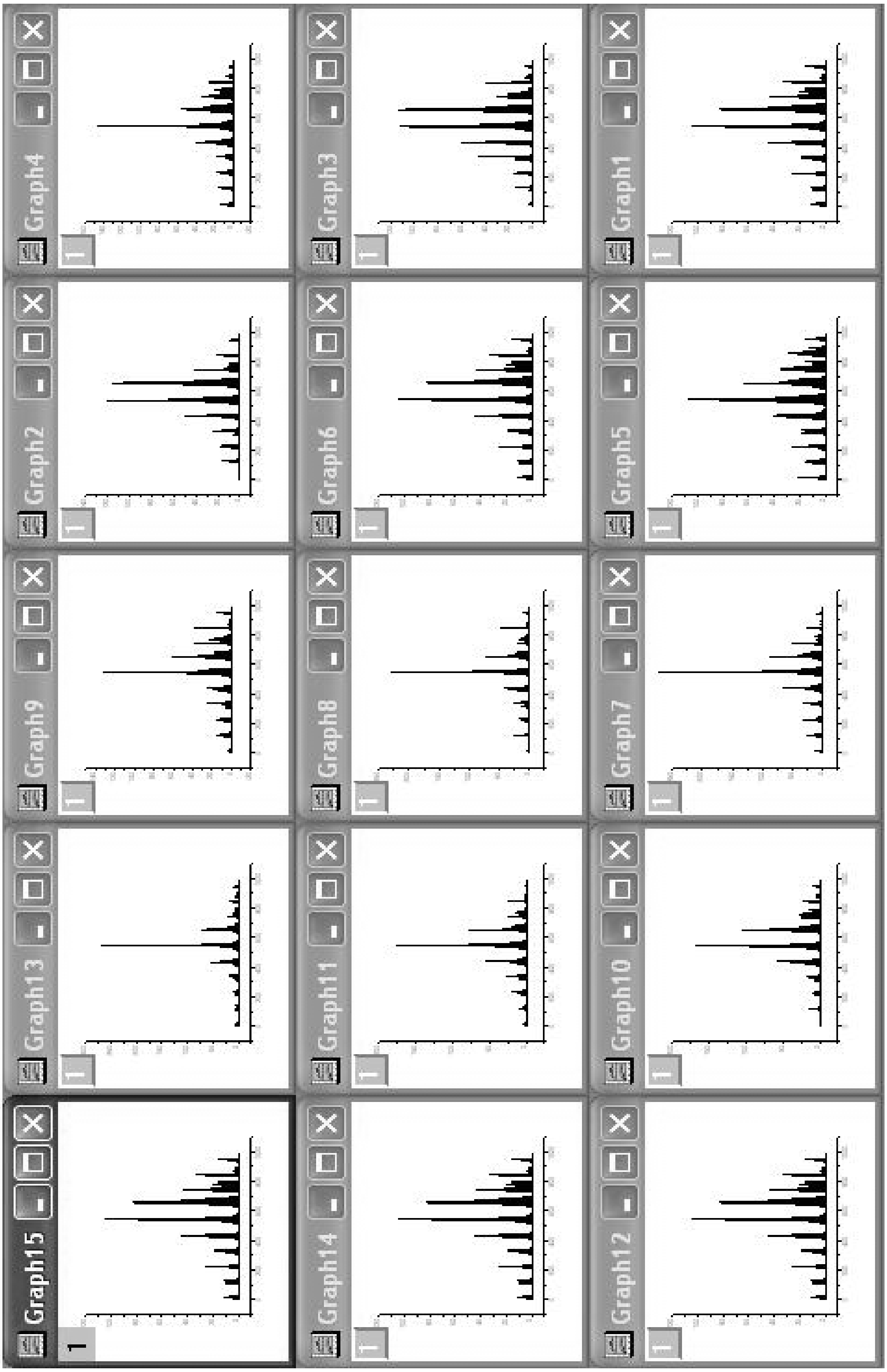


**Fig.7.13(d)** :Modified RPS Distribution Parameter for 15 repeated utterances (Vowel **ഒ/o/**)

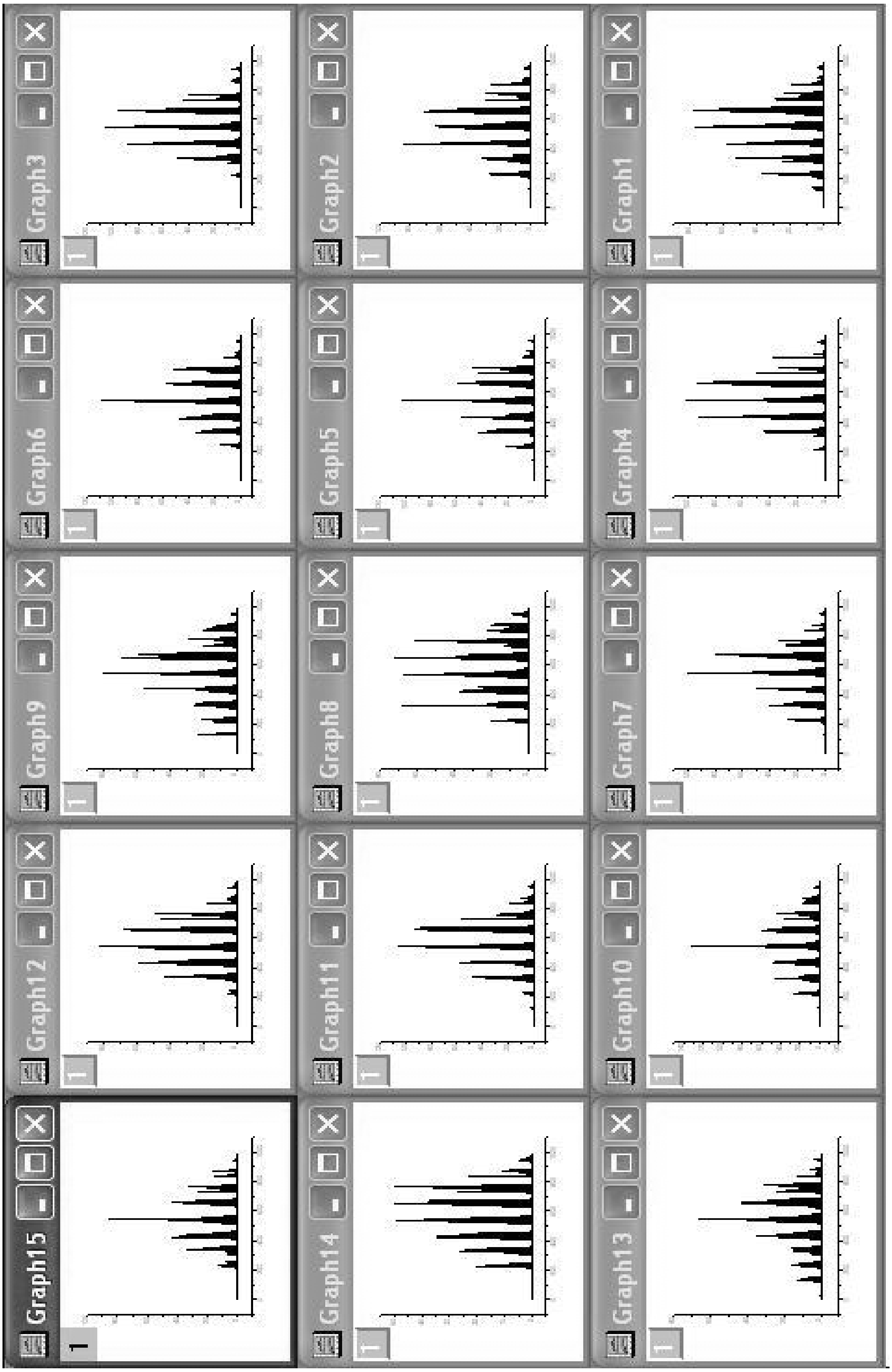


**Fig.7.13(e)** :Modified RPS Distribution Parameter for 15 repeated utterances (Vowel ഉ /**u**/)

## 7.7 Conclusion

The problem of choosing the optimal time delay and the minimum embedding dimension for the reconstruction of phase space using the method of delays are addressed in this chapter. From the discussions of the methods of determining proper time delay, it can be concluded that the optimal delay depends upon the details of the time series as well as the dynamics of the underlying system. We developed a simple procedure that quantifies expansion from the identity line of embedding space. Such a procedure may be more useful in the estimation of the proper time delay. For determining the minimum embedding dimension we have used Cao's method, which does not contain any subjective parameters except time delay for the embedding and is computationally efficient. This method gives consistent results with Malayalam vowels. With these optimum-embedding parameters, Reconstructed Phase Space Distribution Parameter (RPSDP) is modified, whose discriminatory power is illustrated in the next chapter.

## Chapter 8

# Vowel Recognition Using k-NN Classifier and Artificial Neural Network

### 8.1 Introduction

Automatic Speech recognition (ASR) has a history of more than 50 years. With the emerging of powerful computers and advanced algorithms, speech recognition has undergone a great amount of progress over 25 years. Fully automatic speech-based interface to products, which would encompass real-time speech processing as well as language understanding, is still considered to be many years away.

Basic approaches adopted for speech recognition are :

1. Acoustic phonetic approach
2. Pattern recognition approach
3. Artificial Intelligence approach

The **acoustic phonetic** approach is based on the theory of acoustic phonetics that postulates that there exists finite, distinctive phonetic unit in spoken language and that phonetic units are broadly characterized by a set of properties that are manifested in the speech signal, or its spectrum, over time. Even though the acoustic properties of a phonetic unit are highly variable, both with speakers and with neighboring phonetic units (it is called co-articulation of sound), it is assumed that the rules governing the variability are straightforward and can readily be learned and applied in practical situations.

However for a variety of reasons, this approach has limited success in practical systems [Rabiner.L.R and Juang.B.H, 1993]

In **Pattern recognition** approach to speech recognition, the method has two steps namely, training of the speech patterns and recognition of pattern via pattern comparison. This is explained in detail in the later sessions.

The **artificial intelligence** approach to speech recognition is a hybrid of acoustic – phonetic and pattern recognition approaches. The artificial intelligence approach attempts to mechanize the recognition procedure according to the way a person applies his intelligence in visualizing, analyzing, and finally making a decision on the conceived acoustic features.

Pattern recognition is the study of how machines can observe the environment, learn to distinguish pattern of interest from their background, and make sound and reasonable decisions about the categories of the patterns. Automatic (machine) recognition, description, classification and grouping of patterns are important problems in a variety of engineering and scientific disciplines. Pattern recognition can be viewed as the categorization of input data into identifiable classes via the extraction of significant features or attributes of the data from the background of irrelevant details. Duda and Hart [Duda.R.O and Hart.P.E, 1973] define it as a field concerned with machine recognition of meaningful regularities in noisy or complex environment. It encompasses a wide range of information processing problems of great practical significance from speech recognition, handwritten character recognition, to fault detection in machinery and medical diagnosis. Today, pattern recognition is an integral part of most intelligent systems built for decision making.

Normally the pattern recognition processes make use of one of the following two classification strategies.

1. Supervised classification (e.g., discriminant analysis) in which the input pattern is identified as a member of a predefined class.
2. Unsupervised classification (e.g., clustering) in which the pattern is assigned to a hitherto unknown class.

In the present study the well-known approaches that are widely used to solve pattern recognition problems including statistical pattern classifier (*k*-Nearest Neighbor classifier), and connectionist approach (Multi layer Feed forward Artificial Neural Networks) are used for recognizing Malayalam vowels. Here classifiers are based on supervised learning strategy.

The Reconstructed Phase Space Distribution Parameter (RPSDP) extracted as explained in chapter 5 and Modified RPS Distribution Parameter (MRPSDP) using optimum embedding parameters as discussed in chapter 7 are used as input features for recognition study. This chapter is organized as follows.

The first session provides the general description of the pattern recognition approach to speech recognition. The second session deals with recognition experiments conducted using *k*-NN statistical classifier. The third session describes the multi layer feed forward neural network architecture and the simulation experiments conducted for the recognition of Malayalam vowels.

### 8.2 Pattern recognition approach to speech recognition

The block diagram of a typical pattern recognition system for speech recognition is shown in Figure 8.1.

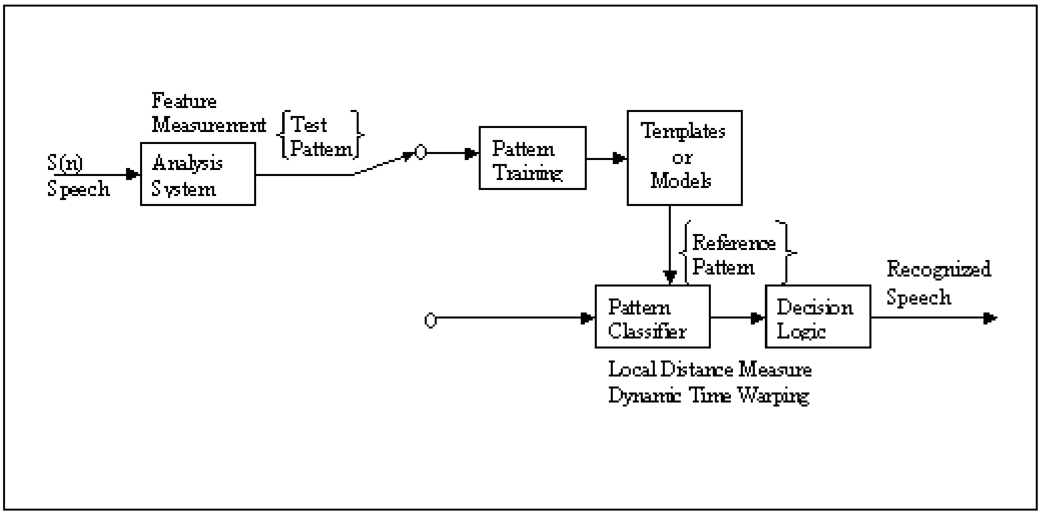


**Fig.8.1:**Block diagram of a pattern recognition system for speech recognition

The pattern recognition paradigm has four steps, namely:

1. **Feature extraction**, in which a sequence of measurements is made on the input signal to define the 'test pattern'. For speech signals the conventional feature measurements are usually the output of some type of spectral analysis technique, such as a filter bank analyzer, a linear predictive coding analysis, or a discrete Fourier transform analysis.
2. **Pattern training**, in which one or more test patterns corresponding to speech sounds of the same class are used to create a pattern, representative of the features of the class. The resulting pattern,

generally called a reference pattern, can be an exemplar or template, derived from some type of averaging technique, or it can be a model that characterizes the statistics of the features of the reference pattern.

3. **Pattern classification**, in which the unknown test pattern is compared with each (sound) class reference pattern and a measure of similarity (distance) between the test pattern and each reference pattern is computed. To compare speech patterns (which consist of a sequence of spectral vectors), we require both local distance measure, in which local distance is defined as the spectral "distance" between two well – defined spectral vectors, and a global time alignment procedure (often called a dynamic time warping algorithm), which compensates for difference of speaking (time scales) of the two patterns.
4. **Decision logic**, in which the reference pattern's similarity scores are used to decide which reference pattern (or possibly which sequence of reference patterns) has best match to the unknown test pattern.

The factors that distinguish the different pattern-recognition approaches are the types of feature measurement, the choice of templates or models for reference patterns, and the method used to create reference patterns and to classify the unknown test pattern. The general strengths and weaknesses of the pattern recognition models include the following:

1. The performance of the system is sensitive to the amount of training data available for creating sound class reference patterns; generally the more training, the higher the performance of the system.

2. The reference patterns are sensitive to the speaking environment and transmission characteristics of the medium used to create the speech. This is because the speech characteristics are affected by transmission and background noise.
3. No speech-specific knowledge is used explicitly in the system; hence, the method is relatively insensitive to choice of the vocabulary of words, task, syntax and semantics.
4. The computational load for both pattern training and pattern classification is generally linearly proportional to the number of patterns being trained or recognized; hence, computation for a large number of sound classes could, and often does, become prohibitive.
5. It is relatively straightforward to incorporate syntactic (and even semantic) constraints directly into the pattern-recognition structure, thereby improving recognition accuracy and reducing the computation.

## 8.3 Statistical Pattern Classification

In the statistical pattern classification process, a ‘d’ dimensional feature vector represents each pattern and it is viewed as a point in the *d*-dimensional space. Given a set of training patterns from each class, the objective is to establish decision boundaries in the feature space, which separate patterns belonging to different classes. The recognition system is operated in two phases, training (learning) and classification (testing). The following section describes the pattern recognition experiment conducted for the recognition of five basic Malayalam vowels using *k*-NN classifier.

### 8.3.1 *k*-Nearest Neighbor Classifier for Malayalam vowel Recognition

Pattern classification by distance functions is one of the earliest concepts in pattern recognition [Tou.J.T and Gonzalez.R.C, 1974], [Friedman.M. and Kandel.A, 1999]. Here the proximity of an unknown pattern to a class serves as a measure of its classification. A class can be characterized by single or multiple prototype pattern(s). The *k*-Nearest Neighbour method is a well-known non-parametric classifier, where a posteriori probability is estimated from the frequency of nearest neighbours of the unknown pattern. It considers multiple prototypes while making a decision and uses a piecewise linear discriminant function. Various pattern recognition studies with first-rate performance accuracy are also reported based on this classification technique [Ray.A.K. and Chatterjee.B, 1984], [Zhang.B and Srihari.S.N, 2004], [Pernkopf.F, 2005].

Consider the case of $m$ classes $c_i$, $i$ =1,….., $m$ and a set of $N$ samples patterns $y_i$, $i$ =1,…, $N$ whose classification is *a priory* known. Let $\mathbf{x}$ denote an arbitrary incoming pattern. The nearest neighbour classification approach classifies $\mathbf{x}$ in the pattern class of its nearest neighbour in the set $y_i$, $i$ = 1,….., $N$ i.e.,

$$\text{If } \| \mathbf{x} - \mathbf{y}_j \|^2 = \min \| \mathbf{x} - \mathbf{y}_i \|^2 \text{ where } 1 \le i \le N$$

$$\text{then } x \,\varepsilon\, c_j \,.$$

This scheme can be termed as 1-NN rule since it employs only one nearest neighbour to $x$ for classification. This can be extended by considering the $k$ nearest neighbours to $x$ and using a majority-rule type classifier. The following algorithm summarizes the classification process.

**Algorithm: Minimum distance *k*-Nearest Neighbor classifier**

```
Input: $N$ – number of pre-classified patterns
    $m$ – number of pattern classes.
    $(y_i, c_i)$, $1 \leq i \leq N$ - $N$ ordered pairs, where $y_i$ is the $i$th pre-classified
    pattern and $c_i$ it's class number ( $1 \leq c_i \leq m$ for all $i$ ).
    $k$ - order of NN classifier (i.e. the $k$ closest neighbors to the incoming
    patterns are considered).
    $x$ - an incoming pattern.
Output: $L$ – class number into which $\boldsymbol{x}$ is classified.
Step 1: Set $S = \{ (y_i, c_i) \}$, where $i = 1, \ldots, N$
Step 2: Find $(y_j, c_j)$ ε $S$ which satisfies
    $\| \boldsymbol{x} - \boldsymbol{y}_j \|^2 = \min \| \boldsymbol{x} - \boldsymbol{y}_i \|^2$ where $1 \leq i \leq m$
Step 3: If $k = 1$ set $L = c_j$ and stop; else initialize an
    $m$ -dimensional vector $I$
    $I( i' ) = 0, i' \neq \mathrm{c}_j$ ; $I(\mathrm{c}_j) = 1$ where $1 \leq i' \leq m$ and
    set $S = S - \{ (y_j, c_j) \}$
Step 4: For $i_0 = 1, \ldots, k-1$ do steps 5-6
Step 5: Find $(y_j, c_j)$ ε $S$ such that
    $\| \boldsymbol{x} - \boldsymbol{y}_j \|^2 = \min \| \boldsymbol{x} - \boldsymbol{y}_i \|^2$ where $1 \leq i \leq N$
```

Step 6: Set $I(c_j) = I(c_j) + 1$ and $S = S - \{ (y_j, c_j) \}$.

Step 7: Set $L = \max \{I(i')\}$, $1 \leq i' \leq m$ and stop.

In the case of $k$-Nearest Neighbor classifier, we compute the distance of similarity between the features of a test sample and the features of every training sample. The class of the majority among the $k$ - nearest training samples is deemed as the class of the test sample.

### 8.3.2 Simulation Experiments and Results

The recognition experiment is conducted by simulating the above algorithm using MATLAB. The Reconstructed Phase Space Distribution Parameter (RPSDP) extracted as discussed in Chapter 5, and Modified RPS Distribution Parameter (MRPSDP) as explained in chapter 7 are used in the recognition study. Here we used the database consisting of 250 samples of five Malayalam vowels collected from a single speaker for training and a disjoint set of vowels of same size from the database for recognition purpose.

The recognition accuracies obtained for Malayalam vowels using the above said features using $k$-NN classifier are tabulated in Table 8.1. The graphical representation of these recognition results based on the features using k-NN classifier is shown in figure 8.2.

The overall recognition accuracies obtained for Malayalam vowels using k-NN classifier with RPSDP and MRPSDP features are 83.12%, and 86.96% respectively. This algorithm does not fully accommodate the small variations in the extracted features. In the next section we present a recognition study conducted using Multi layer Feed forward neural network that is capable of adaptively accommodating the minor variations in the extracted features.

| Vowel Number | Vowel Unit | Average Recognition Accuracy (%) | |
|---|---|---|---|
| | | RPSPD Feature | MRPSPD Feature |
| 1 | അ/Λ/ | 90.4 | 94.8 |
| 2 | ഇ/I/ | 79.2 | 84.4 |
| 3 | എ/ae/ | 70.8 | 73.6 |
| 4 | ഒ/o/ | 82.8 | 86 |
| 5 | ഉ/u/ | 92.4 | 96 |
| Overall Recognition Accuracy (%) | | 83.12 | 86.96 |

**Table 8.1**: Recognition Accuracies of Malayalam Vowels based on RPSPD and MRPSPD features using *k*-NN Classifier

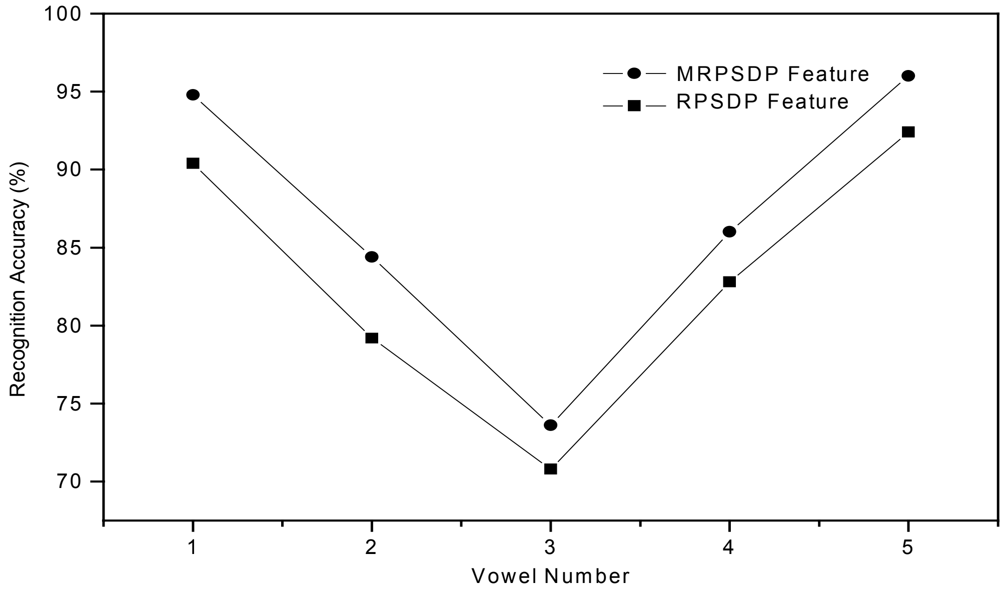


**Fig. 8.2**: Vowel No. Vs. Recognition Accuracies of Malayalam Vowels based on RPSPD and MRPSPD features using *k*-NN Classifier

## 8.4 Application of Neural Networks for Speech Recognition

Neural network is a mathematical model of information processing in human beings. A neural network, which is also called a connectionist model or a Parallel Distributed Processing (PDP) model, is basically a dense interconnection of simple, nonlinear computation elements. The structure of digital computers is based on the principle of sequential processing. These sequential based computers have achieved only little progress in the area like speech and image recognition. An adaptive system having a capability comparable to the human intellect is needed for performing better results in the above said areas. In human beings these types of processing are done using massively parallel-interconnected neuron systems.

A set of processing units when assembled in a closely interconnected network, offers a surprisingly rich structure, exhibiting some features of biological neural network. Such a structure is called an Artificial Neural Network (ANN). The ANN is based on the notion that complex "computing" operations can be implemented by massive integration of individual computing units, each of which performs an elementary computation. Artificial neural networks have several advantages relative to sequential machines. First, the ability to adapt is at the very center of ANN operations. Adaptation takes the form of adjusting the connection weights in order to achieve desired mappings. Furthermore ANN can continue to adapt and learn, which is extremely useful in processing and recognition of speech. Second, ANN tend to move robust or fault tolerant than Von Neumann machines because the network is composed of many interconnected neurons, all computing in parallel, and failure of a few processing units can often be compensated for by the redundancy in the network. Similarly ANN can often generalize from incomplete or noisy data. Finally ANN when used as classifier does not require strong statistical characterization or parameterization of data. Since the advent of Feed Forward Multi Layer Perception (FFMLP) and error-back propagation training algorithm, great improvements in terms of recognition performance and automatic training have been achieved in the area of recognition applications. These are the main motivations to choose artificial neural networks for speech recognition.

The following sections deal with the recognition experiments conducted based on the feed-forward neural network for Malayalam vowels. A brief description about the diverse use of neural networks in pattern recognition followed by the general ANN architecture is presented first. In the next section the error back propagation algorithm used for training FFMLP is illustrated. The Final section deals with the description of simulation experiments and recognition results.

### 8.4.1 Neural Networks for Pattern Recognition

Artificial Neural Networks (ANN) can be most adequately characterized as computational models with particular properties such as the ability to adapt or learn, to generalize, to cluster or organize data, based on a massively parallel architecture. The history of ANNs starts with the introduction of simplified neurons in the work of McCulloch and Pitts [McCulloch.W.S and Pitts.W, 1943]. These neurons were presented as models of biological neurons and as conceptual mathematical neurons like threshold logic devices that could perform computational task. The work of Hebb further developed the understanding of this neural model [Hebb.D.O, 1949]. Hebb proposed a qualitative mechanism describing the process by which synaptic connections are modified in order to reflect the learning process undertaken by interconnected neurons, when they are influenced by some environmental stimuli. Rosenblatt with his perceptron model, further enhanced our understanding of artificial learning devices [Rosenblatt.F., 1959]. However, the analysis by Minsky and Papert in their work on perceptrons, in which they showed the deficiencies and restrictions existing in these simplified models, caused a major set back in this research area [Minsky .M.L and Papert.S.A., 1988]. ANNs attempt to replicate the computational power (low level arithmetic processing ability) of biological neural networks and, there by, hopefully endow machines with some of the (higher-level) cognitive abilities that biological organisms possess. These networks are reputed to possess the following basic characteristics:

- Adaptiveness: the ability to adjust the connection strengths to new data or information
- Speed : due to massive parallelism
- Robustness: to missing, confusing, and/ or noisy data

- Optimality: regarding the error rates in performance

Several neural network learning algorithms have been developed in the past years. In these algorithms, a set of rules defines the evolution process undertaken by the synaptic connections of the networks, thus allowing them to learn how to perform specified tasks. The following sections provide an overview of neural network models and discuss in more detail about the learning algorithm used in classifying Malayalam vowels, namely the Back-propagation (BP) learning algorithm.

### 8.4.2 General ANN Architecture

A neural network consists of a set of massively interconnected processing elements called neurons. These neurons are interconnected through a set of connection weights, or synaptic weights. Every neuron *i* has $\mathrm{N_i}$ inputs, and one output $Y_i$. The inputs labeled $s_{i1}$, $s_{i2}$, …, $s_{iNi}$ represent signals coming either from other neurons in the network, or from external world. Neuron *i* has $N_i$ synaptic weights, each one associated with each of the neuron inputs. These synaptic weights are labeled $w_{i1}$, $w_{i2}$,…,$w_{iNi}$, and represent real valued quantities that multiply the corresponding input signal. Also every neuron *i* has an extra input, which is set to a fixed value $\theta$ , and is referred to as the threshold of the neuron that must be exceeded for there to be any activation in the neuron. Every neuron computes its own internal state or total activation, according to the following expression,

$$\mathrm{x}_j = \sum_{\mathrm{i=1}}^{\mathrm{N_i}} \mathrm{w}_{ij}\mathrm{s}_{ij} + \theta_i \qquad j = 1,2,\ldots\ldots.,\mathrm{M}$$

where $M$ is the total number of Neurons and $N_i$ is the number of inputs to each neuron. Figure 8.3 shows a schematic description of the neuron. The total activation is simply the inner product of the input vector $S_i$= [$s_{i0}$, $s_{i1}$, …, $s_{iNi}$]$^T$ by the weight vector $W_i$ = [$w_{i0}$, $w_{i1}$,…$w_{iNi}$]$^T$. Every neuron computes its output according to a function $Y_i = f(x_i)$, also known as threshold or activation function. The exact nature of $f$ will depend on the neural network model under study.

In the present study, we use a mostly applied sigmoid function in the thresholding unit defined by the expression,

$$S(x) = \frac{1}{1 + e^{-ax}}$$

This function is also called S-shaped function. It is a bounded, monotonic, non-decreasing function that provides a graded nonlinear response as shown in figure 8.4

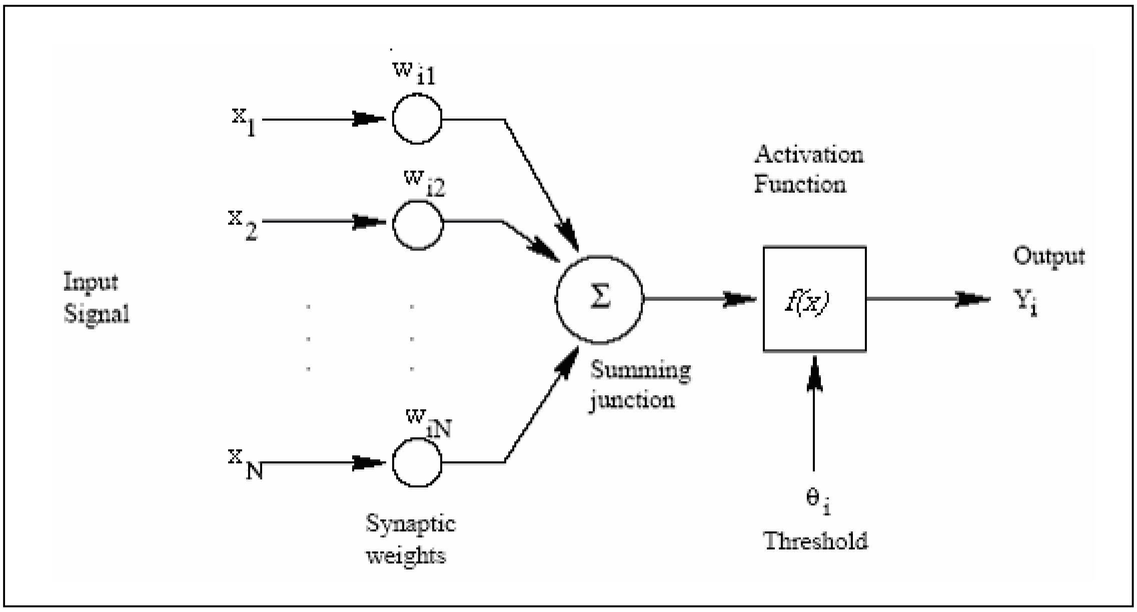


**Fig.8.3:** Simple neuron representation

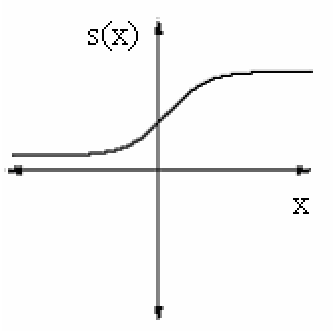


**Fig.8.4:** Sigmoid threshold function

The network topology used in this study is the feed forward network. In this architecture the data flow from input to output units strictly feed forward, the data processing can extend over multiple layers of units but no feed back connections are present.

This type of structure incorporates one or more hidden layers, whose computation nodes are correspondingly called hidden neurons or hidden nodes. The function of the hidden nodes is to intervene between the external input and the network output. By adding one or more layers, the network is able to extract higher-order statistics. The ability of hidden neurons to extract higher-order statistics is particularly valuable when the size of the input layer is large. The structural architecture of the neural network is intimately linked to the learning algorithm used to train the network. In this study we used Error Back-propagation learning algorithm to train the input patterns in the multi layer feed forward neural network. The detailed description of the learning algorithm is given in the following section.

### 8.4.3 Back-propagation Algorithm for Training FFMLP

The back propagation algorithm (BP) is the most popular method for neural network training and it has been used to solve numerous real life problems. In a multi layer feed forward neural network Back Propagation algorithm performs iterative minimization of a cost function by making weight connection adjustments according to the error between the computed and desired output values. Figure 8.5 shows a general three layer network, where $o_k$ is the actual output value of the output layer unit k, $o_j$ is the output of the hidden layer unit j, $w_{ij}$ and $w_{ik}$ are the synaptic weights.

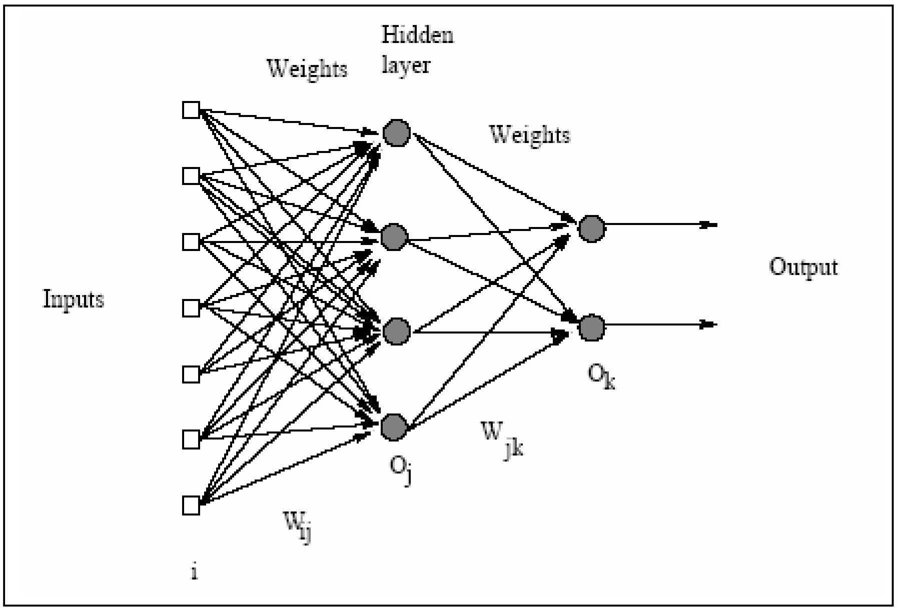


**Fig.8.5:** A general three layer network

The following relationships for the derivation of the back-propagation hold :

$$o_k = \frac{1}{1+e^{-net_k}}$$

$$net_k = \sum_k w_{ik} o_j$$

$$o_j = \frac{1}{1+e^{-net_j}}$$

$$net_j = \sum_j w_{ij} o_i$$

The cost function (error function) is defined as the mean square sum of differences between the output values of the network and the desired target values. The following formula is used for this error computation [Haykins.S, 2004],

$$\mathrm{E} = \frac{1}{2}\sum_p \left( \sum_k \left(t_{pk} - o_{pk}\right)^2 \right)$$

where *p* is the subscript representing the pattern and *k* represents the output units. In this way, t$_{pk}$ is the target value of output unit *k* for pattern *p* and o$_{pk}$ is the actual output value of layer unit *k* for pattern *p*. During the training process a set of feature vectors corresponding to each pattern class is used. Each training pattern consists of a pair with the input and corresponding target output. The patterns are presented to the network sequentially, in an iterative manner. The appropriate weight corrections are performed during the process to adapt the network to the desired behavior. The iterative procedure continues until the connection weight values allow the network to perform the required mapping. Each presentation of whole pattern set is named an *epoch*.

The minimization of the error function is carried out using the gradient-descent technique [Haykins.S, 2004]. The necessary corrections to the weights of the network for each iteration *n* are obtained by calculating the partial derivative of the error function in relation to each weight $w_{jk}$, which gives a direction of steepest descent. A gradient vector representing the steepest increasing direction in the weight space is thus obtained. Due to the fact that a minimization is required, the weight update value $\Delta w_{jk}$ uses the negative of the corresponding gradient vector component for that weight. The delta rule determines the amount of weight update based on this gradient direction along with a step size,

$$\Delta w_{jk} = -\eta \frac{\partial E}{\partial w_{jk}}$$

The parameter $\eta$ represents the step size and is called the learning rate. The partial derivative is equal to,

$$\frac{\partial E}{\partial w_{jk}} = \frac{\partial E}{\partial o_k} \frac{\partial o_k}{\partial net_k} \frac{\partial net_k}{\partial w_{jk}} = -(t_k - o_k) o_k (1 - o_k) o_j$$

The error signal $\delta_k$ is defined as

$$\delta_k = (t_k - o_k) o_k (1 - o_k)$$

so that the delta rule formula becomes

$$\Delta w_{jk} = \eta \delta_k o_j$$

For the hidden neuron, the weight change of $w_{ij}$ is obtained in a similar way. A change to the weight, $w_{ij},$ changes $o_j$ and this changes the inputs into each unit *k,* in the output layer. The change in *E* with a change in $w_{ij}$ is therefore the sum of the changes to each of the output units. The change rules produces:

$$\frac{\partial E}{\partial w_{ij}} = \sum_k \frac{\partial E}{\partial o_k} \frac{\partial o_k}{\partial net_k} \frac{\partial net_k}{\partial o_j} \frac{\partial o_j}{\partial net_j} \frac{\partial net_j}{\partial w_{ij}}$$

$$= \sum_k -\left(t_k - o_k\right) o_k \left(1 - o_k\right) w_{jk} o_j \left(1 - o_j\right) o_i$$

$$= -o_i o_j \left(1 - o_j\right) \sum_k \delta_k w_{jk}$$

so that defining the error $\delta_j$ as

$$\delta_j = o_j \left(1 - o_j\right) \sum_k \delta_k w_{jk}$$

we have the weight change in the hidden layer is equal to

$$\Delta w_{ij} = \eta \delta_j o_i$$

The $\delta_k$ for the output units can be calculated using directly available values, since the error measure is based on the difference between the desired output $t_k$ and the actual output $o_k$. However, that measure is not available for the hidden neurons. The solution is to back-propagate the $\delta_k$ values, layer by layer through the network, so that finally the weights are updated.

A momentum term was introduced in the back-propagation algorithm by Rumelhart [Rumelhart.D.E. *et al.*, 1986]. Here the present weight is modified by incorporating the influence of the passed iterations. Then the delta rule becomes

$$\Delta w_{ij}(n) = -\eta \frac{\partial E}{\partial w_{jk}} + \alpha \Delta w_{ij}(\mathrm{n}-1)$$

where α is the momentum parameter and determines the amount of influence from the previous iteration on the present one. The momentum introduces a *damping* effect on the search procedure, thus avoiding oscillations in irregular areas of the error surface by averaging gradient components with opposite sign and accelerating the convergence in long flat areas. In some situations it possibly avoids the search procedure from being stopped in a local minimum, helping it to skip over those regions without performing any minimization there. Momentum may be considered as an approximation to a second–order method, as it uses information from the previous iterations. In some applications, it has been shown to improve the convergence of the back propagation algorithm. The following section describes the simulation of recognition experiments and results for Malayalam vowels.

### 8.4.4 Simulation Experiments and Results

Present study investigates the recognition capabilities of the above explained FFMLP-based Malayalam vowel recognition system. For this purpose the multi layer feed forward neural network is simulated with the Back propagation learning algorithm. A constant learning rate, 0.01, is used (Value of $\eta$ was found to be optimum as 0.01 by trial and error method). The initial weights are obtained by generating random numbers ranging from 0.1 to 1. The number of nodes in the input layer is fixed according to the feature vector size. Since five Malayalam vowels are analyzed in this experiment, the number of nodes in the output layer is fixed as 5. The recognition experiment is repeated by changing the number of hidden layers and number of nodes in each hidden layer. After this trial and error experiment, the number of hidden layers is fixed as two, the number of nodes in the hidden layer is set to fifteen and the number of *epochs* as 10,000 for obtaining the successful architecture in the present study.

The network is trained using the RPSDP features and MRPSDP features extracted for Malayalam vowels separately. Here we used a set of 250 samples each of five Malayalam vowels for iteratively computing the final weight matrix and a disjoint set of vowels of same size from the database for recognition purpose. The recognition accuracies obtained for the Malayalam vowels based on above said features using multi layer feed forward neural network classifier are tabulated in Table 8.2. The graphical representation of these recognition results based on different features using neural network is shown in figure 8.6.

| Vowel Number | Vowel Unit | Average Recognition Accuracy (%) | |
|---|---|---|---|
| | | **RPSPD Feature** | **MRPSPD Feature** |
| **1** | അ/Λ/ | 96.4 | 97.2 |
| **2** | ഇ/I/ | 87.6 | 90 |
| **3** | എ/ae/ | 82.4 | 86.4 |
| **4** | ഒ/o/ | 89.6 | 92.4 |
| **5** | ഉ/u/ | 96.8 | 98.8 |
| Overall Recognition Accuracy (%) | | 90.56 | 92.96 |

**Table 8.2**: Recognition Accuracies of Malayalam Vowels based on RPSPD and MRPSPD features using Neural Network

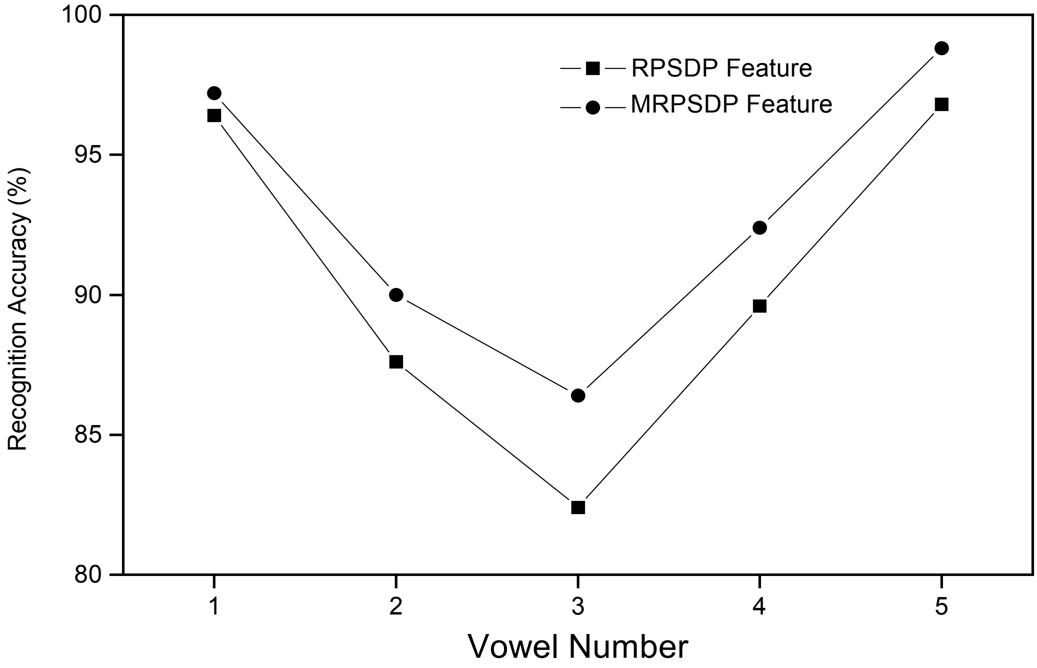


**Fig. 8.6**: Vowel No. Vs. Recognition Accuracies of Malayalam Vowels based on RPSPD and MRPSPD features using Neural Network

The overall recognition accuracies obtained for Malayalam vowels using Multi layer feed forward Neural Network with RPSDP and MRPSDP features are 90.56%, and 92.96% respectively.

From the above classification experiments, the overall highest recognition accuracy (92.96%) is obtained for the MRPSDP features using Multi layer feed forward neural network. Compared to the recognition results, obtained for *k*-NN classifier (86.96%) based on MRPSDP feature, the neural network gives better performance. These results indicate that, for pattern recognition problems the connectionist model based learning is more adequate than the existing statistical classifiers.

### 8.5 Conclusion

Malayalam vowel recognition studies based on the parameters developed in chapter 5 and 7 using different classifiers are presented in this chapter. The credibility of the extracted parameters is tested with the k-NN classifier. A connectionist model based recognition system by means of multi layer feed forward neural network with error back propagation algorithm is then implemented and tested using RPSDP features and MRPSDP features extracted from the vowels. The highest recognition accuracy (92.96%) is obtained with MRPSDP feature using neural network classifier. These results specify the discriminatory strength of the Reconstructed Phase Space derived features for isolated Malayalam vowel classification experiments. The above said RPS derived features are time domain based features. The performance of the recognition experiments can be further improved by combing these features with the traditional frequency domain based Mel frequency cepstral coefficient features (MFCCs). Performance of this hybrid parameter is demonstrated in the next chapter.

# Chapter 9

# Vowel Recognition Using Joint Feature Vectors

## 9.1 Introduction

As mentioned in the earlier chapters, for the study presented in this thesis, speech is modeled as a nonlinear system and Reconstructed Phase Space is taken as the processing space. The classification experiments presented in the previous chapter demonstrated that the features extracted from the Reconstructed Phase Space contain substantial discriminatory power. But current speech recognition systems use frequency domain features, such as Mel-frequency cepstral coefficients (MFCCs), which are based upon a switched linear model of the human speech production mechanism. This model describes human speech production as an excitation source and a linear time invariant filter representing the vocal tract. Cepstral analysis allows the excitation source energy to be separated from the frequency response characteristics of the vocal tract.

The purpose of the work given in the present chapter is to extend the nonlinear methods we developed, by combing the nonlinear based RPS derived features with the traditional MFCC feature set to achieve a boost in the accuracy of recognition experiments. If the frequency domain features contain different discriminatory information than the RPS derived features, we should get better results.

This chapter is organized as follows. In the first session, a detailed description of cepstral analysis and the Mel frequency cepstral coefficients are presented. In the next session, recognition experiments with MFCC feature set alone as input parameter is described. Following this the formation of hybrid feature vector by combing the frequency domain and Reconstructed Phase Space derived features is presented. Finally the simulation experiments with this hybrid feature set and results are demonstrated.

## 9.2 Introduction to cepstral analysis

The aim of cepstral analysis methods is to extract the vocal tract characteristics from the excitation source, because the vocal tract characteristics are what contain the information about the phoneme utterance [Deller. J. R., Proakis. J. G. *et al*, 2000]. Cepstral analysis is a form of homomorphic signal processing, where nonlinear operations are used to give the equations having the properties of linearity [Deller. J. R., Proakis. J. G. *et al*, 2000]. One typical model used to represent the entire speech production mechanism is given in figure below.

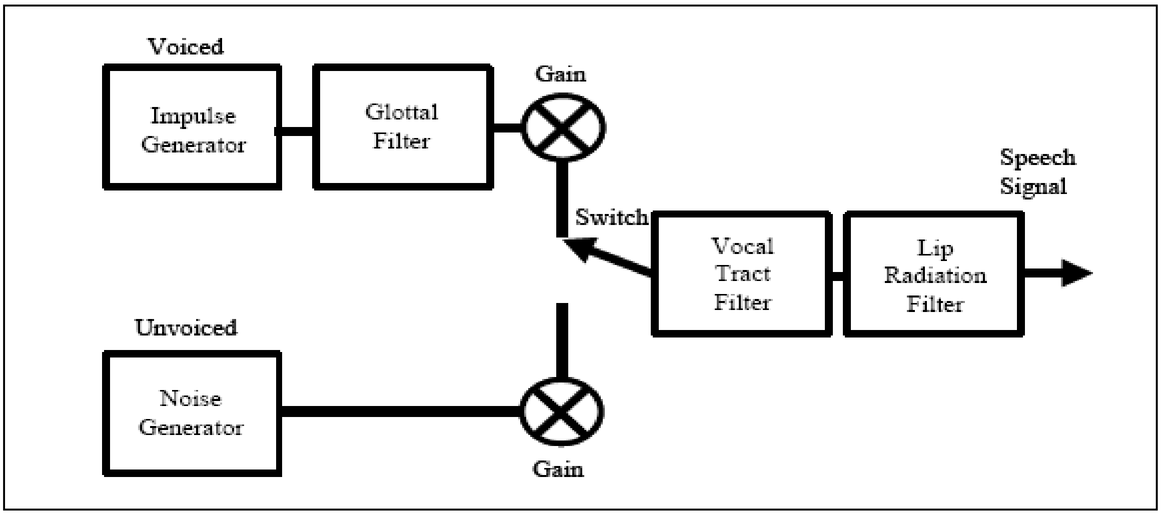


**Fig.9.1:** Block diagram of speech production model

Although this model is accurate, the analysis can be made simpler by replacing the glottal, vocal tract, and lip radiation filters, by a single vocal tract filter as shown in figure 9.2. This model is obtained by collapsing all these separate filters into a vocal tract filter by the convolution operation.

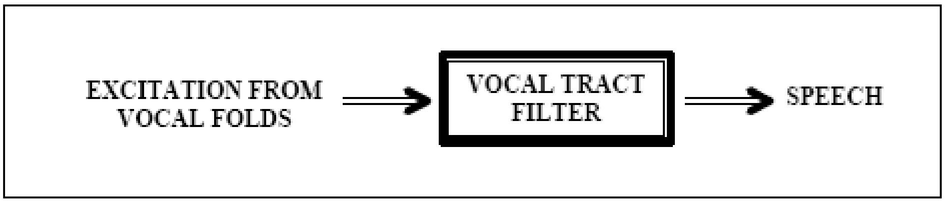


Fig.9.2: Block diagram of source-filter model

An analytical model of this block diagram can be formulated in the following way. According to the conventional source-filter model representation, a speech signal is composed of an excitation source convolved with the vocal tract filter as,

$$s[n] = h[n] \otimes e[n] \qquad \ldots\ldots\ldots\ldots\ldots..(9.1)$$

where, h[n] is the excitation signal, the signal coming from the glottis and e[n] is the vocal tract filter parameter. In the frequency domain (after Fourier transform) this is a product:

$$S(\omega) = H(\omega)\, E(\omega) \qquad \ldots\ldots\ldots\ldots\ldots(9.2)$$

By taking the logarithm on both sides, the equation is converted from multiplication to a simple addition.

$$\log|S(\omega)| = \log\ |H(\omega)E(\omega)|$$

$$\log|S(\omega)| = \log|H(\omega)| + \log|E(\omega)| \ldots\ldots\ldots.(9.3)$$

Then, by taking the inverse discrete Fourier Transform (IDFT) of $\log|S(\omega)|$, the cepstrum is obtained in what is known as the quefrency domain [Deller. J. R., Proakis. J. G. *et al*, 2000].

$$C(q) = \text{IDFT}\{\log|S(\omega)|\}$$

$$C(q) = \text{IDFT}\{\log|H(\omega)|\} + \text{IDFT}\{\log|E(\omega)|\}. \quad \ldots\ldots..(9.4)$$

The cepstrum then allows the excitation signal to be completely isolated from the vocal tract characteristics, because the multiplication in the frequency domain has been converted to an addition in the quefrency domain. Cepstral coefficients, C(q), can be used as features for speech recognition for several reasons. First, they represent the spectral envelope, which is the vocal tract. The vocal tract characteristics are understood to contain information about the phoneme that is produced. Second, cepstral coefficients have the property that they are uncorrelated with one another, which simplifies subsequent analysis [Gold.B. and Morgan. N, 2000], [Deller. J. R., Proakis. J. G. *et al*, 2000]. Third, their computation can be done in a reasonable amount of time. The last and most important reason is that cepstral coefficients have empirically demonstrated excellent performance as features for speech recognition for many years [Deller. J. R., Proakis. J. G. *et al*, 2000].

The first few coefficients are used as the features for acoustic modeling, as they relate more strongly to the vocal tract, which has more smoothly varying spectral characteristics, resulting in lower quefrency characteristics. The higher indexed points correspond more to the excitation, as the excitation has more jagged spectral characteristics, and consequently higher quefrency characteristics. This phenomenon can be seen in figure 9.3 The cepstral values at the beginning represent features of the vocal tract, whereas the humps farther to the right represent the pitch characteristics.

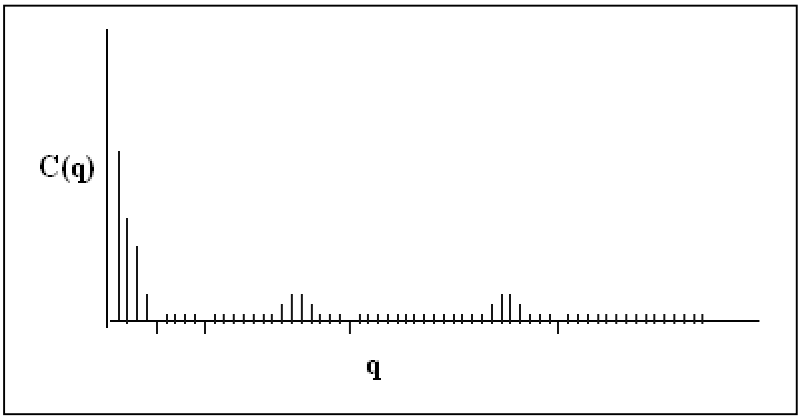


**Fig.9.3:** Cepstrum of a speech signal

### 9.2.1 Mel frequency Cepstral Coefficients (MFCCs)

The most popular form of cepstral coefficients are known as Mel frequency cepstral coefficients (MFCCs). MFCCs are computed in a similar way as the methods described in the previous section, but have been slightly modified. The procedure is as follows:

1. The signal is filtered using a pre-emphasis filter to emphasize the frequency contents.

$$s'[n] = h[n] \otimes s[n]$$

2. The signal is multiplied by overlapping windows and divided into frames usually using 20-30 ms windows with 10-15ms of overlap between windows.

$$s''[n] = s'[n] . w[n]$$

A hamming window is typically used for w[n]

$$w[n] = \begin{cases} 0.54 - 0.46 cos(2\pi n / N), 0 \le n \le N \\ 0, otherwise \end{cases}$$

3. The discrete Fourier Transform (DFT) is taken of every frame s'' followed by the logarithm.

$$S(\omega) = \log \left| DFT\{ s''[n]\} \right|$$

4. Twenty four triangular-shaped filter banks with approximately logarithmic spacing (called Mel banks) are applied to S(ω). A filter bank, in its simplest form, is a set of band pass filters with different frequencies covering the interesting part of the spectrum (the fundamental frequency and the formants). The output of the filters during a frame can be used as features. The center frequencies of the filters can be chosen in several ways. Usually they are set according to some perceptually motivated scale. One commonly used scale is the Mel-scale. It is an empirical scale developed by Stanley Smith Stevens, John Volkman, and Edwin Newman. It was created by modeling human auditory system and has been showed that this model

empirically improves recognition accuracy. The Mel frequency is given by:

$$F_{mel} = 2595 \log_{10}(1+ F_{Hz} / 700)$$

MFCCs are computed by using filter banks. The filter banks consists of triangular filters as shown in figure 9.4. Such filters compute the spectrum around each center frequency with increasing bandwidths. The log energy (S[m]) at the output of each filter is computed afterwards.

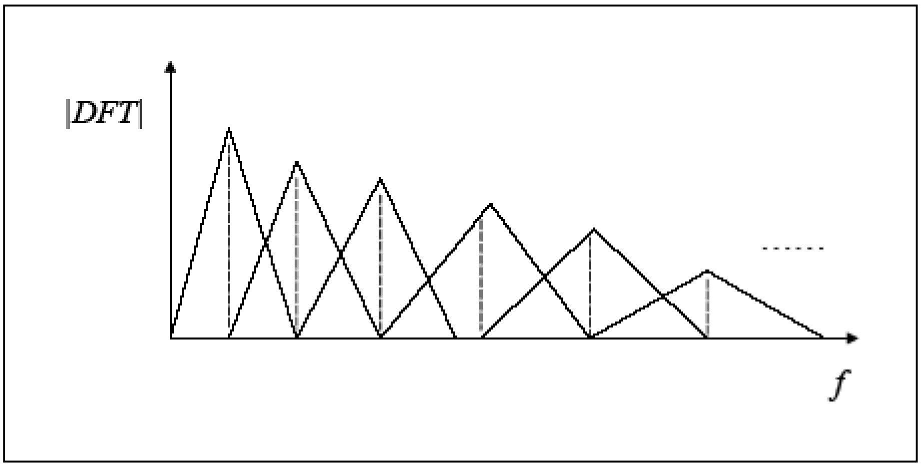


**Fig. 9.4 :** Triangular filters used in the MFCC computation

5. The Discrete Cosine Transform (DCT) is taken to give the cepstral coefficients

$$C[n] = DCT \{ S[m]\}$$

6. The first 12 coefficients, excluding the 0th coefficient are the features that are known as MFCCs ($C_{mel}$)

The advantage of computing MFCCs by using filter energies is that they are more robust to noise and spectral estimation errors.

### 9.3 Simulation Experiments with MFCC feature set as Input Parameter

Here we investigate the recognition capabilities of the Mel Frequency Cepstral Coefficient -based Malayalam vowel recognition system. For this purpose the multi layer feed forward neural network is simulated with the Back propagation learning algorithm. A summary of the parameters used for the simulation experiment is as follows : A constant learning rate, 0.01, is used ( value of learning rate was found to be optimum as 0.01 by trial and error method). The initial weights are obtained by generating random numbers ranging from 0.1 to 1. The number of nodes in the input layer is fixed according to the feature vector size. Since five Malayalam vowels are analyzed in this experiment, the number of nodes in the output layer is fixed as 5. The recognition experiment is repeated by changing the number of hidden layers and number of nodes in each hidden layer. After this trial and error experiment, the number of hidden layers is fixed as three, the number of nodes in the hidden layer is set to Eighteen and the number of *epochs* as 10,000.

The recognition accuracies obtained for the Malayalam vowels based on MFCC feature set using multi layer feed forward neural network classifier are tabulated in Table 9.1. The graphical representation of these recognition results based on different features using neural network is shown in figure 9.5.

| Vowel Number | Vowel Unit | Average Recognition Accuracy (%) |
|---|---|---|
| **1** | അ/Λ/ | 94.4 |
| **2** | **ഇ/I/** | 91.2 |
| **3** | **എ/ae/** | 89.2 |
| **4** | **ഒ/o/** | 93.2 |
| **5** | **ഉ/u/** | 90.4 |
| Overall Recognition Accuracy (%) | | 91.68 |

**Table 9.1**: Recognition Accuracies of Malayalam Vowels based on MFCC feature vector using Neural Network

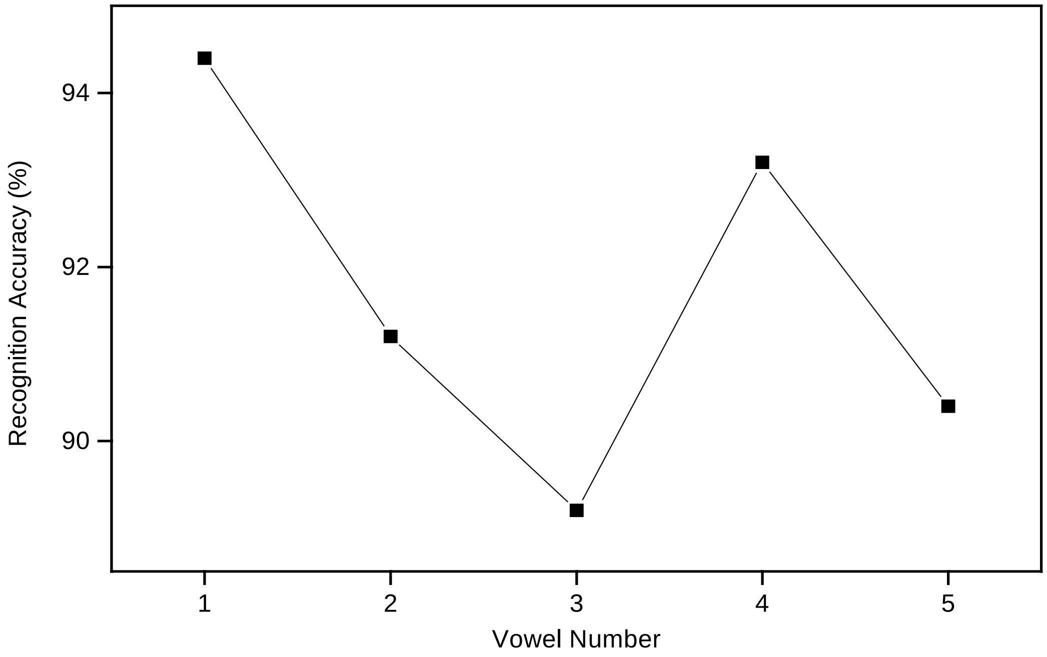


**Fig. 9.5**: Vowel No. Vs. Recognition Accuracies of Malayalam Vowels based on MFCC feature vector using Neural Network

## 9.4 Joint Feature vector

The RPS derived features can be used in unison with the MFCC feature set to create a joint or composite feature vector. The motivation for such a feature vector is that, MFCC feature set has been successful for speech recognition in the past, and utilizing them with the RPS derived feature set will increase classification accuracy, if the information content between the two is not identical. It will be interesting to see how the incorporation of two different sets of features extracted from radically dissimilar processing spaces and methodologies will fuse together to help ascertain the precise information content and discriminatory power of the RPS derived feature set.

From the previous chapter, it is clear that better accuracy is obtained for recognition experiments, when multi layer feed forward neural network classifier is used with modified RPS Distribution Parameter (MRPSDP) as input parameter. Therefore in this work, MFCC feature set is concatenated with the RPS feature vector MRPSDP to make the joint feature set, represented by :

$$y_n = [\ \mathrm{MRPSDP} \mid C_{mel}\ ]$$

And multi layer feed forward neural network is used as the classifier. A block diagram of joint feature vector computation is given in figure 9.6

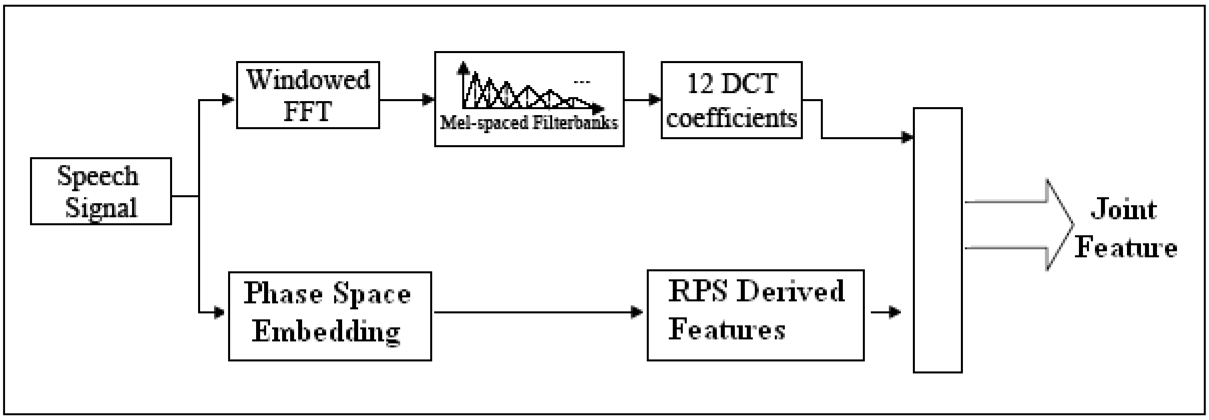


**Fig.9.6:** Block diagram of joint feature vector computation

### 9.5 Simulation Experiments and Results

Present study investigates the recognition capabilities of the above explained joint feature-based Malayalam vowel recognition system. Multi layer feed forward neural network is simulated with the Back propagation learning algorithm. A constant learning rate, 0.001, is used. The initial weights are obtained by generating random numbers ranging from 0.1 to 1. The number of nodes in the input layer is fixed according to the feature vector size and the number of nodes in the output layer is fixed as 5. The recognition experiment is repeated by changing the number of hidden layers and number of nodes in each hidden layer. After this trial and error experiment, the number of hidden layers is fixed as two, the number of nodes in the hidden layer is set to fifteen and the number of *epochs* as 12,000 for obtaining the successful architecture in the present study.

The network is trained using the joint feature vectors extracted for Malayalam vowels. Here we used a set of 250 samples each of the five

Malayalam vowels for iteratively computing the final weight matrix and a disjoint set of vowels of same size from the database for recognition purpose.

The recognition accuracies obtained for the Malayalam vowels based on above said features using multi layer feed forward neural network classifier are tabulated in Table 9.2. The graphical representation of these recognition results based on different features using neural network is shown in figure 9.7.

| **Vowel Number** | **Vowel Unit** | **Average Recognition Accuracy (%)** |
|---|---|---|
| **1** | അ/Λ/ | 100 |
| **2** | **ഇ/I/** | 93.6 |
| **3** | **എ/ae/** | 91.6 |
| **4** | **ഒ/o/** | 96 |
| **5** | **ഉ/u/** | 100 |
| Overall Recognition Accuracy (%) | | 96.24 |

**Table 9.2**: Recognition Accuracies of Malayalam Vowels based on joint feature vector using Neural Network

The overall recognition accuracies obtained for Malayalam vowels using Multi layer feed forward Neural Network with the joint feature vector is 96.24%,

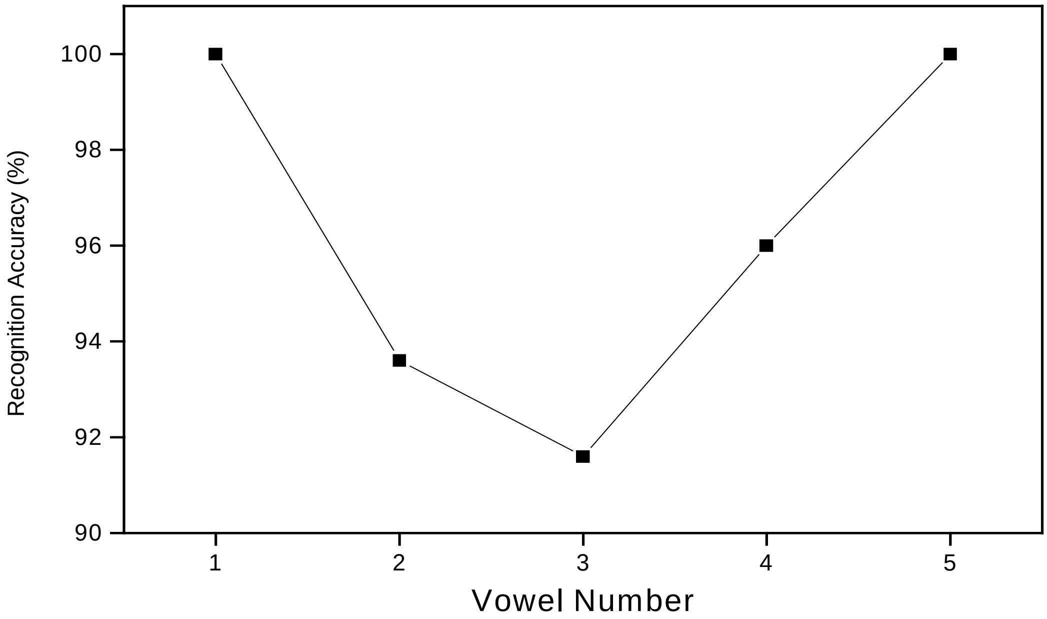


**Fig. 9.7**: Vowel No. Vs. Recognition Accuracies of Malayalam Vowels based on joint feature vector using Neural Network

As explained in the previous chapter, when RPS feature alone is used as the input parameter for neural network, the maximum recognition accuracy obtained is 92.96%. With the hybrid feature vector formed by combing the frequency domain and the Reconstructed Phase Space derived features, the overall recognition accuracy becomes 96.24%. That is by using RPS derived features in unison with traditional MFCC features yield improvement in recognition experiments.

### 9.6 Direct comparisons of the Feature sets

In this study, we have used different feature vectors for the recognition experiments. Here we compare the performance of these feature vectors by simply examining the recognition accuracies. The summary of the performance is tabulated below and is displayed in figure 9.8

| **Vowel Number** | **Vowel Unit** | **Average Recognition Accuracy (%)** | | | |
|---|---|---|---|---|---|
| | | **RPSDP Feature** | **MRPSDP Feature** | **MFCC Feature** | **Joint Feature** |
| **1** | **അ/Λ/** | 96.4 | 97.2 | 94.4 | 100 |
| **2** | **ഇ/I/** | 87.6 | 90 | 91.2 | 93.6 |
| **3** | **എ/ae/** | 82.4 | 86.4 | 89.2 | 91.6 |
| **4** | **ഒ/o/** | 89.6 | 92.4 | 93.2 | 96 |
| **5** | **ഉ/u/** | 96.8 | 98.8 | 90.4 | 100 |
| Overall Recognition Accuracy (%) | | 90.56 | 92.96 | 91.68 | 96.24 |

**Table 9.3**: Comparison of the Performance of Feature sets with Neural Net classifier

As apparent from the table, modifying the Reconstructed Phase Space Distribution Parameter ( RPSDP) with the optimum embedding parameter, boost the recognition accuracy by 2.4%. Also appending the Modified RPS feature with the traditional MFCC feature set, improves the accuracy by 4.56%.

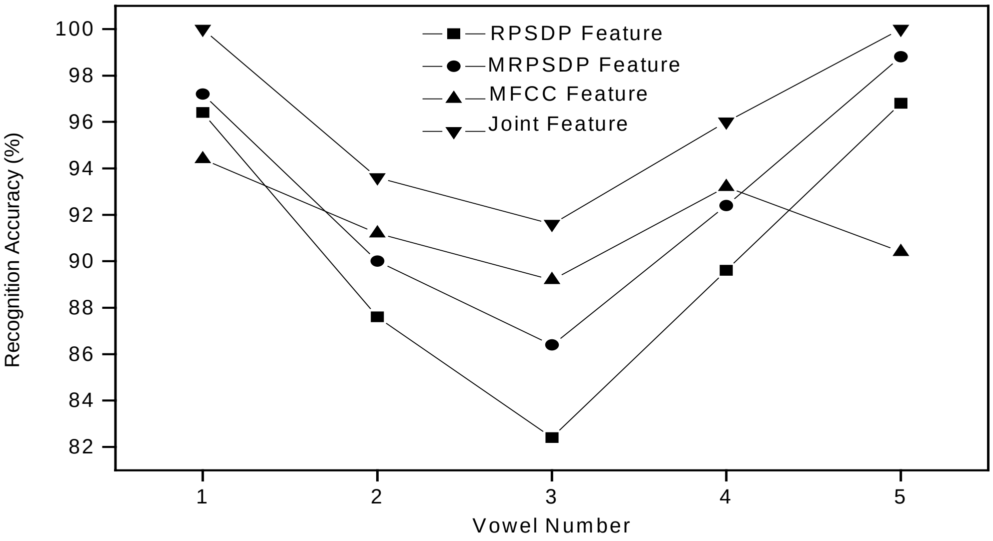


**Fig.9.8**:Comparison of the Performance of Feature sets with Neural Net classifier

## 9.7 Conclusion

Malayalam vowel recognition study based on MFCC feature set and the joint feature vector developed by combining frequency domain and Reconstructed Phase Space derived features are presented in this chapter. Multi layer feed forward neural network with error back propagation algorithm is implemented and tested with these feature sets extracted from the vowels. An overall recognition accuracy of 96.24% is obtained for the simulation experiments with the hybrid feature set. This shows that when the joint feature vector is used as input parameter, there is a significant boost in the recognition accuracy than that is obtained when RPS derived feature or MFCC feature alone is used. This result suggests that the frequency domain features and the RPS derived features contain different discriminatory information. The performance of different feature vectors are compared and shown graphically.

## Chapter 10

# Conclusion

Speech processing using a dynamical systems approach has been presented in this thesis. This is a novel approach that extracts features from the time domain using Reconstructed Phase Spaces. The study of nonlinear dynamical system shows that the RPS is able to capture the nonlinear information of underlying system that cannot be captured by frequency domain analysis alone.

A low cost data acquisition system is developed for database creation. Here a multimedia based system is converted into an economically viable data acquisition system by connecting an 8th order anti aliasing presampling filter prior to A/D converter of the sound card. A speech database of short vowels in Malayalam is created using the data acquisition system developed.

By measuring nonlinear invariant parameters of Malayalam vowels, it is examined that whether speech (especially vowel sounds) is chaotic.The non-integer attractor dimension and non-zero value of Kolmogorov entropy confirm the contribution of deterministic chaos to the behavior of speech signal. These parameters quantify the chaotic behaviour of the speech signal. As far as recognition application is concerned, we want to go for more robust and computationally simple parameters. Phase space is a tool to analyze the underlying dynamics of a system. It can be exploited as a powerful signal processing domain.

Reconstructed Phase Space is generated for vowel sounds by the method of time delay embedding. From the Reconstructed Phase Space, a promising parameter called Reconstructed Phase Space Distribution Parameter (RPSDP) is extracted. With this parameter we can analyze the geometric structure of the reconstructed attractor. The RPSDPs are found to be similar for same vowel sounds and differ from vowel to vowel. Hence they are further used in the recognition experiments. Here, an entirely different way of viewing the speech processing problem is presented, and offering an opportunity to capture the nonlinear characteristics of the acoustic structure.

Fundamental frequency, $f_0$ is the lowest frequency component, in the signal, which relates well to most of the other frequency components. A general method for pitch estimation using Reconstructed Phase Space in two dimensions is presented. Here methodologies originally developed for analyzing chaotic time series have been successfully applied to pitch determination problem. The proposed new method does not suffer from the limitations of other short-term pitch-estimation techniques. The algorithm is very straightforward and flexible. The experimental results show that the pitch estimated using Reconstructed Phase Space features agrees with that obtained using conventional Pitch Detection Algorithms.

The two parameters of an RPS, time lag and embedding dimension, play an important role both theoretically and practically in building a speech recognition system based on RPS features. The problem of choosing the optimal time delay and the minimum embedding dimension for the reconstruction of phase space using the method of delays are addressed in the thesis. The optimal delay depends upon the details of the time series as well as the dynamics of the underlying system. A simple procedure that quantifies expansion from the identity line of embedding space is developed for choosing proper time delay. Such a procedure may be more useful in the estimation of the proper time delay. For determining the minimum embedding dimension we have used Cao's method, which does not contain any subjective parameters except time delay for the embedding and is computationally efficient. This method gives consistent results with Malayalam vowels. With these optimum-embedding parameters, Reconstructed Phase Space Distribution Parameter (RPSDP) is modified as Modified Reconstructed Phase Space Distribution Parameter (MRPSDP).

The recognition experiments of Malayalam vowels based on the above discussed features are conducted using different classifiers such as k-NN classifier and Neural network. The credibility of the parameters is tested with the k-NN classifier. A multi layer feed forward neural network with error back propagation algorithm is implemented and tested using RPSDP features and MRPSDP features extracted from the vowels. The highest recognition accuracy (92.96%) is obtained with MRPSDP feature using neural network classifier. These results specify the discriminatory strength of the Reconstructed Phase Space derived features for isolated Malayalam vowel classification experiments.

The nonlinear RPS derived features are then combined with the traditional MFCC feature set to achieve an improvement in the accuracy of recognition experiments. Multi layer feed forward neural network with error back propagation algorithm is implemented and tested with this hybrid feature set extracted from the vowels. An overall recognition accuracy of 96.24% is obtained for the simulation experiments. When the joint feature vector is used as input parameter, there is a significant boost in the recognition accuracy than that is obtained when RPS derived feature or MFCC feature alone is used. This result suggests that the frequency domain features and the RPS derived features contain different discriminatory information. The entire system is developed and implemented using MATLAB 7.

This thesis has presented a novel technique for speech processing using features extracted from phase space reconstructions. The methods have a sound theoretical justification provided by the nonlinear dynamics literature. The specific approach transfers the analytical focus from the frequency domain to the time domain, which presents a unique way of viewing the speech recognition problem and offers an opportunity to capture the nonlinear dynamical information present in the speech production mechanism.

**Future Work**

Due to the nature of the technique, the features are based in the time domain, and therefore the dynamic range of the time series affects the range of the data in the RPS. The dynamic range or amplitude of  the signal is known to be irrelevant to the classification process, because the amplitude is affected by the external conditions such as  the distance of the speaker from the microphone during the data collection. One advantage of the MFCC features is that they are totally invariant to this issue, because the energy is normalized out on a frame-by-frame basis. In the case of the RPS derived features though, the problem is partially solved by using the normalization procedure, this issue is of particular concern when performing continuous recognition, because there is no way to normalize the data on a phoneme-by-phoneme basis, since the time boundary information would not be present. This concern seriously hampers effective implementation for a continuous speech task. Future work could resolve this problem by discovering a robust method to make this approach independent of amplitude scaling effects.

The other related issue is that of computing an energy measure. Again, the MFCC features incorporate an energy measure by computing it for each frame. For the proposed RPS derived features, since there is no analysis window, there is no way to compute a meaningful energy measure that can be incorporated directly into the feature vector. One method would be to compute the energy over an entire phoneme. But, once again, for continuous recognition, phoneme boundaries are unknown, and computing such a measure may be difficult.

In addition to the isolated vowel classification experiments presented and described here, the entire phonemes should be examined. Thus this approach can be applied for the task of continuous speech recognition using the RPS derived features. As far as continuous speech recognition task is concerned, the temporal variations of the feature vectors also should be incorporated. These issues must be addressed in order for the RPS derived features to have long-term success for speech recognition applications..

In conclusion, this work has extended the fundamental understanding of the speech processing and simultaneously expanded the knowledge of the nonlinear techniques for classification applications. It strongly deviates from mainstream research in speech processing.  Reconstructed Phase Space analysis is an attractive research avenue for increasing speech recognition accuracy as demonstrated through the results, and future work will determine its overall feasibility and long-term success for both isolated and continuous speech recognition applications.

# REFERENCES

[1] Abarbanel.H.D.I, “Analysis of Observed Chaotic Data”, *Springer Verlag, New York,* 1996.

[2] Abarbanel.H.D.I, Brown.R, Sidrorowich.J.J, and Tsimring.L.S, “The analysis of observed chaotic data in physical systems”, *Tec. Mod. Phys.*, Vol. 65, 1993

[3] Ahn.R and Holmes.W.H, “Voiced/unvoiced/silence classification of speech using 2-stage neural networks with delayed decision input”, *Proc. Fourth International Symposium on Signal Processing and its Applications. ISSPA 96,* Queensland Univ. Technol, Brisbane, Australia, 1996.

[4] Albano.A.M, Muench.J, Schwartz.C, Mees.A.I and Rapp.P.E, “Singular-value decomposition and the Grassberger-Procaccia algorithm”, *Phys. Rev. A,* vol. 38, pp. 3017-3025, 1988.

[5] Alligood.K, Sauer.T, and Yorke.J, “Chaos: An Introduction to Dynamical Systems”, *Spinger–Verlag,* New York, 1997.

[6] Ana I. Garcia Moral, Ruben Solera Urena, Carmen Pelaez Moreno and Fernando Dıaz de Maria, “Hybrid models for automatic speech recognition: a comparison of classical ANN and kernel based methods”, *Proc. Int. Conf. on Non-Linear Speech Processing,* NOLISP, 2007

[7] Bagshaw.P.C, Hiller.S.M and Jack.M.A, “Enhanced pitch tracking and the processing of $f_0$ contours for computer aided intonation teaching”, *Proceedings of the 3rd European Conference on Speech Communication and Technology,* vol.2, pp. 1003-1006, 1993.

[8] Baken.R.J, “Clinical Measurement of Speech and Voice”, *Taylor and Francis Limited*, London, 1987.

[9] Banbrook.M and McLaughlin.S, “Is speech chaotic?,” *Proc. of IEE Colloquium on Exploiting Chaos in Signal Processing*, pp. 8/1-8/8, 1994.

[10] Banbrook.M, McLaughlin.S and Mann.I, "Speech characterization and synthesis by nonlinear methods," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 1-17, 1999.

[11] Birgmeier.M, “A fully Kalman-trained radial basis function network for nonlinear speech modeling”, *Proc. IEEE int. Conf. Neural Networks, ICNN’95*, Perth, 1995.

[12] Birgmeier.M, “Nonlinear prediction of speech signals using radial basis function networks”, *EUSIPCO*, vol. 1, pp. 459-462, 1996

.

[13] Bogner.R.E and Li.T, “Pattern search prediction of speech” *Proc. Int. Conf. Acoustics, Speech & signal processing*, Glasgow, vol.1 pp.180-183, 1989.

[14] Bogner.R.E, “Signal prediction by pattern search”, *J. Institution Electron. Telecomm. Engineers*, vol. 34, no 1, pp.43-49, 1988.

[15] Brookes.D.M and Naylor.P.A, “Speech production modelling with variable glottal reflection coefficient,” *Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 671 – 674, 1988.

[16] Broomhead.D.S and King.J.P, “Extracting qualitative dynamics from experimental data”, *Physica D*, vol.20, pp. 217-224, 1986.

[17] Cao.L, “Practical method for determining the minimum embedding dimension of a scalar time series”, *Physica D*, vol. 110, pp. 43-50, 1997.

[18] Casdagli.M ,Des Jardins.D, Eubank.S, Farmer.J.D, Gibson.J, Theiler.J, and Hunter.N “Nonlinear modelling of chaotic time series: theory and applications”, in *applied chaos* (J. Kim and J.Stringer ed.), pp.335-380, New York and Chichester, *Wiley* 1992

[19] Casdagli.M, “Nonlinear prediction of chaotic time series”, *Physica D*, Vol. 35 D, pp.335-356,1989.

[20] Casdagli.M, Eubank.S, Farmer.J.D and Gibson.J, “State space reconstruction in the presence of noise”, *Physica D*, vol. 51,1991.

[21] Chandrasekhar. C and Yegnanarayana. B, “ Recognition of Stop-Consonant-Vowel (SCV) segments in Continuous Speech using Neural Network Models”, *Jour. Institution of Electronics and Tele-communication Engineers (IETE)*, vol.42, pp.269-280, 1996.

[22] Chandrasekhar. C, “Neural Network Models for Recognition of Stop-Consonant-Vowel (SCV) Segments in Continuous Speech”, *Ph.D thesis*, Department of Computer Science and Engg., IIT, Madras, India, 1996.

[23] Connor. J. D. O, “Phonetics”, *Penguin Books*, 1991.

[24] Crutchfield.J, Farmer.J, Packard.N, and Shaw.R, “Chaos,” *Scientific American*, vol. 255, pp. 38 – 49, 1986.

[25] Davis.K.H, Biddulph.R, and Balashek.S, “Automatic recognition of spoken digits”, *Journal of the Acoustical Society of America*, 1952.

[26] Deller.J.R, Hansen.J.H.L and Proakis.J.G, “Discrete-time processing of speech signals”, *Second ed., IEEE Press*, New York, 2000.

[27] Dhananjaya.N, Guruprasad.S and Yagnanarayana.B, “Speaker Segmentation Features and Neural Network Models”, *Proceedings of 11th International Conference , ICONIP 2004,* Culcutta,India, 2004.

[28] Diaz de Maria.F and Figueiras Vidal.A.R, “Radial basis functions for nonlinear prediction of speech in analysis-by-synthesis coders”, *Proc. IEEE workshop on nonlinear. Signal and image processing NSIP’95,* Haldiki, 1995.

[29] Duda.R.O and Hart.P.E, “Pattern classification and scene analysis”, *Wiley Interscience, New York*, 1973.

[30] Duda.R.O, Hart.P.E and Stork.D.G, “Pattern classification”, *John Wiley & Sons, Second Edition,* 2001.

[31] Dudley. H.W, “The carrier nature of speech”, *Bell Systems Technical Journal*, vol.19, pp. 495- 513, 1940.

[32] Eckmann.J.P and Ruelle.D, “Ergodic theory of chaos and strange attractors,” *Review of Modern Physics*, vol. 57, pp. 617 – 656, 1985.

[33] Fant.G, “Acoustic Theory of Speech Production”, *Mouton,* 1960.

[34] Farmer.J.D and Sidorowich.J.D, “Exploiting chaos to predict the future and reduce noise”, *Evolution, Learning and Cognition* (Y.Lee ed.), pp.277-330, Singapore, World scientific, 1988.

[35] Farmer.J.D, “Information dimension and the probabilistic structure of chaos”, *Z. Naturforsch,* vol. 37a, pp.1304-1313, 1982.

[36] Flanagan.J.L, “Speech analysis and Perception”, *SpringerVerlag,* second edition, 1965.

[37] Forgie. J.W and Forgie.C.D, “Results obtained from a vowel recognition computer program”, *Journal of the British Institution of Radio Engineering*, 1959.

[38] Fraser.A.M and Swinney.H.L, “Independent coordinates for strange attractors from mutual information”, *Phys. Rev. A*, vol. 33, pp.1134-1140, 1986.

[39] Friedman.M. and Kandel.A., “Introduction to pattern recognition statistical, structural, neural and fuzzy logic approach”, *World Scientific*, 1999.

[40] Gersho.A. “Optimal nonlinear interpolative vector quantization”, *IEEE trans. Communication*. Vol. 38, no. 9 , pp.1285-1287, 1989.

[41] Gersho.A and Gray.R.M, “Vector Quantization and signal compression”, *Ed. Kluwer*, 1992.

[42] Gimson.A, “An Introduction to the Pronunciation of English”, *Edward Arnold Ltd.*, 1972.

[43] Gold.B and Morgan.N, “Speech and Audio Signal Processing”, *John Wiley & Sons Inc.*, New York, 2000

[44] Grassberger.P and Procaccia.I, “Estimation of the kolmogorov entropy from a chaotic signal”, *Phys. Rev. A*, Vol. 28, no.4, pp. 346-352, 1983.

[45] Haykin.S and Li.L “Nonlinear adaptive prediction of non stationary signals”, *Signal Process.*,vol.43 , pp.526-535, 1995.

[46] Haykin.S, “Neural Networks: A comprehensive foundation”, *Prentice Hall of India Pvt.Ltd*, 2004.

[47] Hebb.D.O, “The organization of behavior”, *A Neuropsychological Theory, Wiley-New York*, 1949.

[48] Heggar.R and Kantz.H, “Embedding of sequences of time intervals”, *Europhys. Lett*, vol. 38, pp. 267-275, 1999.

[49] Hegger.R, Kantz.H and Matassini.L, "Denoising human speech signals using chaoslike features," *Physical Review Letters*, vol. 84, pp. 3197-3200, 2000.

[50] Hegger.R, Kantz.H and Matassini.L, "Noise reduction for human speech sig-nals by local projections in embedding spaces", *IEEE*

*Transactions on Circuits and Systems - I: Fundamental Theory and Applications*, vol. 48, pp. 1454-1461, 2001.

[51] Hess.W.J, "Pitch and voicing determination", *Advances in speech signal processing*, M. M. Sondhi and S. Furui, Marcel Dekker, Eds., pp. 3-48. Marcel Dekker, Inc., New York, 1992.

[52] Hess.W.J, "Pitch Determination of Speech Signals – Algorithms and Devices", *Springer*, Berlin, 1983.

[53] Hilborn.R, "Chaos and Nonlinear Dynamics", *Oxford: Oxford University Press*, 1994.

[54] Hussain.A, "Novel Artificial Neural-Network Architectures and Algorithms for Non-linear Dynamical System Modelling and Digital Communications Applications", *PhD Thesis*, University of Strathclyde, Glasgow, UK, 1996.

[55] Hutton.L.V, "Using statistics to assess the performance of neural network classifiers", Johns-Hopkins-APL Technical Digest, vol.13, no.2, pp. 291-300, 1992.

[56] Ishizaka.K and Flanagan.J.L, "Synthesis of voiced sounds from a two–mass model of the vocal chords," *Bell System Technical Journal*, vol. 51, pp. 1233 – 1268, 1972.

[57] Itakura.F, "Minimum prediction residual applied to speech recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no.1, pp. 67-72, 1975.

[58] Janssen.R.D.T, Fanty.M and Cole.R.A, "Speaker-independent phonetic classification in continuous English letters", *International Joint Conference on Neural Networks, IJCNN-91,* Seattle, IEEE, New York, NY, USA, 1991.

[59] Jinjin Ye, Michael T. Johnson, Richard J. Povinelli, "Phoneme Classification over Reconstructed Phase Space using Principal Component Analysis", *ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP)*, Le Croisic, France, pp.11-16, 2003

[60] Jinjin Ye, Michael T. Johnson, Richard J. Povinelli, "Phoneme classification using naïve bayes classifier in reconstructed phase space", *Proc. IEEE Signal rocessing Society 10th Digital Signal Processing Workshop,* pp. 2.2, 2002.

[61] Kantz, H. and Schreiber.T, “Nonlinear Time Series Analysis”, *Cambridge University Press*, UK, 2003.

[62] Kearney.M.J and Stark.J, “An introduction to chaotic signal processing,” *GEC Journal of Research*, vol. 10, no. 1, pp. 52 – 58, 1992.

[63] Kennel.M.B, Brown.R and Abarbanel.H.D.I, “Determining embedding dimension for phase-space reconstruction using a geometrical construction”, *Physical Review A*, vol.45, pp. 3403 – 3411, 1992.

[64] Kevin M. Indrebo, Richard J. Povinelli, Michael T. Johnson, “Third-Order Moments of Filtered Speech Signals For Robust Speech Recognition”, *Int. Conf. on Non-Linear Speech Processing* (NOLISP), Barcelona, Spain, pp.151-157, 2005.

[65] Ki-Seok-Kim and Hee-Yeung-Hwang, “A study on the speech recognition of Korean phonemes using recurrent neural network models”,*Transactions-of-the-Korean-Institute-of-Electrical-Engineers,* vol.40, no.8, pp.782-791, 1991.

[66] Koizumi.T, Taniguchi.S, and Hiromitsu.S, “Glottal source–vocal tract interaction,” *Journal of the Acoustical Society of America*, vol. 78, pp. 1541 – 1547, November 1985.

[67] Koizumi.T, Taniguchi.S, and Hiromitsu.S, “Two–mass models of the vocal cords for natural sounding voice synthesis,” *Journal of the Acoustical Society of America*, vol. 82, pp. 1179 – 1192, 1987.

[68] Kostelich.E and Schreiber.T, “Noise reduction in chaotic time series: a survey of common methods”, *Physical Review E*, vol. 48, pp. 1752-1763, 1993.

[69] Kubin.G, "Nonlinear speech processing," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds.: Elsevier Science, 1995.

[70] Kumar.A and Gersho.A, “LD-CELP speech coding with nonlinear prediction”, *IEEE Signal Processing letters*, Vol. 4, no. 4, pp.89-91, 1997.

[71] Kumar.A and Mullick.S.K, "Nonlinear dynamical analysis of speech," *Journal of the Acoustical Society of America*, vol. 100, pp. 615-629, 1996.

[72] Ladefoged and Peter, “Vowels and Consonants: An Introduction to the Sounds of Languages”, Blackwell, 2000.

[73] Langi.A and Kinsner.W, “Consonant characterization using correlation fractal dimension for speech recognition”, *Communications, Power, and Computing. Conf. Proc. IEEE*, New York, USA; 1995

[74] Lapedes.A and Farber.R, “How neural nets work”, *Evolution, learning and cognition, World scientific*, pp.231-346, 1988.

[75] Lawrence Rabiner and Bing Hwang Juang, “Fundamentals of speech recognition”, *Prentice Hall*, 1993.

[76] Liang.C, Yanxin.C, and Xiongwei.Z, "Research on speech recognition on phase space reconstruction theory," presented at *Advances in Multimodal Interfaces-ICMI 2000*, Berlin, Germany, 2000.

[77] Liebert.W and Schuster.H.G, “Proper choice of the time delay for the analysis of chaotic time series”, *Phys. Lett. A*, vol.142, pp.107-114, 1989.

[78] Lindau and Mona, “Vowel features”, *Language*, vol.54, pp. 541–563, 1983.

[79] Lindgren.A.C, Johnson.M.T, and Povinelli.R.J, “Speech recognition using Reconstructed phase space features,” *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 61-63, Hong Kong, 2003.

[80] Lorenz.E.N, “Atmospheric predictability as revealed by naturally occurring analogues”, *J.Atmospheric Sciences*, vol. 26, pp.636-646, 1969.

[81] Ma.Y.G.Wei.N, “Speech coding with nonlinear local prediction model”, pp. 1101-1104, vol. 2, ICASSP, 1998.

[82] Mah.R.S.H and Chakravarthy.V, “Pattern recognition using Artificial Neural Network”, *Computers and Chemical Engineering*, vol. 16, no.4, pp. 371-378, 1992.

[83] Mann.I and McLaughlin.S, “A nonlinear algorithm for epoch marking in speech signals using poincaré maps”, *Proc. of the 9th European Signal Processing Conference, EUSIPCO*, Vol.2 pp.701- 704, 1998.

[84] Mann.I and McLaughlin.S, “Synthesizing natural sounding vowels using a nonlinear dynamical model”, *IEEE Trans. Signal Processing*, vol. 81, pp. 1743-56, 2001.

[85] Marcos Faundez-Zanuy “On the Usefulness of Linear and Nonlinear Prediction Residual Signals for Speaker Recognition”, *Proc. Int. Conf. on Non-Linear Speech Processing*, NOLISP, 2007.

[86] Martinerie.J.M, Albano.A.M, Mees.A.I and Rapp.P.E, “Mutual information, strange attractors, and the optimal estimation of dimension”, *Phys. Rev. A*, vol. 45, pp. 7058-7066, 1992.

[87] McCulloch.W.S and Pitts.W, “A logical calculus of the ideas immanent in nervous activity”, *Bulletin of Mathematical Biophysics*, vol.5, pp.115-133, 1943.

[88] Michael T. Johnson, Andrew C. Lindgren, Richard J. Povinelli, Xiaolong Yuan, “Performance of Nonlinear Speech Enhancement using Phase Space Reconstruction”, *Int. Conf. on Acoustics, Speech and Signal Processing*, Hong Kong, China, vol. I, pp.872-875, 2003.

[89] Minsky.M.L and Papert.S.A, “Perceptrons: An introduction to computational geometry”, *MIT Press, Cambridge-M.A.*, 1988.

[90] Moore.K.L, “Artificial neural networks”, *IEEE-Potentials*. vol.11, no.1, p.23-28, 1992.

[91] Mumolo.E, and Francescato.D, “Adaptive predictive coding of speech by means of volterra predictors”, *Workshop on nonlinear digital signal processing*. Tampere, pp. 2.1-4.1, 1993

[92] Mumolo.E, Carini.A and Francescato.D “ADPCM with nonlinear predictors” *Signal processing VII: Theories and applications*, Vol 1, pp.387-390, Ed. Elsevier, 1994.

[93] Narayanan.N.K and Sridhar.C.S, “Parametric Representation of the Dynamical Instabilities and Deterministic Chaos in Speech Signals”, *Proc. Symposium on Signals, Systems and Sonars*, NPOL, Cochin, 1988.

[94] Narayanan.N.K, “Voiced / Unvoiced classification using Second Order attractor dimension and Second Order Kolmogorov Entropy of Speech Signals”, *J. Acoust. Soc. India*, Vol XXVII, p.p 181-185, 1999.

[95] Narayanan.S.S and Alwan.A.A, “A nonlinear dynamical systems analysis of fricative consonants”, *Journal of the Acoustical Society of America*, vol. 97, pp. 2511-2524, 1995.

[96] Olson.H.F and Belar.H, “Phonetic typewrite”, *Journal of the Acoustical Society of America*, vol. 28, pp.1072-1081,1956.

[97] Ott.E, “Chaos in Dynamical Systems”, *Cambridge University Press*, Cambridge, 1993.

[98] Ott.E, and Sauer.T, “Coping with Chaos”, Wiley, New York,1994.

[99] Packard.N.H, Crutchfield.J.P, Farmer.J.D and Shaw.R.S, “Geometry from a time series”, *Phys. Rev. Lett.* Vol. 45, pp.712-717, 1980.

[100] Parker.T.S and Chua.L.O, “Chaos; a tutorial for engineers,” *Proceedings of the IEEE,* vol. 75, pp. 982 – 1008, 1987.

[101] Pernkopf.F, “Bayesian network classifiers versus selective k-NN classifier”, *Pattern Recognition*, vol.38, pp.1-10, 2005.

[102] Personnaz.L and Dreyfus.G, “Neural networks: state of the art and future prospects”, *Revue-Generale-de-l'Electricite*. no.5, pp.27-34, 1990.

[103] Petry.A, Augusto.D, and Barone.C, “Speaker Identification using nonlinear dynamical features”, *Chaos, Solitons, and Fractals*, vol. 13, pp. 221-231, 2002.

[104] Pitsikalis.V and Maragos.P, “Some advances on speech analysis using chaotic models,” *Proc. ISCA Tutorial and Research Workshop on Nonlinear Speech Processing (NOLISP)*, La Croisic, France, 2003.

[105] Pitsikalis.V and Maragos.P, “Speech analysis and feature extraction using chaotic models,” presented at *EEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, 2002.

[106] Povinelli.R.J, Bangura.J.F, Demerdash.N.A.O, and Brown.R.H, “Diagnostics of bar and end-ring connector breakage faults in poly phase induction motors through a novel dual track of time-series data mining and time-stepping coupled festate space modeling," *IEEE Transactions on Energy Conversion*, vol. 17, pp. 39-46, 2002.

[107] Povinelli.R.J, Michael T. Johnson, Andrew C. Lindgren, Felice M. Roberts, Jinjin Ye, “Statistical Models of Reconstructed Phase Spaces

for Signal Classification", *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2178-2186.,2006.

[108] Priestley.M.B, "Non-linear and non-stationary time series analysis",. *Academic press*, 1988.

[109] Rabiner.L.R., "Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, pp. 257-286, 1989.

[110] Rabiner.L.R., Levinson.S.E., A.E.Rosenberg, and J. G. Wilpon, "Speaker Independent Recognition of Isolated Words Using Clustering Techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 336-349, 1979.

[111] Rabiner.L.R and Juang.B.H, "Fundamentals of Speech Recognition", *Prentice Hall*, 1993.

[112] Rabiner.L.R and Schafer.R.W, "Digital Processing of Speech Signals", *Prentice-Hall*, Eaglewood Cliffs, N.J.,1978.

[113] Ray.A.K. and Chatterjee.B, "Design of a nearest neighbour classifier system for Bengali character recognition", *Journal of Inst. Elec. Telecom. Eng.*, vol.30, pp.226-229, 1984.

[114] Reddy.D.R., "An approach to computer speech recognition by direct analysis of the speech wave," Computer Science Dept., Stanford University Technical Report No. C549, Sept. 1966.

[115] Reininger.H and Wolf.D, "Nonlinear prediction of stochastic processes using neural networks", *Signal processing V: theories and applications, Ed. Elsevier,* pp.1623-1626, 1990.

[116] Roberts.F.M, Povinelli.R.J, and Ropella.K.M, "Identification of ECG arrhythmias using phase space reconstruction," *Proc. Principles and Practice of Knowledge Discovery in Databases (PKDD'01)*, pp. 411-423Freiburg, Germany, 2001.

[117] Rosenblatt.F. "Two theorems of statical separability in the perception", *Symposium on the Mechanization of Thought Process,* pp. 421-456, 1959.

[118] Sada Siva Sarma.A, Strune.H.W, Rajesh Varma and Agarwal.S.S, "Recognition of Hindi Consonants Using Time Delay Neural

Network", *Jour. Acoustical Society of India*, vol.24, pp.III-10.1-10.6, 1996.

[119] Sauer.T, Yorke.J.A and Casdagli.M, "Embedology." *Journal of Statistical Physics*, vol. 65, no.3, pp. 576-616, 1991.

[120] Schoentgen.J, "Non–linear signal representation and its application to the modelling of the glottal waveform," *Speech Communication*, vol. 9, pp. 189 – 201, 1990.

[121] Sicuranza.G.L, "Quadratic filters for signal processing", *Proc. IEEE*, Vol. 80, pp.1263-1285, 1992

[122] Singer.A.C, Wornell.G.W and Oppenheim.A.V, "Nonlinear autoregresive modelling and estimation in the presence of noise", *Digital signal processing*, vol. 4, pp.207-221, 1994.

[123] Singer.A.C, Wornwell.G.W and Oppenheim.A.V, "Codebook prediction: a nonlinear signal model paradigm", *Proc. Int. Conf. Acoustics, Speech & signal processing*, San Francisco, vol. 5, pp.325-328, 1992.

[124] Steinecke.I and Herzel.H, "Bifurcations in an asymmetric vocal–fold model," *Journal of the Acoustical Society of America*, vol. 97, pp. 1874 – 1884, 1995.

[125] Sunil Kumar.R.K, "Vowel Phoneme recognition from zerocrossing based parameters using Artificial Neural Networks", Ph.D Thesis, University of Calicut, 2002.

[126] Takens, F, "Detecting strange attractors in turbulence", in *Lecture Notes in Mathematics*, Vol. 898, Eds. D.A.R and L.S.Young, Springer, Berlin, 1981.

[127] Takens.F, "Detecting strange attractors in turbulence," *Proc. Dynamical Systems and Turbulence*, Warwick, pp. 366-381, 1980.

[128] Talkin.D, "A robust algorithm for pitch tracking (RAPT)", *Speech Coding and Synthesis*, W.B.Kleijin and K.K.Paliwal, Eds., pp 497-518. Elsevier Science, 1995.

[129] Teager.H.M.and Teager.S.M, "Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract", *Speech Production and Speech Modelling, W.J. Hardcastle and A. Marchal, Eds., NATO Advanced Study Institute Series D*, vol.55, Bonas, France, 1989.

[130] Thomas Parsons, “Voice and speech processing”, *McGraw Hill*, New York, 1987.

[131] Thyssen.J, Nielsen.H and Hansen.S.D, “Non-linear short term prediction in speech coding”, *Proc. Int. Conf. Acoustics, Speech & signal processing*, Australia. pp.I-185, I-188, 1994.

[132] Tishby.N, “A dynamical systems approach to speech processing”, *Proc. Int. Conf. Acoustics, Speech & signal processing,* Albuquerque, 1990.

[133] Tong.H, “Non-linear time series- a dynamical system approach”, *Oxford University press,* 1990.

[134] Tou.J.T and Gonzalez.R.C, “Pattern recognition principles”, *Addison-Wesley, London,* 1974.

[135] Townshend.B, “Nonlinear prediction of speech signals”, *Nonlinear modeling and forecasting, SFI studies in the sciences of complexity,* Vol. Proceedings XII, pp.433-453, Addison-Wesley 1992

[136] Townshend.B, “Nonlinear prediction of speech”, *Proc. Inl. Conf. Acoust. Speech, sign. Proc.*, Toronto, pp.425-428, 1991.

[137] Velayudhan.S, “Vowel Duration in Malayalam – as Acoustic Phonetic Study”*, Dravidian Linguistic Association of India,* 1971.

[138] Vintsyuk.T.K, “Speech discrimination by dynamic programming”, *Kibernetika*, vol. 4, no. 2, pp. 81-88, 1968.

[139] Wang.S, Paksoy.E and Gersho.E, “Performance of nonlinear prediction of speech”, *Proceedings of the international conference on Spoken Language processing,* Kobe, pp.29-32, 1990.

[140] Wei Gang, Lu Yiqing and Quyang Jingzheng, “Chaos and fractal theories for speech signal processing”, *Acta-Electronica-Sinica*. vol.24, no.1,1996.

[141] Weibel.A, Hanazava.T, Hinton.G, Shikano.K and Lang.K, “Phoneme recognition : Neural networks vs. Hidden Markov Models”, *IEEE Transactions on Neural Networks,* vol.8, no.2, pp. 107-110, 1988.

[142] William Huang, Richard Lippmann and Ben Gold, “A Neural Net approach to Speech Recognition”, *IEEE Transactions on Neural Networks,* vol.8, no.2, pp. 99-100, 1988.

[143] Wu.L and Niranjan.M, “On the design of nonlinear speech predictors with recurrent nets” Proc. *Int. Conf. Acoustics, Speech & signal processing,* pp.II-529 II-532,Adelaide, 1994.

[144] Wu.L, Niranjan.M and Fallside.F, “Fully vector quantized neural network based code excited nonlinear predictive speech coding”, *IEEE trans. Speech and audio processing,* vol.2, no. 4, 1994.

[145] Xavier Domont, Martin Heckmann, Heiko Wersing, Frank Joublin, Stefan Menzel, Bernhard Sendhoff and Christian Goerick, “Word Recognition with a Hierarchical Neural Network”, *Proc. Int. Conf. on Non-Linear Speech Processing,* NOLISP, 2007.

[146] Yakowitz.S, “Nearest neighbour methods for time series analysis”. *J. Times series anal*. Vol. 8, no. 2, pp 235-247, 1987.

[147] Yee Y.S.P. Haykin, “A dynamic regularized gaussian radial basis function network for nonlinear, nonstationary time series prediction". *ICASSP*,1995.

[148] Yegnanarayana.B “Artificial Neural Network”, Printice Hall of India, 1999.

[149] Yoshua Bengio and Renato De Mori, “Use of Neural Network for the Recognition of Place of Articulation”, *IEEE Transactions on Neural Networks,* vol. 8, no. 2, pp. 103-106, 1988.

[150] Young.S, “Statistical Modelling in Continous Speech Recognition ”, *Proc. 17th Int. Conference on Uncertainty in Artificial Intelligence,* Seattle, WA, 2001.

[151] Zhang.B and Srihari.S.N, “Fast k-Nearest Neighbor classification using cluster-based trees”, *IEEE Trans. on PAMI,* vol.26, no.4, pp.525- 528, 2004.

## LIST OF PUBLICATIONS OF THE AUTHOR

[1] **Prajith.P**, Sasindran.E, Sunil Kumar.R.K and Narayanan.N.K, “Malayalam Phoneme Duration analysis for Computer Speech Recognition Applications”, *Proc. National Seminar on Information Revolution and Indian Languages,* Hyderabad, 2000.

[2] Narayanan.N.K, **Prajith.P** and Sasindran.E, “Applications of Phase Space Point Distribution in Speech Recognition”, *Proc. International Conference on Communications, Control and Signal Processing (CCSP 2000),* IISc. Bangalore, 2000.

[3] Sunil Kumar.R.K, **Prajith.P** and Narayanan.N.K, “Development of an Antialiasing Pre sampling filter for Multimedia card based Data acquisition Systems”, *Proc. National Seminar on Information Revolution and Indian Languages,* Hyderabad, 2000.

[4] **Prajith.P** and Narayanan.N.K, “Phase Space Map and Phase Space Point Distribution in Speech Recognition”, *Jour. Acous. Soc. of India*, vol. 31, 2002.

[5] **Prajith.P**, Sreekanth.N.S and Narayanan.N.K, “Phase Space Parameters for Neural Network Based Vowel Recognition”, *Proc. of 11th International Conference on Neural Information Processing (ICONIP 04),* Lecture Notes in Computer Science, Springer, 2004.

[6] **Prajith.P**, Sreekanth.N.S and Narayanan.N.K, “Vowel Recognition using Multi layer feed forward Neural Network with Phase Space Point Distribution as Input parameter”, *Proc. National seminar on Artificial Intelligence and Neural Networks,* Alwaye, Kerala, 2004.

[7] **Prajith.P** and Narayanan.N.K, “Nonlinear Phase Space Features for Robust Pitch Determination”, *Proc. of the National Symposium on Acoustics, (NSA 2006)*, NPL, NewDelhi, 2006

[8] **Prajith.P** and Narayanan.N.K, “Optimum Embedding Parameters for Phase Space Reconstruction”, Communicated.

[9] **Prajith.P** and Narayanan.N.K, “Nonlinear and linear Hybrid Parameters based Speech Recognition”, Communicated.