Analysis of Voice Biometric Attacks: Detection of Synthetic *vs.* Natural Speech

by

ADARSA.S
201211027

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
INFORMATION AND COMMUNICATION TECHNOLOGY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

June, 2014

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

ii) due acknowledgment has been made in the text to all the reference material used.

Adarsa.S

Certificate

This is to certify that the thesis work entitled, "Analysis of Voice Biometric Attacks: Detection of Synthetic *vs.* Natural Speech" has been carried out by Adarsa S. for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under my supervision.

Dr. Hemant A. Patil
Thesis Supervisor

Acknowledgments

I would like to express my deepest gratitude to Dhirubhai Ambani Institute of Information and Communication Technology for providing necessary facilities and environment to carry out this work.

I avail this opportunity to express my sincere gratitude to Dr. Hemant A. Patil for his guidance, advice and encouragement at every step of this endeavour.

I am thankful to Dr. Laxminarayan Pillutla, subject expert, for his suggestions and willingness to help.

I extend profound thanks to members of speech lab, especially Ms. Tanvina Patel, Mr. Nirmesh Shah and Mr. Kuldeep Dhoot for their support extended at different points and sharing their expertise on different text-to-speech systems without which understanding those technology in this short time would not have been possible.

I am greatly indebted to my parents and friends for their wholehearted support which made it possible for me to successfully complete this project in time.

Above all I am always thankful to Almighty for giving me good health and mental strength for completing the project.

Contents

Abstract

The improvement in text-to-speech (TTS) synthesis also poses the problem of biometric attack on speaker verification system. In this context, it is required to analyse the performance of these system using false acceptance rate to impostor using artificial speech and incorporate features in the system to make it *robust* to these attacks. The aim of the study here, is to understand different aspects and hence extract appropriate features for distinction of natural and synthetic speech. The study focuses on understanding those aspects which gives *naturalness* to human speech that the present day TTS systems fail to capture. Three different aspects, *viz.*,Fourier transform phase, nonlinearity and speech prosody are analysed. The results obtained provides an evaluation of the naturalness of the synthetic speech used and gives direction to improve robustness against biometric attacks in speaker verification systems.

List of Principal Symbols and Acronyms

ASV	Automatic Speaker Verification
CFCC	Cochlear filter Cepstral Coefficients
DCT	Discrete Cosine Transform
F0	fundamental frequency
FT	Fourier transform
GMM-UBM	Gaussian mixture model universal background model
HMM	Hidden Markov Model
HTS	Hidden Markov Model-based text-to speech
LE	Lyaponav
LLE	Largest LE
MFCC	Mel-frequency cepstral coefficient
MODGDF	Modified group delay function
RPS	Relative phase shift
SV	speaker verification
TTS	Text-to-Speech
USS	Unit-Selection Synthesis

List of Tables

List of Figures

CHAPTER 1
Introduction

Sound waves and their perception has fascinated humans from time immemorial. Excavations in the oldest archeological sites of the world have resulted in discovery of musical instruments such as *The Divje Babe flute* that dates back to around 40000 BC [1]. The scientific understanding of sound waves as we know today, was first recorded in the studies of Pythagoras and later Aristotle. Understanding of speech as signals and their production mechanism happened much later. Speech synthesis is the artificial production of human speech. Literature from the early 2^{nd} millennium refer to legends of machines to produce human speech. With the advent of computers and the tremendous growth in computational ability, today speech synthesis is one of the demanding research area with giants like Google, Microsoft and militaries of the nations spending millions on speech-based research every year.

1.1 Text-to-Speech Synthesis

A Text-to-Speech (TTS) system is a speech synthesiser that convert normal text to speech [2]. One approach to generating synthetic speech is by concatenating speech segments from a large database. The units are chosen to minimise the acoustic distortion due to difference between output of the TTS system and the target speech [3], [4]. In [5], units in the database is considered as a state transition network, transition cost being the distance between a database unit and a target and an estimate of the quality of concatenation of two speech sound units. Viterbi search was proposed to select the best units for concatenation. Units chosen be *phoneme, diphone, syllable*, etc. Hence, the name unit-selection-synthesis (USS). Over the years, TTS systems using many approaches have evolved. For example, formant synthesis, statistical parametric synthesis (SPS) are some of the approaches for speech synthesis.

1.2 Automatic Speaker Verification (ASV)

Speaker recognition is the identification or verification of a person from his or her voice with the help of machines. Speaker recognition encompasses speaker verification and identification. Speaker verification is the technique to judge whether input speech is the same as the claimed speaker's speech. Use of machines to verify a person's claimant identity from his or her voice is called *Automatic Speaker Verification* (ASV) [6]. Figure 1.1 shows block diagram for a typical ASV system. ASR at a fundamental level is a pattern recognition problem, consisting of two blocks, feature extraction and pattern classification. Ideally the feature extracted should have characteristics that occur naturally or can be extracted without signal distortion, be easy to measure, stable over time, environment and be robust to attack etc. The pattern classifier matches the input features with the feature in the database to make a decision on a claimant speaker. The output of the classifier is the percentage identity in comparison with the claim and a decision is made based on a threshold criteria.

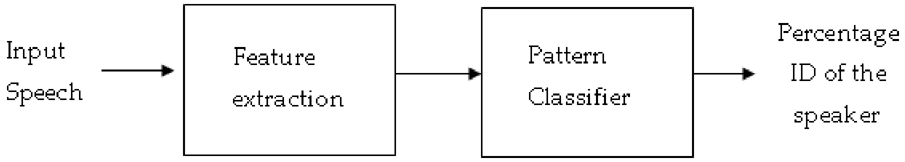


Figure 1.1: Block diagram of a Typical ASV system after [7].

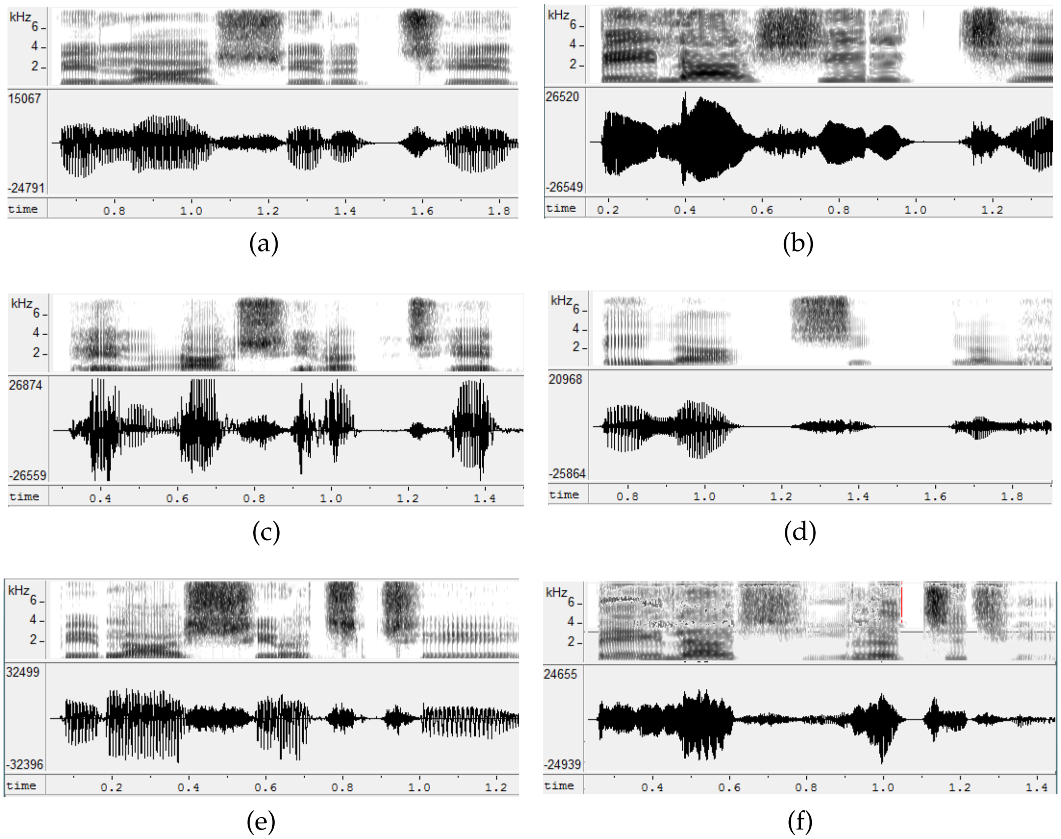
1.3 Motivation

With the improvement in understanding speech signals and widening of their application, *speech recognition* and *speaker verification* has also evolved into vast research areas. For use of ASV systems in practical application, threat to ASV systems in the form of biometric attacks are also to be examined. In addition to impostor techniques such as mimicking a speaker, playback of voice recording and voice modification, with improvement in TTS systems, the analysis of performance of ASV systems in their context and making the systems robust to deliberate attacks using synthetic speech is also an important factor to be considered in designing ASV systems. Chapter 2 discusses in detail, different work carried out in recent years in the context of impostor attack.

1.4 Problem Overview

The problem of distinction of natural and synthetic speech is essentially, finding the component of natural speech that cannot be mimicked in synthetic speech. From a broader perspective, rather than confining to distinction from the synthetic systems, it is required to view the problem as understanding the *naturalness* in speech.

The study presented here is restricted to artificial speech and do not deal with attacks such as mimicking, voice modification, identical twins, etc. In this context, the existing state-of-the-art automatic speaker recognition systems fails to capture a wealth of longer-range and linguistic information that also resides in the signal and the finer nonlinear components of natural speech that are seldom used in modelling. To begin with, USS-based speech is closest to natural speech. The difference lies only at the point of concatenation. Figure 1.2 shows natural, HTS-based and USS based synthetic speech for the same Gujarati utterance.



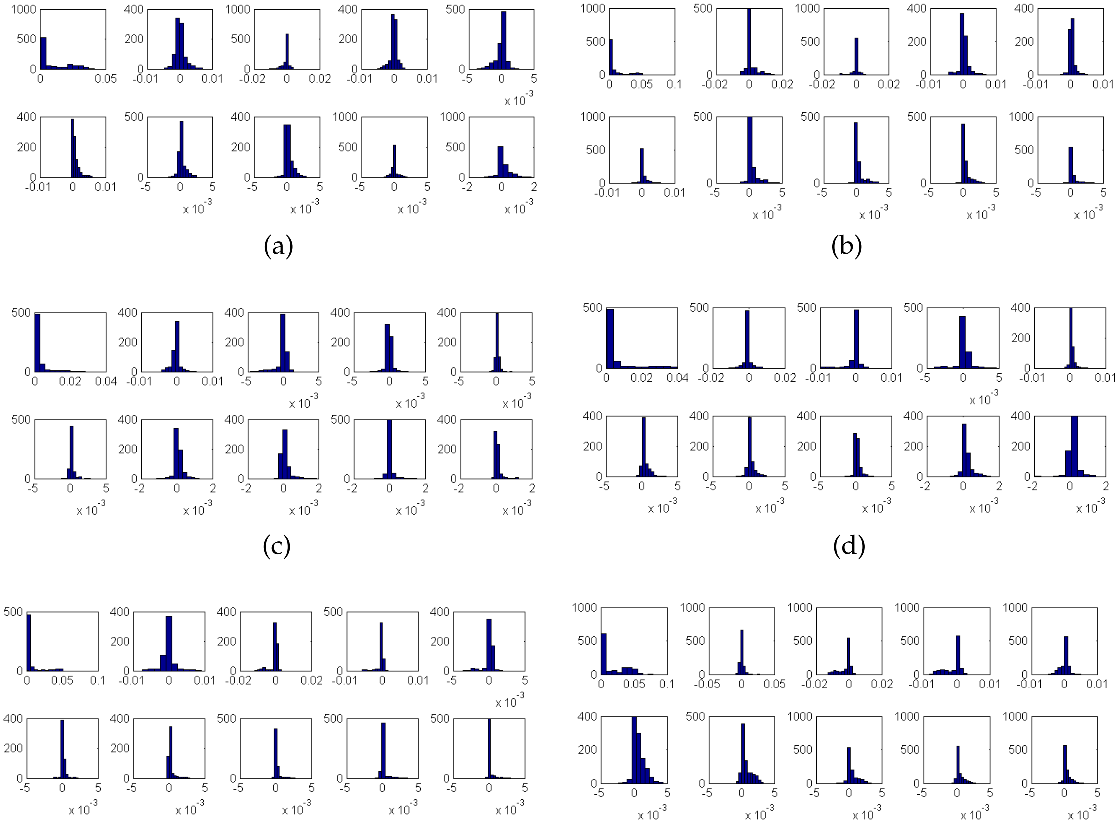
(a) (b)

(c) (d)

(e) (f)

Figure 1.2: Natural, HTS and USS speech and their respective spectrogram for same Gujarati utterance. (a), (b) corresponds to male and female natural speech, (c), (d) to male and female HTS-based speech and (e), (f) to male and female USS-based speech, respectively.

A basic difference that can be observed is the prolonged pauses and mixing between different phonemes in natural speech. This prosodic aspect is expected to be captured in *Fujisaki parameter extraction* discussed in chapter 5. Apart from this, the transitions at the points of concatenation are pretty *smooth* that it is not distinguishable with human eyes in most of the cases. To start with, basic speech processing tools such as Mel Frequency Cepstral Coefficients (MFCC), cochlear filter cepstral coefficients (CFCC), etc. were used to try to capture *spectral* features. The histogram of the MFCC, for natural and synthetic speech for both male and female speaker is shown in Figure 1.3. It can be observed that while MFCC values of natural speech are concentrated near zero, it is not so for USS and HTS speech. Furthermore, MFCC coefficients for female speaker is more spread than corresponding MFCC's for male voice . However there is not much difference between MFCC's of natural and synthetic speeches for the same utterance. This



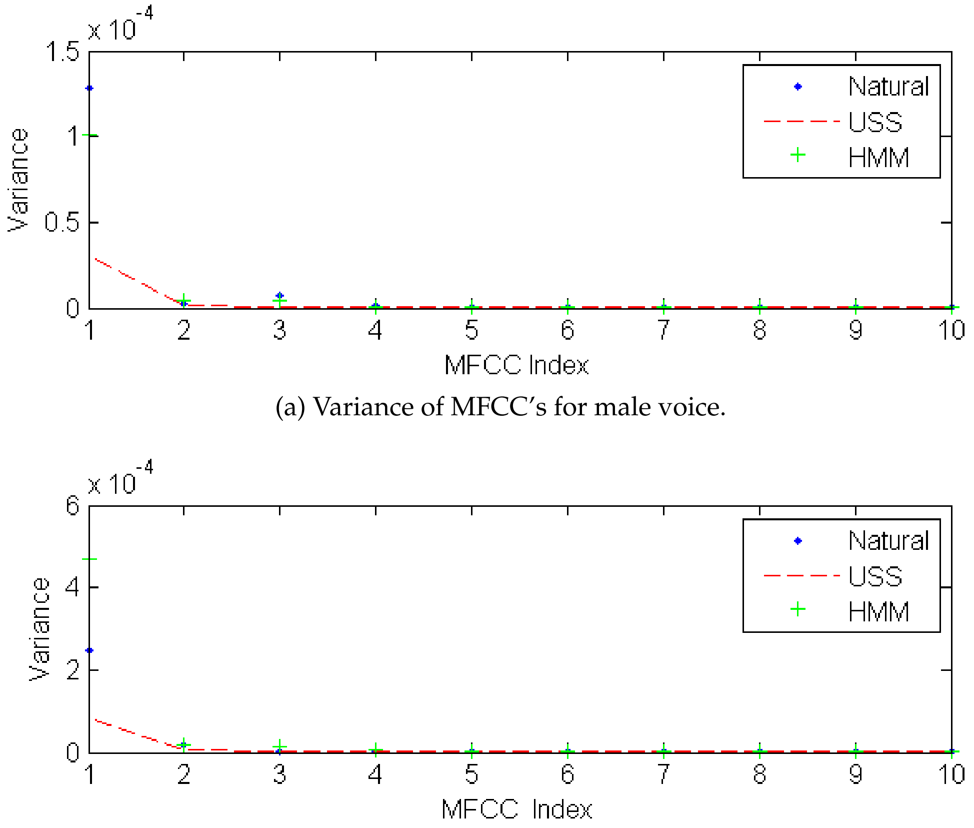
(a) (b)

(c) (d)

(e) (f)

Figure 1.3: Histogram of 10 MFCC values corresponding to (a) male and (b) female natural speech, (c), (d) male and female USS-based speech and (e), (f) male and female HTS-based speech, respectively.

can be further seen in the plot of variances of the coefficients as shown in Figure 1.4. It is seen that the plots for all the three utterances coincide for both male and female voice. This is justified by the fact that the basic system used in *Festival* software for USS speech (ref, Appendix:A) uses MFCC with several other spectral features to improve its performance. Similarly cepstral coefficients are used in HTS systems. This prompts one to look at features that are generally not used in TTS synthesisers.



(a) Variance of MFCC's for male voice.

(b) Variance of MFCC's for female voice.

Figure 1.4: Comparison of variance of MFCC's

The synthetic speech used in the research presented here is from two synthesisers, *viz.*, USS speech from system trained using Gujarati language, both male and female speakers and Hidden Markov Model (HMM)-based Text-To Speech(HTS) also trained for same speakers in Gujarati. Appendix A and appendix B gives a detailed description of the training and synthesis. Unless specified, the natural data used in the study is studio recorded speech sampled at 16 kHz. The same speaker data for different utterance is used to train the TTS systems.

1.5 Organisation of the Thesis

The organisation of the thesis through chapters from $1 - 4$ is presented as follows. Chapter 2 gives a review of the literature survey involved in arriving at the problem statement. Furthermore, existing methodologies that have been researched in this area are discussed. In Chapter 3, group delay as a feature for analysis of natural and synthetic speech is discussed. Since it is a century old observation that human ear is *deaf* to phase changes, TTS systems generally ignore the phase information. The study involving signal processing-based method of *relative-phase-shift*(RPS) has been treated as a reference point for this study [8].

Chapter 4 discusses extracting nonlinearity from speech data. The traditional approach to speech signal modeling has been the linear, i.e., source-filter model. This involves approximating the nonlinear aspects of speech production using assumptions of linear acoustics of sound wave in the vocal tract. Using *Lyapunov Exponent* (LE) as the feature to capture amount of *chaos* inherent in the speech samples, study has been carried out to understand the difference between natural speech and the synthetic speech. The chapter concludes with using LE to classify phoneme instances of synthetic and natural speech data.

Looking into how to distinguish natural and synthetic speech, the most interesting initial observation is that our ears do it with ease. This is due to the *prosodic* aspects such as positional, contextual and phonological information always present in natural speech. Chapter 5 discusses Fujisaki model for prosodic parameter extraction. The *phrase* and *accent* component of natural and synthetic speech has been studied. The speaker-dependent parameters in both the cases is found to have considerable difference. This in turn enables using these parameters for detection of synthetic speech. Finally, the thesis is concluded in Chapter 6. A summary of the work done and the results obtained is discussed here. Furthermore, the chapter discusses application of the insights that the study has provided and the scope of future work in the area.

1.6 Chapter Summary

This chapter has provided the problem formulation and motivation of the thesis. The chapter gives an introduction to the various aspects of impostor detection to be discussed further. In addition a road map of the work done and layout of the thesis is provided here. The next chapter provides a brief literature search on the problems addressed in the thesis.

CHAPTER 2

Literature Survey

2.1 Introduction

This chapter intends to give an overview of different terminologies and technologies discussed in this thesis. A background study of the prior work carried out in different areas related to the problem, that has formed a base for the work carried out in the research is presented here. Though the particular problem of impostor attack is relatively untapped, different researches that has motivated this study is discussed. Figure 2.1 shows a tree of thesis overview and different aspects of the problem.

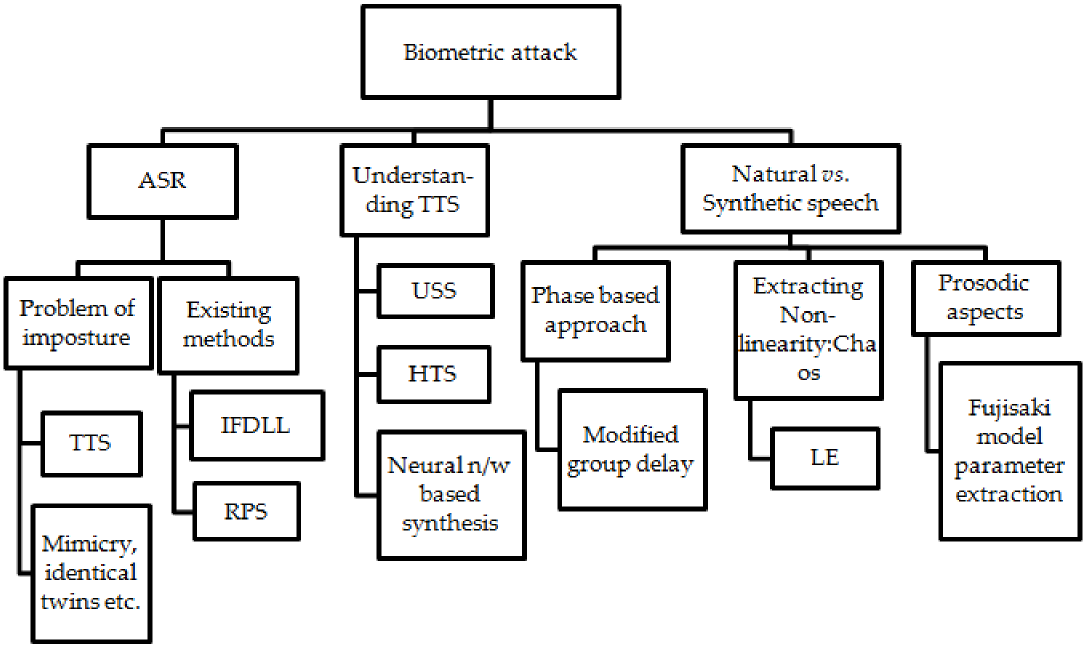

Biometric attack
ASR
Understan-ding TTS
Natural vs. Synthetic speech
Problem of imposture
Existing methods
TTS
Mimicry, identical twins etc.
IFDLL
RPS
USS
HTS
Neural n/w based synthesis
Phase based approach
Modified group delay
Extracting Non-linearity:Chaos
LE
Prosodic aspects
Fujisaki model parameter extraction

Figure 2.1: Thesis Overview

2.2 Automatic Speaker Verification

The objective in speaker verification (SV) is to accept or reject a claim of identity based on a voice sample. The problem of identification is to decide if a speaker is a specific person or is among a group of persons and hence the problem is of pattern recognition. Gaussian mixture model universal background model (GMM-UBM) is one of the most used approach for text-independent speaker recognition applications [9]. In GMM, *expectation maximisation* algorithm is used to estimate model parameters from feature vector of the speaker. A *log-likelihood* detection is used to verify the claimant speakers identity [10]. As ASV systems gain widespread use, it is crucial to understand the errors made by these systems, *viz.* the false acceptance of an invalid user and the false rejection of a valid user [6]. In the study concerning synthetic speech, the aim is to reduce false acceptance of invalid user.

2.3 Various Biometric Attacks

The simplest impostor is playback of a voice recording for a targeted speaker and the logical solution is a text-prompted approach. In addition, the vulnerability of SV to voice mimicking by humans using twins has also been examined [7]. Technical impostor attacks include, synthetic speech production by concatenation of recorded speech. This method would require large recorded speech database of the target speaker. Another plausible way for an impostor to attack is to try and re-synthesise his/her own voice into that of the claimant speaker through some kind of transform [11]. Today's state-of-art systems include, HMM, USS and artificial neural networks.

2.4 Problem of Impostor

There is considerable amount of work in industry, national laboratories and universities in the area of speaker-verification. Few benchmark work in the specific area of impostor attacks is presented in Table 2.1. "Source" refers to a citation in the reference and "year" refers to the year of publication. "Attack" refers to the type of impostor attack used and "method" to the SV system used in the study. "Error" column shows the percentage of false acceptance for male speaker data for each of the system.

The earliest and most used method in SV systems is the likelihood detector. [12] uses it to analyse the performance of a linear predicted signal. Lindberg and

Table 2.1: Selected chronology of published work in analysis of attack on SV systems

Source	Year	Attack	Method	Error
F.K.Soong, et al. [12]	1985	Linear prediction analysis	Likelihood ratio detector	5 %
Lindberg and Blomberg [13]	1999	concatenation and resynthesis	log-Likelihood	42.6 %, 6.8 %
Zhizheng, et al. [14]	2012	converted speech using GMM and USS	MODGDF-phase	9.13 % and 4.6 % resp.
De Leon, et al. [8]	2012	HMM-based speech	GMM with RPS	10.1 %
Zhizheng, et al. [15]	2013	converted speech using JD-GMM and USS	GMM-UBM	42.5 % and 39.2 % resp.

Blomberg in their 1999 paper gives a detailed analysis of the then popular concatenated target speech, re-synthesis of the target speech and diphone synthesis of the target. Later, the performance of concatenated speech against the present state-of-the-art GMM-UBM system was done in 2012 [14]. In the following year, studies have also been reported on the evalutation of GMM-UBM against HMM-based speech [8]. This paper for the first time proposes dedicated feature of RPS for SV system.

The general trend shows that, with the improvement in quality of artificial speech production, the performance of state-of-art SV systems with additional features is also low. This is because, the design of TTS systems concentrate on reducing false rejection rate. The GMM-UBM used in [8] has acceptance rate of true claim as 100 %.

2.5 Phase-based Approaches

Since the human auditory system is insensitive to the phase of speech signal's [16], TTS is normally based on a *minimum-phase* vocal tract model. [17] discusses the processing of phase of *Fourier transform*(FT) to derive smooth log-magnitude spectrum corresponding to the vocal tract system and [18] discusses an application for *modified group delay* in phoneme recognition. Another phase based feature, *relative phase shift*(RPS) for impostor detection was published in [8].

2.6 Nonlinear Analysis

The enhancement in understanding the speech production mechanism has motivated many studies involving analyzing nonlinear behaviour in speech production systems. Studies reported in [19] and [20] explore models suitable for extracting information about modulation, *fractals* and *chaotic structure* of speech signals and use it in applications such as recognition and synthesis. In addition, quantitative measures of chaos such as *Correlation Dimension* (CD) and LE have been used in [21] for speech decomposition. With improvement in TTS systems, nonlinearity-based feature can be used to improve ASV systems.

2.7 Chapter Summary

This chapter is a background study on the work done in area of impostor attack and the different methods used in the study for analysis of natural and synthetic speech. Different aspects of SV that has lead to the thesis topic has been discussed. It is noted that though the area of SV and TTS systems is subjected to a lot of extensive study, the specific problem of natural and synthetic speech is relatively untapped. In addition with fast improvement in speech production techniques and increased real life application of speech based systems, the problem requires to be addressed.

CHAPTER 3

Phase-based Approach

3.1 Introduction

Studies reported in [8] shows the relevance of Fourier transform phase in the context of speaker verification. This motivates one to use phase-based features for distinction of synthetic and natural speech. However, using the phase spectrum directly has the disadvantage that it requires unwrapping. Another measure of phase that is relevant in this context is *group delay*. This chapter explores group delay based features for detection of synthetic speech.

3.2 Modified Group Delay

Group delay is the *negative* derivative of the phase function and can be directly calculated from the signal. If $x(n)$ is a speech frame, whose FT $X(\omega)$ is given by

$$X(\omega) = |X(\omega)|e^{j\theta(\omega)}, \tag{3.1}$$

then group delay is given by,

$$\tau(\omega) = -\frac{d\theta(\omega)}{d(\omega)}. \tag{3.2}$$

A multiplication in magnitude spectrum becomes multiplication in group delay domain. This additivity property is made use of in formant extraction. Closely spaced formants are better resolved in group delay-domain as shown in figure 3.1 below.

When dealing with digital signal, the computation is accurately represented [22] by,

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2}, \tag{3.3}$$

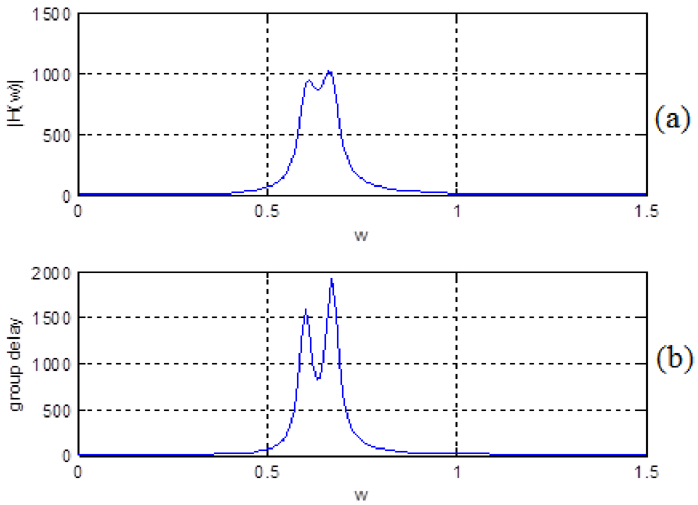


Figure 3.1: The top figure shows the plot of magnitude spectrum and the bottom one shows formants in group delay plot for a cascaded system with poles at $\omega = 0.6$ *rad* and 0.8 *rad* .

where $y(n) = nx(n)$ and its FT $Y(\omega) = Y_R(\omega) + jY_I(\omega)$ and $X(\omega) = X_R(\omega) + jX_I(\omega)$. $|X(\omega)|$ represents the magnitude of $X(\omega)$. Due to excitation of source and short-term processing, zeros occur in the vocal tract modelling and at these points, the denominator of the expression given in eq.(3.3) becomes zero. Computation of group delay near these zeros hence gives large values and masks the formant structure. This issue is addressed by *modified group delay function* (MODGDF) proposed in [22]. MODGDF for an all-pole system is defined as,

$$\tau_m(\omega) = sign|\tau^{'}(w)|^{\alpha}, \tag{3.4}$$

where

$$\tau^{'}(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\gamma}}. \tag{3.5}$$

Here, $|S(\omega)|$ is the *cepstrally smoothed* version of $|X(\omega)|$. α and γ, control the dynamic range of MODGDF, $0 \leq \alpha, \gamma \leq 1$.

To compute MODGDF-based feature from the speech signal, first it is required to represent the speech signal as an all-pole system. This is done using *linear prediction* analysis [23]. Speech sound with impulsive or periodic sources, are loosely categorised as "deterministic", while speech sounds with noise sources as "stochastic" sounds. Estimating all-pole model parameters for these systems provides a desirable signal in both time and frequency-domain. A signal S_n is considered to be linear combination of the past outputs and the inputs from time

1 to n.

$$S_n = \sum_{k=1}^{p} a_k \widehat{S}_{n-k} + G \sum_{l=0}^{q} b_l u_{n-l}, b_0 = 1, \tag{3.6}$$

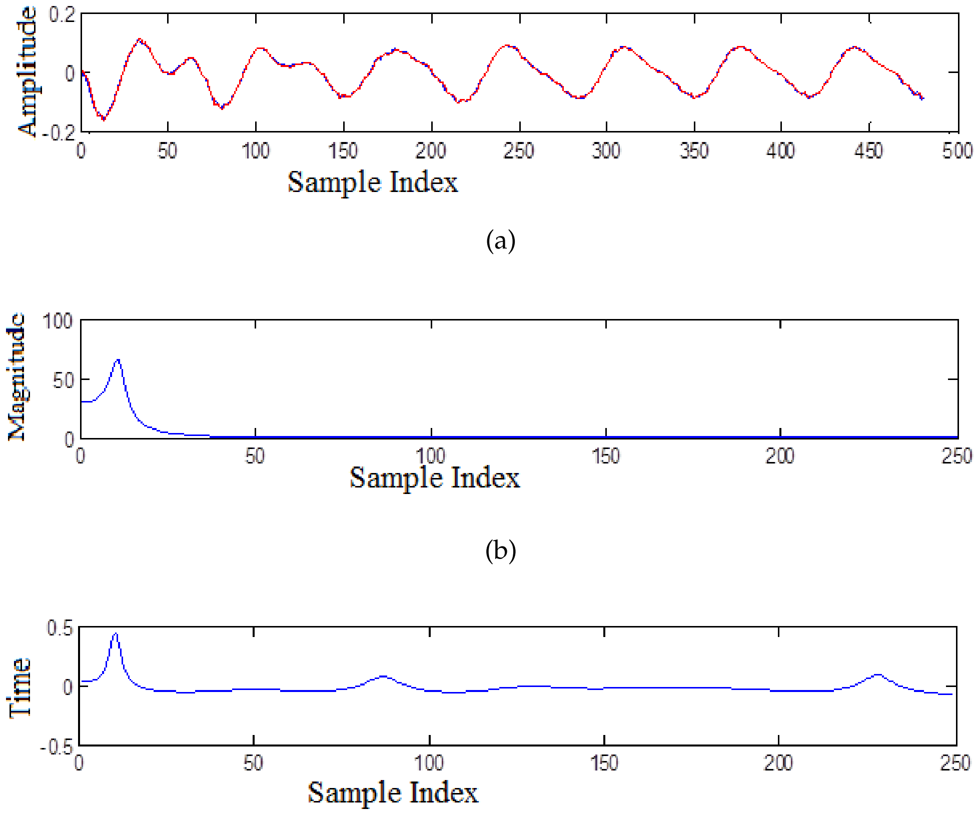
where u_n is the assumed input signal, a_k, $1 \leq k \leq p$, b_l, $1 \leq l \leq q$ and G are parameters of the hypothesised system. To minimise the error of approximation, the error signal, i.e., the difference between original signal and predicted signal $S_n - \widehat{S}_n$ is differentiated with respect to the coefficients a_k's. This gives $R_n\alpha = r_n$, where R refers to the *autocorrelation* matrix which is a *Toeplitz* matrix and α is the matrix of all coefficients and r is a $p \times 1$ matrix of autocorrelation values for lag 1 to p. The equation is solved using *Levinson's recursion algorithm*.

3.3 Algorithm and Analysis

A given speech signal is first fragmented into frame sizes corresponding to more than a pitch period, typically 30 *ms* with more than 50 % overlap between the frames. Phase-based analysis has used natural speech and speech from USS, HTS synthesisers. These would be henceforth referred to as USS-based and HMM-based speech. After finding LP coefficients and hence the system function, corresponding to the speech frame, the modified group delay is calculated. Figure 3.2 shows the MODGDF for the all pole model of a speech segment. To quantify MODGDF, its discrete cosine transform (DCT) is taken. DCT represents the energy in a signal. Figure 3.3 shows the entire process in extraction of feature.

Since the maximum energy in the signal is concentrated in its larger DCT coefficients, the coefficients were sorted and the probability distribution of the prominent coefficients were observed for samples of natural, HMM-based and USS-based speech signals. Figure 3.4 shows the probability distribution of largest 12 DCT coefficients over all the frames for the three cases.

For all the coefficients, it can be noted that the maximum DCT value is for the natural speech closely followed by USS and least for HMM. This implies that natural speech has maximum amount of inherent phase mismatches. HTS being based on parametric synthesis, fails to capture those. Since this pattern is recurring in value of PDF, the DCT of modified group delay with the maximum value could be used as a parameter for distinction of synthetic and natural speech.



(a)

(b)

(c)

Figure 3.2: LP analysis using 16 prediction coefficients. (a) Shows the original signal in blue and the predicted signal in red, (b) the frequency spectrum of the predicted signal and (c) shows group delay plot derived from the predicted signal.

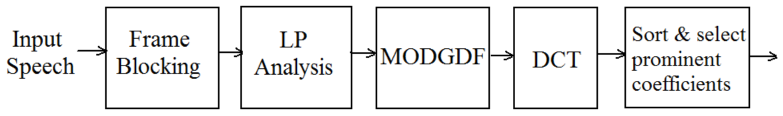


Figure 3.3: Algorithm for parameter extraction using MODGDF.

3.4 MODGDF for USS-based Synthetic Speech

Since the USS speech is composed of concatenated units of natural speech, their difference lies only at the point of concatenation. To further understand the effect of MODGDF, USS speech from different hours of training data was analysed. As the number of hours of training increases, the chance of adjacent units being selected from recordings farther in time increases. A delay in time-domain corresponds to addition of phase of the signal. It is observed that even though there is

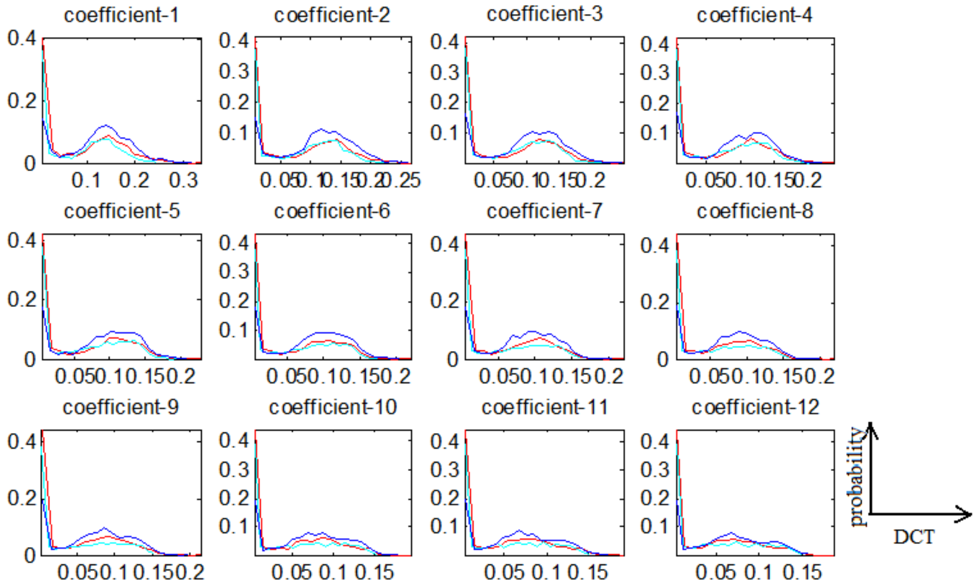


Figure 3.4: PDF's over the frames of DCT coefficients : x-axis-value of DCT of MODGDF, y-axis-Probability over all the frames.

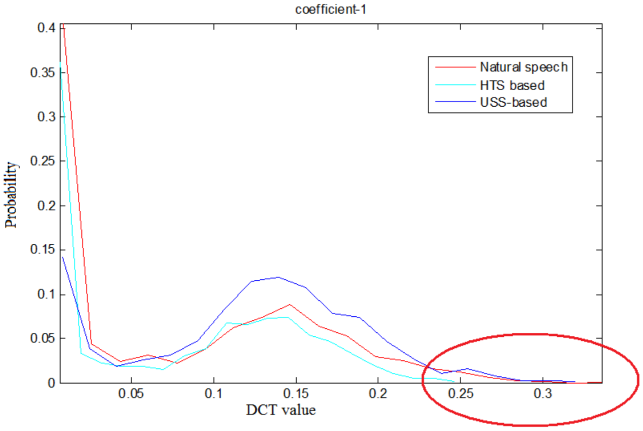


Figure 3.5: PDF over all the frames for the first DCT coefficient for male voice. The circled portion shows the largest DCT coefficient values.

no difference in hearing or in the spectrogram of say,1 and 8 hours training-based synthesis, the maximum value of modified group delay is considerably different. Figure 3.6 shows the value of maximum DCT value for USS speech for a particular text for different hours of training from 1-8 hrs.

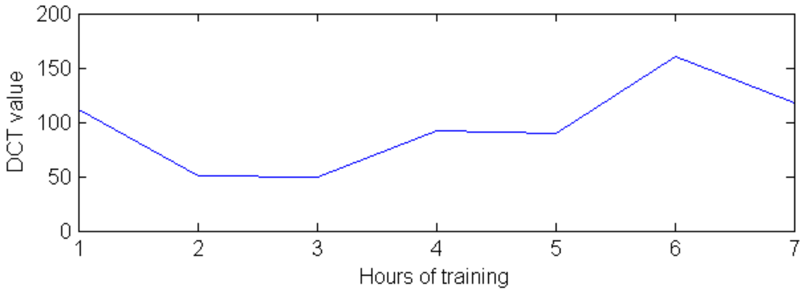


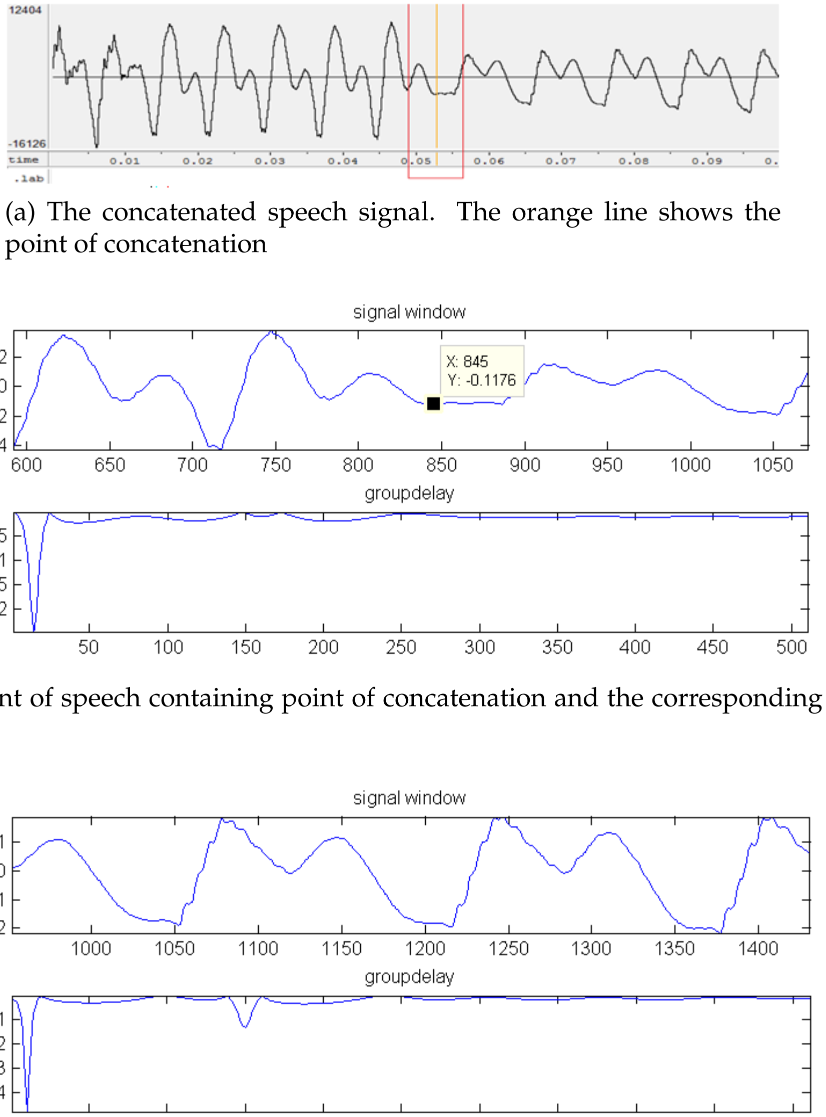
Figure 3.6: Maximum value of DCT coefficient of MODGDF for USS-based speech for different hours of training for male voice

It is noted here that hours of training has no direct relation on the largest DCT coefficient. This is in accordance with the observation that DCT values for natural speech is greater than USS-based speech implying the concatenation introduces no greater phase mismatches. Figure 3.7 shows MODGDF of a speech segment containing concatenation point. The maximum value of MODGDF for a speech segment containing the point of concatenation shown in figure 3.7b is -0.25 while for adjacent window not containing concatenation point, it is -0.4. Since concatenation points does not necessarily correspond to maximum group delay, MODGDF does not provide a robust enough feature for distinction of USS-based and natural speech. Further exploration is required in this regard.

3.5 MODGDF for HTS-based Synthetic Speech

From the *pdf* of DCT coefficients, it is noted that, the value of DCT is considerably smaller for HMM-based speech as compared to natural and USS-based speech. To explore this further, 100 samples each of HMM-based and natural speech for male voice was considered. Figure 3.8 shows the histogram of the DCT values for the samples. The mean DCT value for natural speech is *0.22* and that of HMM-based speeches is *0.19*. The same for USS was observed to be close to natural speech at *0.21*.

The DCT values for natural speech is sparsely distributed. Furthermore, for each individual utterance, the value of natural speech is higher than that of HMM synthetic speech. Hence, it can be concluded that the MODGDF is always greater for natural speech.



(a) The concatenated speech signal. The orange line shows the point of concatenation

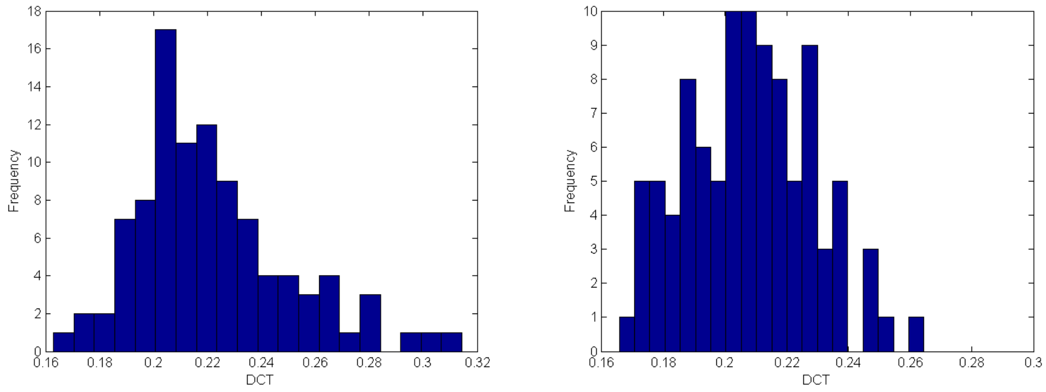
(b) Segment of speech containing point of concatenation and the corresponding MOD-GDF

(c) Segment of speech that is from a single unit and the corresponding MODGDF

Figure 3.7: Study of point of concatenation

3.6 Chapter Summary

The results obtained shows that the amount of phase mismatch inherent natural speech is higher than those produced artificially. Since USS speech contains majorally natural speech and since, the phase mismatches happening at the points of concatenation is relatively less, group delay is not a good measure of distinction. HMM-based synthetic speech is produced depending on certain fixed parameters that are learned. Since phase-related measures are not explicitly included as a training parameter, the produced speech fails to capture the phase deviations. Hence, given a speech sample for comparison, HMM speeches are distinguish-



(a)

(b)

Figure 3.8: Histogram of largest DCT of MODGDF of (a) natural speech and (b) HMM-based speech

able from natural speech. This is in accordance with the results obtained using RPS though in the aspect of implementation, MODGDF is easier to be extracted.

The results obtained in this section motivates one to look further into exploring natural components of speech that TTS systems fail to incorporate. Moving a step ahead, the next chapter tries to understand those aspects of speech that makes natural systems unique. Even with the exponential rate of advancement of technology, we still do not understand life forms completely. The attempt is to understand *chaos* and nonlinearity deeply embedded in speech signals.

CHAPTER 4

Nonlinear Analysis

4.1 Introduction

Many investigations on speech nonlinearities have been carried out and these studies provide strong evidences to support nonlinear system modelling of speech production. The nonlinear characteristics that these studies point to are analogous to chaotic systems. Chaos is the phenomenon of occurrence of bounded non periodic evolution in completely deterministic nonlinear dynamical systems with high sensitive dependence on *initial* conditions. For a system to be chaotic, it should have sensitive dependence on initial conditions. It has aperiodic orbit. Though it seems to be random, system is actually deterministic. The rate of separation of nearby trajectories is positive [24].

Studying real-life nonlinear deterministic systems involve modeling them using an attractor model, i.e., a set values to which the system evolves independent of the starting point. Hence, measures that are invariant in this approximation are required for quantizing the nonlinear time series. Some such measures are *correlation dimension*(CD) and Lyapunov exponent. CD gives lower boundary for the degrees of freedom a signal possesses and hence, a measure of complexity of the system [25]. LE estimates the mean exponential divergence or convergence of nearby trajectories in phase-space.

4.2 Chaos and Lyapunov Exponent

The similarities of chaotic systems to speech production were discussed in [26] coincidentally following another study reported in [27] which proposed an algorithm for finding largest LE (LLE) (though the algorithm was applied for speech signals later in [28]). The Rosenstein's algorithm is used for finding largest LE in this study. Since understanding chaoticity requires understanding the speech signal at sample-level, the choice of algorithm that works well for smaller data-sets is an appropriate choice. The analysis is done using natural speech and HMM systems trained on Gujarati data from a single speaker with sampling frequency 16 *kHz*. Overlapping speech windows of *400-600* samples were considered.

4.3 Rosenstein's Algorithm for Largest Lyapunov Exponent

The Rosenstein's algorithm involves reconstructing the attractor dynamics from the speech sample by using method of delays [28]. The first step involves expressing the reconstructed trajectory say X, as a matrix where each row X_i is a phase-space vector at discrete-time i. For an N-point time series $(x_1, x_2, ...x_n)$, we have

$$X_i = (x_i, x_{i+\tau}, ..., x_{i+(m-1)\tau}), \tag{4.1}$$

where τ is the *reconstruction delay* and m is the embedding dimension. The reconstruction delay τ was taken as the time at which the autocorrelation function has the first zero. This would make the coordinates linearly uncorrelated. After reconstructing the dynamics, the algorithm locates the nearest neighbour of each point on the trajectory. The nearest neighbour, X_i, is found by searching for the point that minimizes the distance to the particular reference point, X_j. The LLE is then estimated as the mean rate of the nearest-neighbour separation. The LLE λ is defined using

$$d(t) \approx Ce^{\lambda t}, \tag{4.2}$$

where $d(t)$ is the average divergence at time t and C is a constant that normalizes the initial separation [24]. Assuming the j^{th} pair of nearest neighbors diverge approximately at rate given by the LLE,

$$d_j(t) = C_j e^{\lambda_i (i\Delta t)}, \tag{4.3}$$

where Δt is the sampling period of the speech time series, $d_j(t)$ is the distance between the j^{th} pair of nearest neighbors after $i\Delta t$ seconds, and C_j, is the initial separation. Taking logarithm and averaging over all j gives LLE. Using a least-squares fit to the "average" line defined by

$$y(i) = \frac{1}{\Delta t} \langle \ln d_j(i) \rangle \tag{4.4}$$

where $\langle . \rangle$ denotes the average over all values of j gives largest LE. Table 4.1 shows

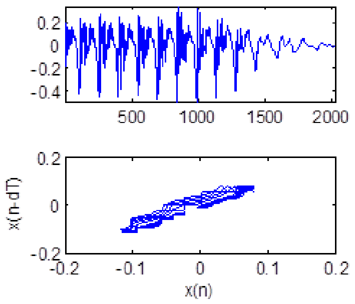
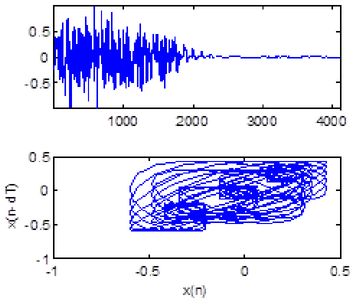
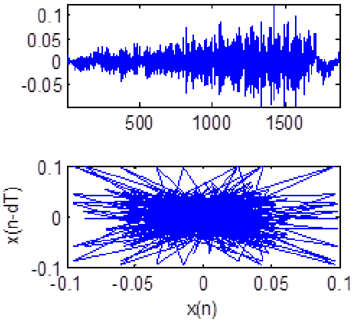
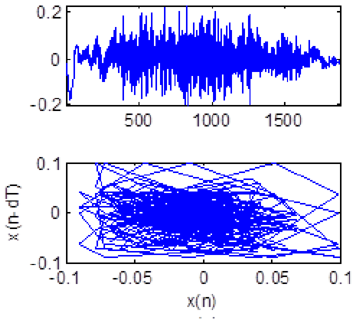
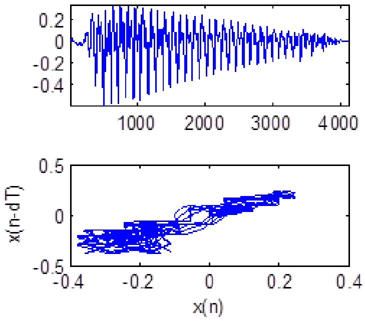
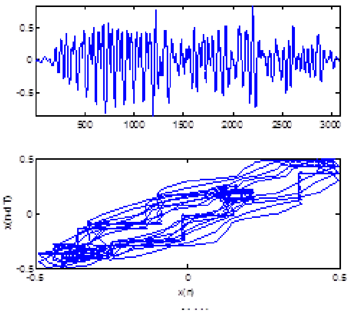
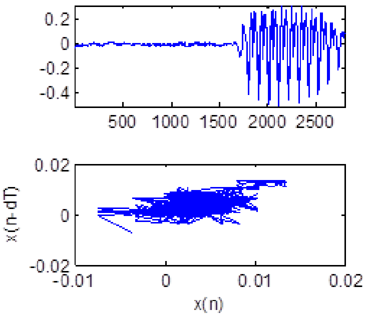
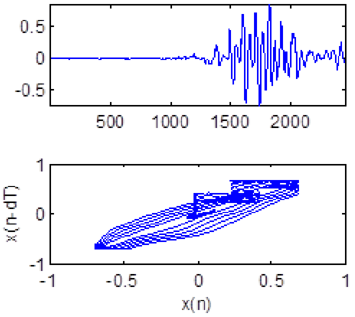
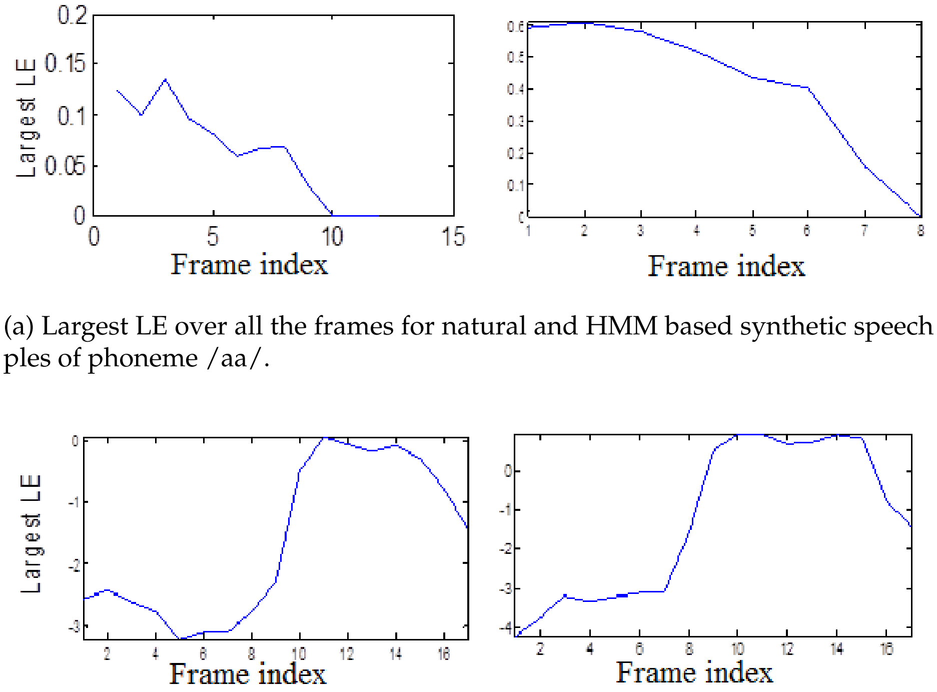
Phoneme	Natural	HTS-based
aa		
s		
p		
f		

Table 4.1: Phase plots of different phonemes.

the phase-plot for phoneme /*aa*/, /*p*/ and /*f*/ in Gujarati, sampled at 16 *kHz* for delay in the range *1-5*. Phase-plot shows all possible values of the system. The system value is plotted against time delayed values of the same system. It can be observed for /*aa*/ and /*p*/ that the orbits are deterministic and aperiod. In addition, for different initial conditions, they do not follow the same orbit, exhibiting all the characteristics of a typical chaotic attractor. More importantly, in case of HTS-based speech, the attractor can be observed for all the phonemes. This implies that HMM speech is more chaotic than natural speech for all phonemes under consideration.

4.4 Experimental Setup and Results

This section intends to discuss experiments conducted and their results. Instances of phoneme from natural speech samples and HTS-based speech for single Gujarati speaker are considered. The initial stage of study involved, understanding the nature of the speech samples. The speech segments were fragmented into overlapping windows and largest LE of each frame was calculated. Plot of /LLE

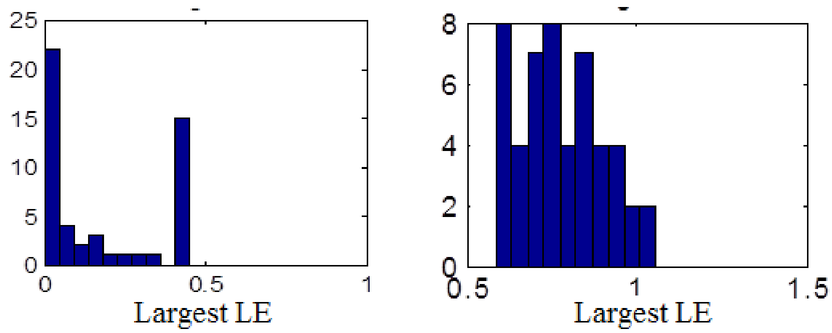


(a) Largest LE over all the frames for natural and HMM based synthetic speech samples of phoneme /aa/.

(b) Largest LE over all the frames for natural and HMM-based synthetic speech samples of phoneme /f/.

Figure 4.1: Largest LE comparison.

over all the frames for vowel /*aa*/ and fricative /*f*/ for overlapping windows of



(a) (b)

Figure 4.2: Histogram showing distribution of largest LE for samples of (a) natural and (b) HTS-based synthetic speech.

size 480 samples is shown in figure 4.1a and figure 4.1b. It can be noted that the value of LLE is more in the case of vowel /*aa*/. Furthermore, for all the phonemes, it is observed that the value of LLE is higher in the case of HTS-based speech. This is explained by the fact that HMM uses statistical parametric speech synthesis. Hence, the modeling of a particular speech for the input text is based on *context- dependent phonemes* from the training phase. Since the speech is *vocoded*, the modeling of the voice source is naive in the sense that periodic and aperiodic excitations are switched unlike a natural source where they are mixed.

For the usefulness of the observation in the context of speaker verification, the experiment was extended to *50* instances of each of the phoneme. Phonemes were extracted from natural speech signal and the HTS-based speech signal from the HTS system mentioned above. The LLE among all its frames was used to represent a particular sample. The distribution of the LE for phoneme /aa/ for natural and synthetic speech is shown in figure 4.2.

It is observed that the distribution of the LLE for natural and synthetic speech is different. The distributions are entirely non-overlapping. In addition, the average value of LLE is higher in the case of HTS-based synthetic speech. This difference between synthetic and natural speech can be used in ASV systems. Results obtained in this study shows means to improve the performance of the system.

Figure 4.3 shows the distribution of the data points corresponding to 50 instances of natural and synthetic phoneme /aa/. In the $2-D$ plot, the separation of the natural and HTS speech data points into different clusters is clearly observed. Fixing a threshold at an optimum value of maximum LE *0.25* and second largest LE *0.2* would give a false acceptance rate of synthetic speech to be *4* %. Though the results could vary based on the extraction of phonemes, use of LE is certainly a huge improvement over the phase-based methods proposed earlier.

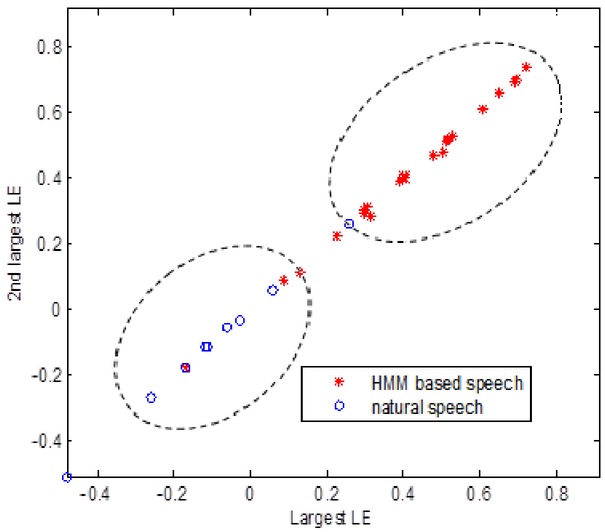


Figure 4.3: Classification of natural and synthetic signal corresponding to of phoneme/aa/. The blue points represent natural speech and red ones correspond to synthetic speech. The dotted lines show the clustering of data points.

4.5 Chapter Summary

With a different synthetic speech production, the error rate would increase. For instance, using speech from USS-based system for impostor attack would require a different approach than directly using the phonemes. Since the phonemes used are directly from the training natural speech, the property of difference in chaoticity made use of here is not valid. Hence, point of concatenation becomes of significance. Nevertheless, the experimental results obtained paves way to look at the problem of speaker verification from a different perspective. With the increasing use of speech-based systems in real-life application, the linear approximations used in speech modeling misses out some factors that might prove important. Furthermore, study on chaoticity brings one closer to understanding the naturalness in the human speech production.

CHAPTER 5
Prosodic Aspects: Fujisaki Model-based

5.1 Introduction

Looking into how to distinguish natural and synthetic speech, the most interesting initial observation is that our ears do it with ease. A convincing hypothesis is that, it is through linguistic context and production constrains such as positional, contextual and phonological information always present in natural speech. These are depended on the physical traits of the speaker as well as the learned traits. In speech processing, these characteristics are broadly called prosody. Common parameters used are pitch, duration and energy dynamic.

5.2 Fujisaki Model

The fundamental frequency (F_0) contour, conveys a lot of information about the linguistic, para-linguistic and non-linguistic information in speech signal. The frequency of vibration of the vocal cords mainly through various intrinsic and extrinsic laryngeal muscles, accounts for the fluctuations in the F_0. The slow changes(global components) in the F_0 contour conveys the linguistic information on the syntactic structure, while information on the word accent/syllable tone is expressed by relatively rapid changes (local components) of the F_0 contour. Fujisaki in his paper elaborately explained the role of cryconoid muscles and how its rotational and translational motion is responsible for *accent* and *phrase* component of speech respectively [29].

He further states that the command-response model represents the contour of the logarithm of fundamental frequency, i.e., $\ln F_0(t)$, as the sum of phrase components, accent/tone components, and a speaker dependent baseline level $\ln F_b$.

Thus the F_0 contour as a function of time can be expressed by the following

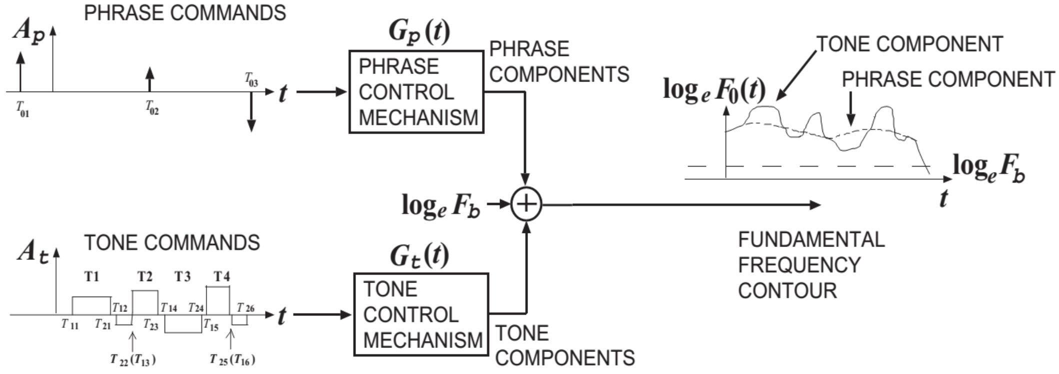


Figure 5.1: A command-response model for the process of F_0 contour generation. Adapted from [29]

equations:

$$\begin{aligned}\ln F_0(t) = \ln F_b(t) + \Sigma_{i=l}^{I} A_{pi} G_p(t - T_{0i}) + \Sigma_{j=l}^{J} [A_{t1j} G_t(t - T_{1i}) - G_t(t - T_{2i}) \\ + A_{t2j} G_t(t - T_{2i}) - G_t(t - T_{3i})],\end{aligned} \tag{5.1}$$

where

$$G_p(t) = \begin{bmatrix} \alpha^2 t \exp(-\alpha t) & t \geq 0, \\ 0 & t < 0 \end{bmatrix}, \tag{5.2}$$

$$G_t(t) = \begin{bmatrix} min[1 - (1 + \beta_1 t) \exp(-\beta_1 t), \gamma_1], & t \geq 0, \\ 0, & t < 0 \end{bmatrix} \tag{5.3}$$

(for positive tone commands),

$$G_t(t) = \begin{bmatrix} min[1 - (1 + \beta_2 t) \exp(-\beta_2 t), \gamma_2], & t \geq 0, \\ 0, & t < 0 \end{bmatrix} \tag{5.4}$$

(for negative tone commands),

$G_p(t)$ represents the impulse response function of the phrase control mechanism and $G_t(t)$ represents the step response function of the tone control mechanism. The symbols in equation indicate,
F_b-baseline value of fundamental frequency
I-number of phrase commands
J-number of syllables
A_{pi}-magnitude of the ith phrase command

A_{t1j}-amplitude of the first command in the jth syllable
A_{t2j}-amplitude of the second command in the jth syllable
T_{0i}-timing of the i^{th} phrase command
T_{1j}-onset of the first command in the j^{th} syllable
T_{2j}-end of the first command (and onset of the second command if the second command exists) in the j^{th} syllable
T_{3j}-end of the second command if the second command exists in the j^{th} syllable
α- natural angular frequency of phrase control mechanism, set empirically at $3/s$
β- natural angular frequency of tone control mechanism, set empirically at $20/s$
γ- relative ceiling level of tone components, set empirically at 0.9

5.3 Algorithm and Experimental Results

Its seen that the pitch contour of a speech signal can be decomposed into the slow varying phrase, fast varying accent component and constant speaker-dependent parameter. Since each of these capture different prosodic aspects of a signal, Fujisaki model parameter extraction could prove to be a promising method to understand prosodic aspects of natural speech. The TTS systems used in the study uses F_0, pitch etc. to capture prosody during synthesis. An analysis of the hence produced artificial speech in comparison with natural speech is done here. Figure 5.2 below shows algorithm used in extraction of phrase and accent components.

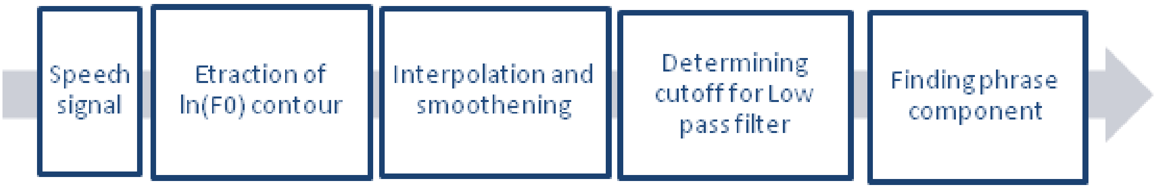
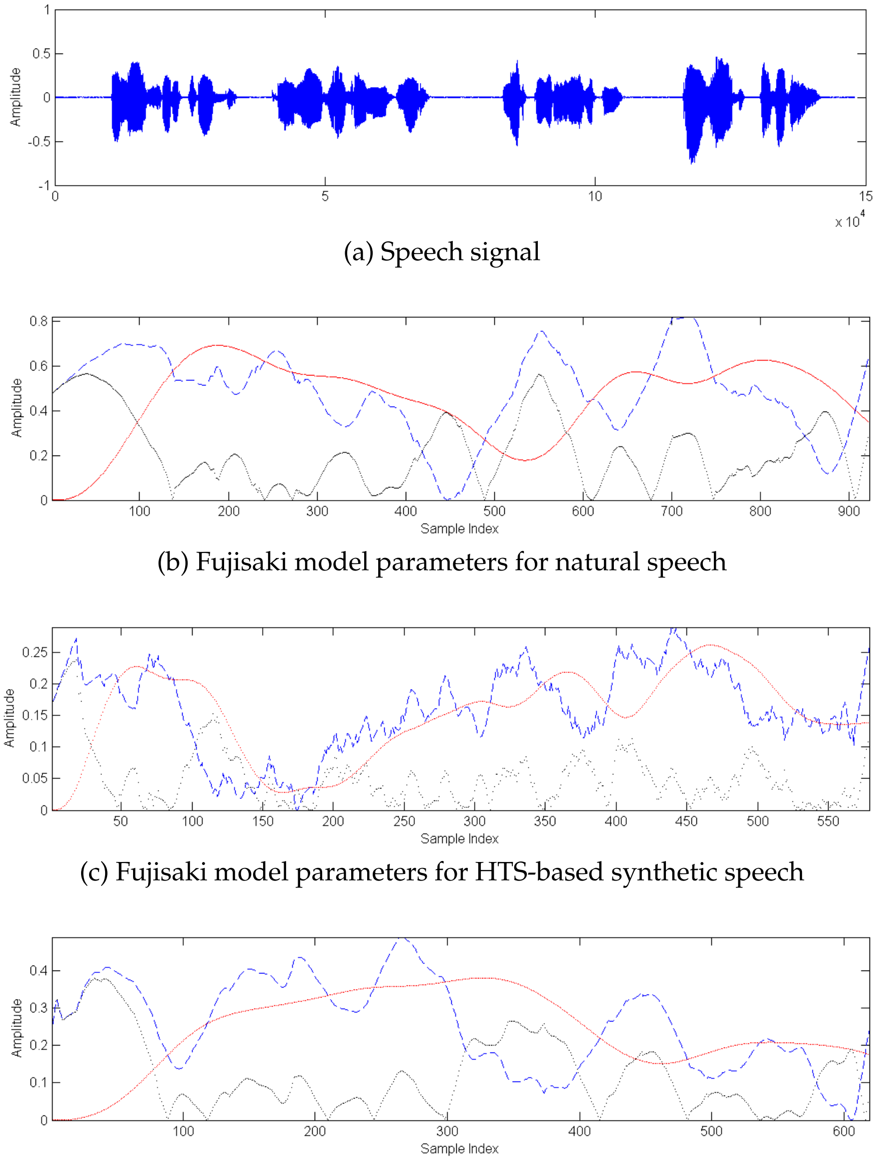

Speech signal
Etraction of ln(F0) contour
Interpolation and smoothening
Determining cutoff for Low pass filter
Finding phrase component

Figure 5.2: Algorithm for detection of phrase components

The extraction of F_0 contour was done using autocorrelation method and the cut off frequency for low pass filter was found from the spectrum of the contour. The F_0 contour, and corresponding low-pass component (LPC) and high pass component (HPC), representing phrase and accent respectively, obtained for natural, HMM-based and USS-based syntheticspeech are shown in figure 5.3

The negative peaks in the LPC corresponds to a phrase component. Three ob-



(a) Speech signal

(b) Fujisaki model parameters for natural speech

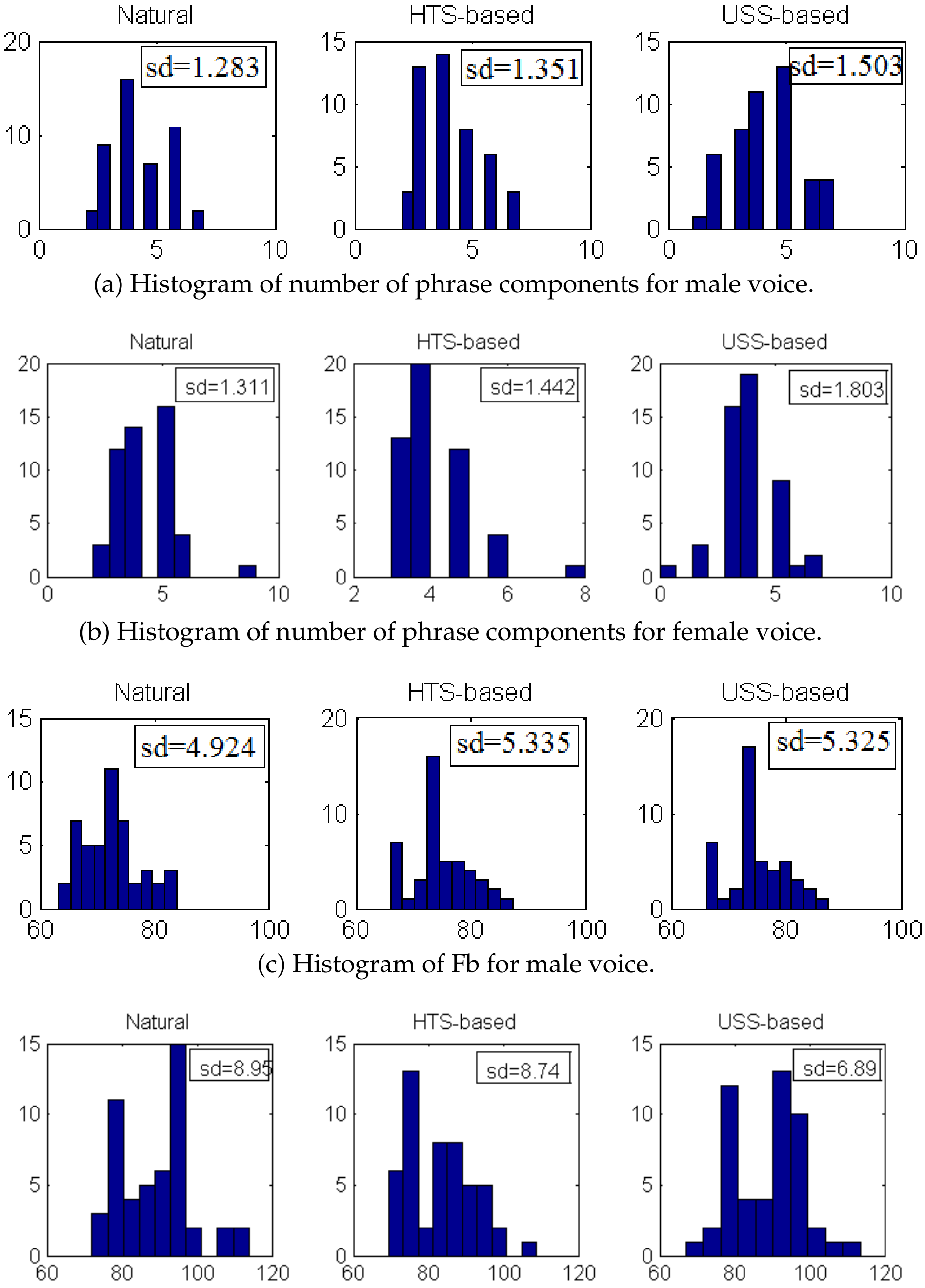
(c) Fujisaki model parameters for HTS-based synthetic speech

(d) Fujisaki model parameters for USS-based synthetic speech

Figure 5.3: Blue dashed lines- F_0 contour, red dotted line-LPC and black dotted line-HPC. In each figure, the x-axis shows time in *ms* and y-axis the magnitude.

servable parameters are the accent components, number of phrase components and the value of F_b. Figure 5.4 shows the distribution of number of phrase components and the value of F_b for 50 speech samples each of natural, USS-basd and HTS-based speech. The average value of number of phrase components of both type of synthetic speech are comparable. Natural speech has slightly higher mean of 4.4. The USS speech has distributed values and hence higher variance.

This is different from experiments conducted earlier which has shown closeness of USS-based synthetic speech to natural speech. Further the mean of speaker dependent factor F_b is highest for USS-based speech.This can be accounted for by



(a) Histogram of number of phrase components for male voice.

(b) Histogram of number of phrase components for female voice.

(c) Histogram of Fb for male voice.

(d) Histogram of Fb for female voice.

Figure 5.4: In each case the parameter value and corresponding frequency is plotted. The first histogram corresponds to natural speech, then HMM-based speech and that corresponding to USS-based speech is the right most. ′sd′ refers to the standard deviation of the data.

the fact that phrase occurs additionally at the concatenation points. However, due to the constrain added in counting the phrase components, that distance between adjacent components should be greater that pitch period, this effect is nullified.

5.4 Chapter Summary

Prosody is one of the most obvious factor that distinguishes natural speech for human ear. The human speech depends also on many psychological and emotional factors of the speaker. Fujisaki model is a widely acclaimed model for understanding prosody. Since the model incorporate the physiological aspects of the production mechanism, it provides a very detailed analysis. Because of this reason, the implications of the results obtained here are far reaching though the study conducted is very much confined to the problem of distinction. It is noted that the speaker-dependent characters of prosody, does not act as a robust method and further breaking down of different prosodic constrains is required for use of Fujisaki model parameters in real-life applications.

CHAPTER 6
Summary and Conclusions

Different features discussed in the study, potentially create flexibility in the analysis of natural speech and provides a rich variety of behaviours that can be utilized for more versatile performance. The results obtained also provides an evaluation of the TTS systems used in the study. It is seen that phase-based features are not captured in parametric modelling. The algorithm proposed using modified group delay is an improvement over the existing RPS based method in the sense that the computation is simpler. Though the method does not guarantee to work for USS speech or in that case any concatenated speech, the classification using MODGDF features is not seen to increase the false acceptance rate (FAR).

Further the initial assumption of chaotic nature of speech is validated. The positive LE indicates chaotic nature of the studied vowels in Gujarati language. Furthermore, the research illustrates the accountability of LE for distinction of synthetic and natural speech. Though the extraction of feature is at phoneme-level and the results are dependent on the phoneme used, with present day algorithms for phoneme extraction, LE would prove to improve the performance of ASV systems. The study needs to be extended to other phonemes and by incorporating LE as a feature in ASV system; it may be made robust to impostor attack using synthetic speech.

Final chapter of the thesis discusses prosodic aspects of the problem. Using Fujisaki model for parameter extraction, it is observed that though the TTS systems do not explicitly use these parameters, use of f_0 and pitch does capture aspects such as speaker information and high frequency components corresponding to accent. Unlike the observation for other feature, USS-based synthetic speech is seen to be more closer to HTS-based synthetic speech for the analysed prosodic aspects.

The study presented here is aimed at particular application of distinction of synthetic and natural speech in ASV systems. Additionally, exploring different aspects of speech signals has helped in understanding speech signal better. TTS systems often looks at the signal production from the aspect of hearing. However with the computational ability present today, finer details of the natural system of speech production can be studied in greater depths.

6.1 Future Work

1. An appropriate extension to the work presented in this thesis is implementing the proposed features on ASV systems. This would provide an evaluation of the robustness of the features proposed. Further, LE and MODGDF can be used to improve the TTS systems. Chaotic and prosodic features could be exploited in further understanding of natural speech.

2. The study has been carried out in low resource language of Gujarati. The results obtained can be verified for other Indian languages. In the present day context where multilingual processing is widely discussed and speech-based application are fast improving, the natural phenomenon of speech production and speech signal, still holds many aspects to be explored.

References

[1] M. Brodar, “"piscalka" iz divjih bab ni neandertalska" [the divje babe "flute" is not neanderthal], (in slovene),” September 2008.

[2] J. Allen, M. S. Hunnicutt, H. K. Dennis, C. A. Robert, and B. P. David, “From text to speech: The mitalk system,” *Cambridge University Press*, 1987.

[3] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, “Atr v-talk speech synthesis system,” pp. 483–486, 1992.

[4] I. N, K. N, and S. Y, “Concatenative speech synthesis by minimum distortion criteria,” *ICASSP ’92*, pp. 11–65–68, 1992.

[5] H. Andrew, J and B. Alan, W, “Unit selection in a concatenative speech synthesis system using a large speech database,” *IEEE International Conference On Acoustics, Speech, And Signal Processing,Icassp-96*, vol. 1, pp. 373 – 376, May 1996.

[6] J. Cammbell, P and JR., “Speaker recognition: A tutorial,” vol. 85, no. 9, pp. 1437–1462, September 1997.

[7] A. Patil, Hemant, “Speaker recognition in indian languages: A feature based approach,” Ph.D. dissertation, Indian Institute of Technology Kharagpur (IIT-K),, 2005.

[8] L. Leon, P. Michael, Y. Junichi, H. Inma, and I. Saratxaga, “Evaluation of speaker verification security and detection of hmm-based synthetic speech,” *IEEE Transactions On Audio, Speech, And Language Processing*, vol. 20, no. 8, pp. 2280 – 2290, October 2012.

[9] D. A Reynolds, T. F Quatieri, and R. B Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[10] B. L. Higgins, A. and J. Porter, “Speaker verification using randomized phrase prompting,” *DigitalSignalProcess*, vol. 1, pp. 89–106, 1991.

[11] K. Simon, “A beginnerâĂŹs guide to statistical parametric speech synthesis,” *Invited paper, Sadhana - Academy Proceedings in Engineering Sciences, Indian Institute of Sciences*, 2010.

[12] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, “A vector quantization approach to speaker recognition,” *in Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 387–390, 1985.

[13] J. Lindberg and M. Blomberg, “Vulnerability in speaker verification- a study of technical impostor techniques,” *In Proc of Eurospeech 99*, pp. 1211–1214, 1999.

[14] W. Zhizheng, C. Eng, Siong, and L. Haizhou, “Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition,” *Interspeech*, September 2012.

[15] W. Zhizheng, L. Anthony, L. Kong, Aik, C. Eng, Siong, K. Tomi, and L. Haizhou, “Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints,” pp. 950–954, 2013.

[16] T. F. Quatieri, “Discrete-time speech signal processing principles and practice,” 2002.

[17] A. M. Hema and Y. B, “Speech processing using group delay functions,” *Speech Processing ,ELSEVIER*, vol. 22, no. 3, pp. 259–267, March 1991.

[18] A. M. Hema and R. G. Venkata, Ramana, “The modified group delay function and its application to phoneme recognition,” *Proc.ICASSP '03*, vol. 1, pp. 68–71, April 2003.

[19] K. Iasonas and M. Petros, “Nonlinear speech analysis using models for chaotic systems,” *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 6, pp. 1098–1109, November 2005.

[20] A. D. P. Maragos and I. Kokkinos, “Some advances in nonlinear speech modelling using modulations, fractals, and chaos,” *Proc. Int. Conf. on DSP*, vol. 1, pp. 325–332, 2002.

[21] H. N. Teodorescu, F. Grigoras, and V. Apppei, “Analysis of nonlinear and nonstacionary processes in speech production,” *Proc. of: Appl. of Signal Processing to Audio and Acoustics*, vol. 5, no. 3, pp. 1453–1457, 1997.

[22] K. Murthy and B. Yegnanarayana, “Effectiveness of representation of signals through group delay functions,” *Signal Processing*, vol. 17, no. 2, pp. 141–150, June 1989.

[23] M. John, “Linear prediction: A tutorial review,” *proceedings of the IEEE*, vol. 63, no. 4, April 1975.

[24] K.T.Alligood and T. D. Sauer, “Chaos: An introduction to dynamical systems, 1st ed.” *Springer*, 1997.

[25] P. Grassberg and L. Procaccia, “Characterization of the strange attractors,” *Phys. Rev. Lett.*, vol. 50, no. 5, pp. 346–349, 193.

[26] H. Teager and S. Teager, “Evidence for nonlinear sound production mechanisms in the vocal tract,” *in Proc. NATO ASI on Speech Pmduchon and Speech Modelling*, pp. 241–261, 1990.

[27] M. Rosenstein, J. Collins, and C. C.J.De, “A practical method for calculating largest lyapunovs exponentes for small data sets,” *Physica D*, vol. 65, pp. 117–134, 1993.

[28] F. Takens, “Detecting strange attractors in turbulence,” *Lecture Notes in Mathematics*, vol. 898, p. 366, 1981.

[29] F. Hiroya, O. Sumio, and G. Wentao, “Physiological and physical mechanisms for fundamental frequency control in some tone languages and a command-response model for generation of their f0contours,” *In Proc. of International Symposium on Tonal Aspects of LanguagesâĂŤwith Emphasis on Tone Languages, Beijing*, pp. 61–64, 2004.

[30] O. Douglas, “Speech communication: Humans and machines,” *IEEE Computer Society Press*, November 1999.

[31] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for hmm-based speech synthesis,” *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP ’00.*, vol. 3, pp. 1315–1318, 2000.

[32] M. E. Hamon C. and C. F., “A diphone synthesis system based on time-domain modifications of speech,” *in Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 238 – 241, May 1989.

[33] V. Werner, C. Dirk, Van, and W. Patrick, “A unified view on synchronized overlap-add method for prosodic modification of speech,” *In proc. of thr international conference on spoken language processing, 2000*, vol. 2, pp. 63–66, 2000.

Chapter A
Hidden Markov Model(HMM)-based Speech Synthesis

Rather than problem of converting parameters into speech, the limiting factor in producing high quality speech is picking the right parameter for a synthesis specification. Hidden Markov Model(HMM) synthesis uses statistical machine learning techniques for this purpose. This has the advantage over synthesis by concatenation that the memory required for storing data is less and that modification of the models is possible. To quote Simon King, "No-one would claim that the HMM is a true model of speech. But the availability of effective and efficient learning algorithms (Expectation-Maximisation), automatic methods for model complexity control (parameter tying) and computationally efficient search algorithms (Viterbi search) make the HMM a powerful model." [11].

HMM is a finite-state machine which generates a sequence of discrete-time observations. At each time unit, the HMM changes its state depending on a state transition probability, and then generates observational data in accordance with an output probability distribution of the current state [30].A number of representations can be used as observations; a common set up is to use MFCCs, and F0 and their delta values, and perhaps additional information about the source. HMM with self state transition is simplistic and a better model of duration is required for high-quality speech synthesis. Once an explicit duration model is added to the HMM, we would only have a semi-Markov-transitions between states. Hence, we most often really mean HSMM speech synthesis [3].

The training phase feature vectors corresponding to different observations is defined and a separate model is trained for each unique feature combination and decision tree clustering is used to merge parameters for different states. The HMM is the concatenation of the models corresponding to the full context-label sequence, which has been predicted from text by the front end. Choice of duration model determines how many frames will be generated from each state in the model.

In the synthesis phase, the input text is analyzed to produce a sequence of full context labels. The sequence of models corresponding to this sequence of labels is then joined together into a single long chain of states. From this model, the linguistic parameters are generated and these are used to drive the output stage to produce a speech waveform. Generating parameters from the model is based on the principal of maximum likelihood. An example is the naive method as described in [11]. This method considers only the static parameters and generates the most likely observation from each state which is the mean of the Gaussian in that state. This would give piecewise constant parameter trajectories, that change value abruptly at each state transition and so this will not sound natural when used. This problem is solved by the MLPG algorithm [31] which also considers the speed with which parameters change value. MLPG finds the most likely sequence of generated parameters, given the distributions for different parameters.A number of refinements to the basic HMM technique have been proposed, including more realistic duration modelling and accounting for global variance.

The HMM speech used in the study is from Hidden Markov Model Toolkit (HTK) developed by the Machine Intelligence Laboratory of the Cambridge University Engineering Department. The system was trained using 5 hour duration studio recorded voice in Gujarati language sampled at *16KHz*.

CHAPTER B

Unit Selection-based Speech Synthesis

In this technique, synthetic speech is produced by concatenating the waveforms of units selected from large, single-speaker speech databases. The primary motivation for the use of large databases is that with a large number of units available with varied prosodic and spectral characteristics it should be possible to synthesize more natural-sounding speech than can be produced with a small set of controlled units [5].

The first stage of USS is creating a unit inventory. Choice of units could be, phoneme, diphone, syllable or their combinations. For instance, if the unit used is diphones, the set of all different diphones in the training data is acquired. Each recorded database is segmented into some or all units.

Typically the division into segments is done using speech recogniser with manual correction afterwards using visual representations such as waveform and spectrogram. An index of the units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency, duration, position in the syllable and neighbouring phones . Figure

After training the system, the desired target utterance is created by determining the best chain of candidate units from the database (unit selection).This is achieved using a weighted decision tree and catalogue that is created during training of the system. The input to synthesiser is typically text, though this may be augmented with structural and discourse information. In the first stage of synthesiser, the input is transformed into a string of phonemes or syllables depending on the text input, and annotate it with prosodic features (pitch, duration and power) which specify the desired speech output.the pitch and timing of each of the unit is modified to match the prosodic part of the specification. Some popular techniques for performing this are pitch-synchronous overlap and add(PSOLA), time-domain PSOLA(TD-PSOLA) [32], [33],residual-excited linear prediction(RELP) etc. These techniques, isolate the individual pitch periods of samples from training, modify them and resynthesise the waveform corresponding to input text.

With present day quality training data the key aspects for quality synthesis is ensuring continuity of acoustics features(spectral envelope, amplitude, fundamental frequency, speaking rate) at concatenation points without degrading the source and vocabulary independent synthesis with phone or syllable-like units. The USS synthesiser used in the entire study is the Festival system designed by Carnegie Mellon University's speech group trained using studio recorded speech for Gujarati language.