Speech Enhancement Using Voice Source Models

by

Anisa Yasmin

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical Engineering

Waterloo, Ontario, Canada, 1999

I hereby declare that I am the sole author of this thesis.

I authorize the University of Waterloo to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the University of Waterloo to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

Autoregressive (AR) models have been shown to be effective models for speech signals. However, although it is the most common model of speech, an AR process excited by white noise for speech enhancement, fails to capture the effects of source excitation, especially the quasi periodic nature of voiced speech. Speech synthesis researchers have long recognized this problem and have developed a variety of sophisticated excitation models. Such models have yet to make an impact in speech enhancement. We have concentrated our research on modifying the conventional white noise excited AR model for various speech classes and on establishing performance benchmarks by studying speech-enhancement, using the proposed models, in detail for individual phonemes under arbitrarily well-characterized circumstances.

We have proposed three different types of impulsive excitation models for an AR model for various phoneme classes based on the type of excitation with which each class is associated. For voiced speech, the effect of the glottal excitation is simulated by a train of impulses spaced according to pitch periods. For unvoiced stops and unvoiced affricates, the excitation source is modeled by a single impulse marking the instant of the onset of the burst and a white noise term. For voiced stops and voiced affricates, a mixed excitation of the plosive driving term and a quasi-periodic train of impulses are used. For voiced fricatives a mixed excitation of white noise and a quasi-periodic train of impulses separated by pitch periods is used. Due to the inclusion of impulsive driving terms in an AR model, the conventional AR parameter estimation techniques could not be used. We have proposed a novel AR parameter estimation technique for the models with impulsive excitations. In each case, impulsive AR models outperformed their white-noise-driven counterparts.

The success of the tentative impulsive excitation models has motivated us towards applying a more sophisticated excitation model. We have chosen one of the most common excitation source models, the four-parameter model of Fant, Liljencrants and Lin[1], which is also known as an LF model and applied it to the enhancement of individual voiced phonemes. We have proposed a novel two step optimization algorithm for estimating the parameters for an LF model. Among the AR models with three different types of excitation models (a conventional white-noise excitation, an impulsive excitation and an LF model), the LF excitation model yields the best performance in speech enhancement in terms of the output signal-to-noise ratios (SNRs).

Acknowledgments

I have received support from a number of people during my journey through the PhD program.

I am immensely grateful to my parents for planting the seed of higher education into me since my childhood. I don't think I have high enough words of reverence for their constant motivation, encouragement, love and guidance throughout my life. My only regret being their untimely passing away from this world before they could see my major accomplishments in life. May God bless their souls and I strongly believe that they are around me as my guardian angels.

I am greatly thankful to Prof. Li Deng for his supervision and financial support throughout this course of research.

I express my sincere gratitude to Prof. Paul Fieguth for closely supervising me. I am thankful to him for bearing with my numerous (could be trillion) unannounced visits to his office for any problem faced during my PhD. I am indebted to him not only for his invaluable insights on my research but also for being my mentor during the whole process.

I would like to thank the AUCC (Association of Universities and Colleges of Canada) and CBIE (Canadian Bureau for International Education) for offering me with a Canadian Commonwealth Scholarship for my MASc and PhD. I am also grateful to the department of Electrical and Computer engineering, University of Waterloo for financially supporting me after conclusion of my Commonwealth Scholarship.

I must thank my colleagues Gordon, Sorin and Bahman for their valuable support and deep insights into the philosophy of life of graduate students!

I surely have to mention the Minota Hagey residence, the U of W *den* for grad students where most of my social life was spent (more specifically the first floor kitchen, the central gossip headquarter!). Thanks to my very dear friends, Prasad and Jacinte, the other two musketeers from Minota Hagey, who accompanied me through my triumphs, sorrows, various mischiefs and most importantly our PhD-joint-venture! I must thank my friend Poonam for dauntingly bearing with me during my final year of PhD in her first year of PhD and for constantly making me conscious of my ignorance in Philosophy.

My many many gratitudes to my three wonderful brothers and sister-in-laws, who constantly have given me inspiration, support and direction, who have stood beside me after both of my parents have left this world.

Finally I would like to thank Wendy Boles, our graduate Secretary for her cooperation and the Bangladeshi Community in Waterloo who have given me a flavour of my home in Dhaka.

In the loving memory of my mom and dad,
Ayesha & Anisur Rahman.

Contents

List of Figures

List of Tables

Nomenclature

Symbols	Definitions
a_i	ith AR coeffcient
h_{ik}	transition probability for an HMM from the state i to k
$c_{m_t \mid s_t}$	mixture probability for an HMM for the mixture m_t given the state s_t
$b(x_t \mid m_t, s_t)$	observation probability for an HMM of the output vector x_t given the mixture m_t and the state s_t
i, j, k, l, n	indices
m_i	ith mixture variable for an HMM
q_w	variance of white noise process $\{w(t)\}$
q_v	variance of measurement noise process $\{v(t)\}$
$\mathbf{r}_{zx}$	cross-correlation vector of the noisy speech vector $\mathbf{z}^T(t)$ and the clean speech $x(t)$
$\mathbf{r}_{xx}$	auto-correlation vector for $\mathbf{x(t)}$
s_i	ith state variable for an HMM
t	time
$u(t)$	driving term at time t

Symbols	Definitions
$\{v(t)\}$	measurement noise process
$\{w(t)\}$	white noise process
$x(t)$	clean speech signal at time t
$z(t)$	noisy speech signal at time t
AR	auto-regressive
$\mathcal{E}[\cdot]$	expectation
$\mathbf{F}$	state-transition matrix for AR state-space model
$\mathbf{G}$	process matrix for AR state-space model
$\mathbf{H}$	observation matrix for AR state-space model
$H(j\omega)$	filter transfer function
$\mathbf{D}$	input distribution matrix for AR state-space model
lpc	linear predictive coefficents
MMSE	minimum mean squared estimate
MSE	mean square estimation
N_x	lpc order for the speech process $\{x(t)\}$
PDF	probability density function
$\mathbf{R}_{xx}$	auto-correlation matrix of the clean speech signal $\mathbf{x}(t)$
R_u	zero-lag autocorrelation of the excitation u_I
$S(\omega)$	power spectrum
SKM	Segmental k-Means algorithm
W_i	ith Wiener filter wieght
MAP	maximum a posterior
ML	maximum likelihood
N_S	number of states for an HMM
N_M	number of mixtures for an HMM
J	length of the speech signal

Symbols	Definitions
K	frame length
VQ	Vector Quantization
α	parameter for an LF model
β	parameter for an LF model
λ_x	parameter set $(\pi_{hmm}, h, c, a)_x$ for an AR HMM
	modelling the clean speech $\{x_t\}$
$p\lambda_x(x)$	PDF of an Gaussian AR HMM for the clean signal
λ_v	parameter set $(\pi_{hmm}, h, c, a)_v$ for an AR HMM
	modelling the noise $\{v_t\}$
$p\lambda_v(v)$	PDF of an Gaussian AR HMM for noise
σ_v^2	variance for $\{v(t)\}$
σ_w^2	variance for $\{w(t)\}$
$\mathbf{\Sigma_x}(t\|t)$	state estimation error covariance
$\mathbf{\Sigma_x}(t\|t-1)$	one step predicted error covariance
$\eta(t)$	innovation at time t
$\epsilon(t)$	estimation error
$\epsilon_{\mathbf{x}}(\mathbf{t}\|\mathbf{t})$	state estimation error
$\epsilon_{\mathbf{x}}(\mathbf{t}\|\mathbf{t}-\mathbf{1})$	one step predicted error of $\mathbf{x}(t\|t)$
κ	Kalman gain
$\Phi(i,k)$	cross-correlation matrix of clean speech x(t)
$\Psi(i,k)$	cross-correlation between clean speech and the excitation

Chapter 1

Introduction

This thesis deals with the problem of modeling speech for enhancement purposes. Our approach, in general, involves model-based speech enhancement [2] in which prior stochastic models of the clean speech and of the corrupting noise are used for estimation of clean (de-noised) speech from noisy speech. Clearly, accurate estimation requires that these models be robust and faithful representations of reality. By far the two most popular models for speech are Hidden Markov models (HMM) and white noise driven autoregressive (AR) models. We shall discuss the limitations of such models and enhancement systems based on such models. In this thesis, we shall focus our research entirely on modifying the white noise excited AR model based on the concept of the source-filter theory of speech production [3].

Section 1.1 of this chapter presents a general overview of speech enhancement research that has been carried out so far. Section 1.2 discusses motivations, objectives and contributions of this thesis. Finally, Section 1.3 outlines the organization of this thesis.

1.1 Overview of Speech Enhancement Techniques

Broadly speaking, the field of speech enhancement is interested in addressing three (not necessarily compatible) objectives [2]: (a) the improvement of the perceptual quality of noisy speech, (b) the immunization of speech encoders against input noise[4, 5], and (c) the improvement of the performance of speech recognition systems in the presence of noise[6, 7]. This thesis investigates the first of these. In our context, the speech enhancement problem concerns the estimation of "clean" (de-noised) speech $\hat{x}(t)$ from noisy speech $z(t)$. Speech enhancement has applications in a wide variety of speech communication contexts where the quality or the intelligibility of speech has been degraded by the presence of background noise. For example, cellular radio telephone systems are plagued not only by background noise but also by channel noise. Public telephones suffer from environmental disturbances of their location. Air-ground communication systems are corrupted with cockpit noise. Moreover the hearing impaired require an increase of between 2.5 and 12 dB signal-to-noise ratio to achieve similar speech discrimination capabilities to those of normal hearing [8]. These problems call for the use of speech enhancement.

Researchers have been working on devising an efficient way to extract clean speech from noisy speech for the last 30 years. Two broad divisions of speech enhancement techniques are non-parametric and parametric model based approaches [9]. One of the popular digital signal processing (DSP) non-parametric techniques for speech enhancement is spectral subtraction [10, 11]. In 1979, Lim and Oppenheim [12] presented an overview of contemporary speech enhancement techniques. They inferred that spectral subtraction was the most efficient in enhancing speech corrupted by uncorrelated additive noise. The spectral subtraction method estimates the Fourier transform of the clean signal by removing an estimate of the

power spectral density of the noise signal. The basic advantage of this approach is the implementation simplicity and low computational complexity[8]. One major drawback of this technique is the annoying nonstationary "musical noise" which is the residual noise consisting of narrow-band signals with time varying amplitudes and frequencies[2]. A number of modifications of the basic spectral subtraction approach have been proposed to alleviate the effects of the musical noise[2, 11, 13, 14]. Ephraim et al.[15] have proposed a signal subspace approach for speech enhancement. The basic principle of the signal subspace is to decompose the noisy signal space into a signal-plus-noise subspace and a noise subspace. After removal of the noise subspace, the clean signal is estimated from the remaining subspace. They have shown that the spectral subtraction is a special case of this approach. This work provides a theoretical basis for the spectral subtraction approach which is a special case of this signal subspace approach.

The parametric model based approaches have been well received in speech enhancement. One example of such models are AR models [16, 17, 18] which have widely been used for representing speech. Lim and Oppenheim [19] have used maximum *a posteriori* (MAP) estimation techniques for estimating AR parameters for the speech signal contaminated by additive white Gaussian noise. Hansen et al. [20] have used similar iterative MAP estimation techniques as in [19] followed by imposition of interframe and intraframe constraints upon the speech spectra. Such constraints introduce more speech-like formant trajectories and reduce frame-to-frame pole jitter and were applied using line spectral pair transformation of the AR parameters.

Hidden Markov modeling [21, 22, 2] is another common means of parametrically modeling speech. An HMM assumes that speech is composed of a set of statistically independent subsources, where each subsource represents a particular class

of statistically similar sounds [2]. The transition from one subsource to another is controlled by a first-order hidden Markov chain. The HMM based Wiener filter [23, 24, 8, 25] has been a popular choice for robust automatic speech enhancement. Ephraim et al. [23] have used a MAP approach that utilizes the expectation-maximization (EM) algorithm to estimate the clean signal from the noisy speech. Ephraim [24, 25] has used the minimum mean square error (MMSE) method that gives better enhancement results compared with that of MAP estimation which needs to iterate many times to achieve an acceptable result [8]. The MMSE based HMM is modified further by Sameti [8, 26] by incorporating multiple state-mixture based models for speech and noise. This model also incorporates the dynamic nature of the speech signal based on work done by Deng et al. [27, 28]. [29] uses cepstral domain modeling of speech and noise processes with MMSE method.

Dynamic filtering techniques, such as Kalman filtering, also provide a good estimate of clean speech given noisy speech. The Kalman filter is based on a state-space approach whereby a process state equation models the dynamics of the speech signal generation process and an observation state equation models the noisy signal. Paliwal et al. [30] have shown that a autoregressive (AR) model based Kalman filter and the delayed Kalman filter perform better than that of the stationary and the nonstationary Wiener filters. Gibson et al. [5] have implemented AR model based scalar and vector Kalman filters for both white and colored measurement noise assumptions for both speech enhancement and coding. As with any model based enhancement, the parameter estimation problem remains a big issue for AR model based Kalman filters when only noisy speech is available[31, 32, 33]. [31] uses power spectral density of speech signals to calculate the AR parameters. The EM algorithm has been used by [34, 32, 33] for iterative parameter estimation. Lee et al. [35] have proposed a Kalman filter algorithm with a hidden filter model (HFM)

of the clean speech signal. The HFM is an AR model with its parameters associated with first-order Markov chain.

In this thesis, we will review AR model based Kalman filter [30, 5] and HMM based Wiener filter [24, 25, 8], which provide us with insights for the problems associated with an AR and an HMM model for speech. In the following section, we will discuss the motivations that led to modifications of the conventional white noise excited AR model. We shall be considering the problem of enhancing speech corrupted by additive white noise. The proposed dynamical models will be used with Kalman filtering for estimating de-noised speech.

1.2 Thesis Motivations, Objectives and Contributions

In this section, we first present the motivations that fueled our interest in using voice source models for speech enhancement. We then discuss our objectives and follow with an outline of the contributions of this thesis.

1.2.1 Motivations

An AR model excited by a white noise process[16, 18] has traditionally been a favorite choice for modeling speech. One of the advantages of this type of AR model is the existence of efficient parameter estimation procedures known as linear predictive (LP) analysis. Secondly, a white noise excited AR model provides an approximate representation of all speech types, including voiced and unvoiced speech[16]. Finally, such AR models have a state space representation that can be

used with the Kalman filter algorithms for estimating de-noised speech from noisy speech. The main limitation of the white noise driven AR model is that it fails to take into account the effects of the voice source, especially in the case of voiced speech. This *flaw*, which is quite evident in quasi-periodic AR residuals, has been one of the motivations behind our interest in modifying a white noise driven AR model. The development of the source-filter theory of speech production initially proposed by Fant[3] also has an impact on our research. According to this linear speech production theory[36, 37, 38], the speech signal or pressure wave, measured at a microphone, is produced by the combined effects of the voice source excitation, vocal tract articulation and radiation from the lips or nostrils. This theory also provided good motivation for proposing different models for various speech types based on the nature of the associated excitation. The concept of the source-filter theory has been well utilized in speech analysis and synthesis. A precise and versatile model of the voice source is vital for production of natural sounding synthetic speech [39]. A number of deterministic voice source models have been proposed for speech synthesis and analysis [40, 41, 1]. Such deterministic models also provided good motivation for adding a source excitation model to the white noise driven AR model.

1.2.2 Objectives

One of the two main objectives of this thesis is to propose alternative appropriate models for various speech types. Another objective is to establish performance benchmarks or limits by studying speech-enhancement in detail for individual phoneme classes under arbitrarily well-characterized circumstances. For modeling the glottal excitation we shall be making explicit assumptions about known pitch locations for voiced speech. We shall be using clean speech for estimating AR

and Kalman filter parameters. Although such circumstances might appear artificial, they are essential in understanding the intrinsic factors which limit enhancement performance — an understanding which may improve enhancement algorithms in much broader, less constrained conditions.

1.2.3 Contributions

One of the contributions of this thesis is the comparative study of HMMs and AR models. We have investigated advantages and drawbacks of the state-of-the-art HMM based and AR model based enhancement systems. One of the significant contributions of this thesis are implementing impulsive models for individual phoneme classes. As each phoneme class has different production mechanism, we have proposed and implemented three types of Impulsive AR models which include impulsive driving terms which are tentative models for various types of excitation sources. In the first type of model, the glottal excitation, for voiced speech such as vowels, semivowels, diphthongs and nasals, is modeled by a train of impulses spaced according to pitch periods. The second impulsive model, for voiced stops and voiced affricates, models the voiced excitation by an impulse train and the plosive excitation by a single pulse marking the onset of the burst and white noise. The third model, for unvoiced stops and unvoiced affricates, uses the plosive excitation term and white noise.

Due to inclusion of an impulsive term, the conventional AR parameter estimation procedure needed to be modified. We have also proposed a novel AR parameter estimation procedure that takes account of the impulsive driving term and estimates the AR parameters and the amplitudes of the individual impulses. We have demonstrated the appropriateness of our models by applying such models to a wide

variety of phonemes. We also have clearly demonstrated the limits to performance for Kalman filter based enhancement by making a number of model assertions and parameter assumptions. Impulsive models have shown remarkable improvements in output signal-to-noise ratios (SNRs) over the conventional AR model driven by white noise.

The success of impulsive AR models over the conventional white noise driven AR models has motivated us to use a more sophisticated model for the voice source. Another significant contribution is the proposal of using an LF based model for voiced speech. An LF model is based on the glottal model proposed by Fant, Liljencrants and Lin[1]. The LF model has been well received in speech synthesis and analysis for a long time, but it has yet to make an impact on speech enhancement. Parameter estimation problems associated with an LF model for speech enhancement are completely different from those of speech synthesis. A novel parameter optimization algorithm has been proposed for LF models for speech enhancement. The optimization algorithm gives the LF parameter estimates in two steps. The initial step gives the initial estimates using a Minimum Finder Algorithm (MFA). The latter step uses the initial estimates to find the final estimates using a Grid Search Algorithm (GSA) via co-ordinate optimization. For estimating the AR parameters we have used similar parameter estimation procedure proposed for impulsive models only in this case the driving term being an LF model. We have clearly demonstrated the applicability of an LF model, over an impulsive or a conventional white noise model, as an excitation model for voiced speech. Finally, we also discuss the limits to performance for an LF model based Kalman filter.

1.3 Thesis Organization

Chapter 2 presents the background relevant to impulsive and LF model based AR models for speech enhancement. It begins with a brief introduction to the type of speech enhancement problem we are interested in this thesis. It then briefly introduces the white noise excited AR model and the Wiener filter. The next two sections discuss HMM and Kalman filter based systems. Next, we review the anatomy and physiology of the human speech production system followed by the discussion of the phonemes used in North American English. Finally, we present a concise description of the TIMIT database used to supply the speech data for enhancement.

Chapter 3 proposes and implements impulsive AR models for speech enhancement. We begin by discussing the drawbacks of a white noise excited AR model, followed by a review of the performance of the two state-of-the-art speech enhancement systems: AR model based Kalman filter and HMM based Weiner filter. The next section reviews the production mechanism of various phoneme classes. We propose impulsive AR models for various phoneme classes in the next section. The next three sections discuss the Kalman filter algorithms, the AR parameter estimation techniques for impulsive models and the model assertions and assumptions. Finally, enhancement results for impulsive models are presented and discussed.

Chapter 4 concentrates on proposing and implementing an LF excitation model for voiced speech enhancement. This chapter begins by motivating a need for more sophisticated excitation models compared to tentative impulsive models. The next section reviews some of the voice source used in speech synthesis, and discusses the feasibility of the LF voice source model for speech enhancement, followed by a discussion of an LF model. Next, we propose an optimization procedure for LF

parameter estimation. Finally, the results for LF model based enhancement are presented and discussed.

Chapter 5 summarizes the results of this thesis and presents a number of directions for future research.

Chapter 2

Background

The main objective of this chapter is to motivate a foundation for voice source model based enhancement. Section 2.1 discusses the speech enhancement problem in general while Section 2.2 presents the white noise driven autoregressive model most commonly used in speech processing. As mentioned in Chapter 1, we shall be comparing performances of the state-of-the-art HMM based Wiener filter and AR model based Kalman filter systems in Chapter 3. Section 2.3 therefore briefly mentions Wiener filtering before reviewing an HMM based system in Section 2.4. Section 2.5 is dedicated to the autoregressive model based Kalman filter. As we shall be proposing voice source model based enhancement system (in Chapters 3 and 4) based on the production mechanism of the smallest speech units (phonemes or phones), we discuss the speech production system in Section 2.6. We briefly mention the phonemes used in North American English in Section 2.7. Finally, Section 2.8 discusses the TIMIT database in a concise manner as we shall be using the speech data from the TIMIT database.

2.1 Speech Enhancement Problem

Speech enhancement deals with minimizing effects of noise on speech by improving the perceptual quality of noisy speech, improving the performance of machine recognizers in a noisy environment or immunizing speech coders against input noise[2]. As mentioned in Chapter 1, in this thesis we shall be dealing with only the first type of speech enhancement problem. Let $\{z(t)\}, z(t) \in \Re$ be a random process modeling the noisy speech. Let $\{x(t)\}, x(t) \in \Re$ denote a random process modeling the clean speech while let $\{v(t)\}, v(t) \in \Re$ be a random process representing measurement noise modeled as a Gaussian white noise. Let us assume that

$$z(t) = x(t) + v(t) \quad 0 \leq t \leq T \tag{2.1}$$

and that $\{x(t)\}$ and $\{v(t)\}$ are statistically independent and that $\{v(t)\}$ is a white Gaussian process with a zero mean and a variance of σ_v^2. The speech enhancement problem, in the context of this thesis, concerns the estimation of the clean speech $x(t)$ from the noisy speech $z(t)$, given a model for clean speech and a model for noisy speech. As pointed out by (2.1), in this thesis we shall only be dealing with additive noise, or more specifically, with additive Gaussian white noise. The two models, used for representing speech, discussed in this chapter, are an Autoregressive (AR) model and an Hidden Markov model (HMM). Our research is concentrated on modifying the excitation term associated with an AR model. We present in the following section, a detailed discussion of an AR model with white noise excitation.

2.2 Autoregressive (AR) Speech Model

In this section, we present an AR or an all-pole model which is one of the most popular models for representing speech waveforms[18, 16, 42]. This model is based

on an acoustic analysis of the speech production system[16, 43]. The popularity of the AR model stems from its simplicity, and because the human vocal tract during voicing can be modeled by an all-pole system[16]. Furthermore, although unvoiced speech and nasals introduce zeros into the system, since the zeros of the transfer function of the vocal tract lie inside the unit circle, [16] they can be approximated by an all-pole system with sufficiently many poles. In such model, a speech sample is approximated as a linear combination of past speech samples and a white noise term. Let us assume that the clean speech sequence $x(t)$ is generated according to an $N_x th$ order AR model,

$$x(t) = \sum_{i=1}^{N_x} a_i x(t-i) + w(t) \tag{2.2}$$

where $w(t)$ is a zero mean, white Gaussian process with variance σ_w^2 and a_i is the ith AR coefficient. $w(t)$, is also known as process noise.

Another advantage of an AR model is that its parameters can be estimated accurately using the method of linear predictive(LP) analysis. Among many formulations of LP analysis, the covariance method [16, 18] and the autocorrelation method [18, 21, 16] have been used extensively in speech processing. It has been shown in [16] that a predictor order of 12 gives a reasonable estimate for all speech types. Finally, because (2.2) can be rewritten in state-space form, the Kalman filter can be used to compute the optimal estimates $\hat{x}(t)$ of $x(t)$ [30, 5].

2.3 Wiener Filter

Wiener filter based enhancement algorithms have been used widely in speech because of the simplicity of implementing the Wiener filter[12, 44, 2]. A Wiener filter, represented by the coefficient vector $\mathbf{W}$ accepts a noisy signal $z(t)$ and yields the

minimum mean squared estimate (MMSE) $\hat{x}(t)$ of a desired signal $x(t)$. An Optimum solution for the coefficients is obtained by mean square estimation (MSE) only when the input signal is stationary. The filter output is given by

$$\hat{x}(t) = \sum_{i=0}^{J-1} W_i z(t-i) = \mathbf{W}\mathbf{z}^T(t) \tag{2.3}$$

where the filter input vector $\mathbf{z}(t)^T = [z(t), z(t-1), \ldots, z(t-J-1)]$ and the Wiener filter coefficient vector $\mathbf{W}^T = [W_0, W_1, \ldots, W_{J-1}]$. The estimation error signal is given by

$$e(t) = x(t) - \hat{x}(t) = x(t) - \mathbf{W}\mathbf{z}^T(t) \tag{2.4}$$

while the mean squared estimation error is given by,

$$\mathcal{E}[e^2(t)] = \mathcal{E}[x(t) - \mathbf{W}\mathbf{z}^T(t)]^2 = \mathcal{E}[x(t)^2] - 2\mathbf{W}\mathbf{r}_{zx} + \mathbf{W}\mathbf{R}_{zz}\mathbf{W}^T \tag{2.5}$$

where $\mathcal{E}[\cdot]$ denotes the expectation, $\mathbf{R}_{zz} = \mathcal{E}[\mathbf{z}^T(t)\mathbf{z}(t)]$ is the correlation matrix of the noisy signal and $\mathbf{r}_{zx} = \mathcal{E}[\mathbf{z}^T(t)x(t)]$ denotes the cross correlation vector of the desired and input noisy signals. The coefficients of the filter, obtained by minimizing the mean squared error $\mathcal{E}[e^2(j)]$ with respect to filter coefficient vector $\mathbf{W}$, are given by

$$\mathbf{W} = \mathbf{R}_{zz}^{-1}\mathbf{r}_{zx} \tag{2.6}$$

The system of equations in (2.6) are known as Wiener-Hopf equations [45].

The basic Wiener-Hopf equations in (2.6) can only be applied to stationary signals. For nonstationary speech signals, a number of methods have been proposed based on short-time power spectra [8, 30]. These nonstationary Wiener filters can be used only for the signals which are stationary over a small segment of time. We shall not go into details on frequency domain formulation of Wiener filter theory [45, 44] but rather just present the formulation frequently used in speech enhancement. For

the *uncorrelated* signal $x(t)$ with the noise $v(t)$ in (2.1), the autocorrelation matrix $\mathbf{R}_{zz}$ of the noisy signal is given by,

$$\mathbf{R}_{zz} = \mathbf{R}_{xx} + \mathbf{R}_{vv} \tag{2.7}$$

and the cross correlation vector $\mathbf{r}_{zx}$ is expressed as

$$\mathbf{r}_{zx} = \mathbf{r}_{xx} \tag{2.8}$$

where $\mathbf{R}_{xx}$ and $\mathbf{R}_{vv}$ respectively are autocorrelation matrices for clean signal and noise. Substituting (2.7) and (2.8) into Wiener-Hopf equations (2.6), gives an optimal linear filter coefficients for noise filtering,

$$\mathbf{W} = (\mathbf{R}_{xx} + \mathbf{R}_{vv})^{-1}\mathbf{r}_{xx} \tag{2.9}$$

Applying a Fourier transform to both sides of (2.9), we obtain a very useful formulation of Wiener filter used in speech enhancement, the transfer function $H(\jmath\omega)$ for a Wiener filter is given by [8, 12],

$$H(\jmath\omega) = \frac{S_{xx}(\omega)}{S_{xx}(\omega) + S_{vv}(\omega)} \tag{2.10}$$

where $S_{xx}(\omega)$ and $S_{vv}(\omega)$ denote the clean speech and noise power spectra. We shall be using this form of Wiener filter given by (2.10) with the hidden Markov model based enhancement system which is discussed in the following section.

2.4 Hidden Markov Model (HMM) Based Speech Enhancement

One popular parametric statistical model for speech is the hidden Markov model (HMM) [21, 23, 24, 25]. The main reason behind their popularity, is that HMMs

are the most general models known for speech processing. An HMM is a composite statistical model composed of statistically independent subsources where each subsource may represent various phonemes or different configurations of the vocal tract[2]. In Section 2.4.1, we define some of the very basic concepts of stochastic processes that form the basis for an HMM followed by a presentation of a complete parameter set for an HMM. Next in Section 2.4.2, we explain the HMMs specific to speech recognition and enhancement and define the parameter sets for speech enhancement. Sections 2.4.4 and 2.4.5 respectively are dedicated to the training and the filtering procedures involved in an HMM based MMSE (with Wiener filters) speech enhancement.

2.4.1 Hidden Markov Models (HMMs)

An HMM is a double layered finite state stochastic process where the selection of states of an observable process is governed by a hidden Markov chain. A first order Markov chain is a stochastic process where the conditional distribution of any future state, given the past states and the present state, is independent of the past states and dependent only on the present state [46]. In an HMM, each state dependent probability distribution (PD) can be chosen to be a mixture of Gaussian or any other type of PD. Each Gaussian mixture can be further assumed to be an AR process of the given order for convenient parameterization of the covariance matrices characterizing the Gaussian subsources[2]. Based on the state-to-state transition, there are various configurations for HMMs. One such configuration is an ergodic HMM. By ergodic HMM we mean that every state can be reached from every other state of the model in a finite number of steps[21]. Figure 2.1 shows a three state ergodic or fully connected HMM structure. Another example of the HMM structure is the left-right

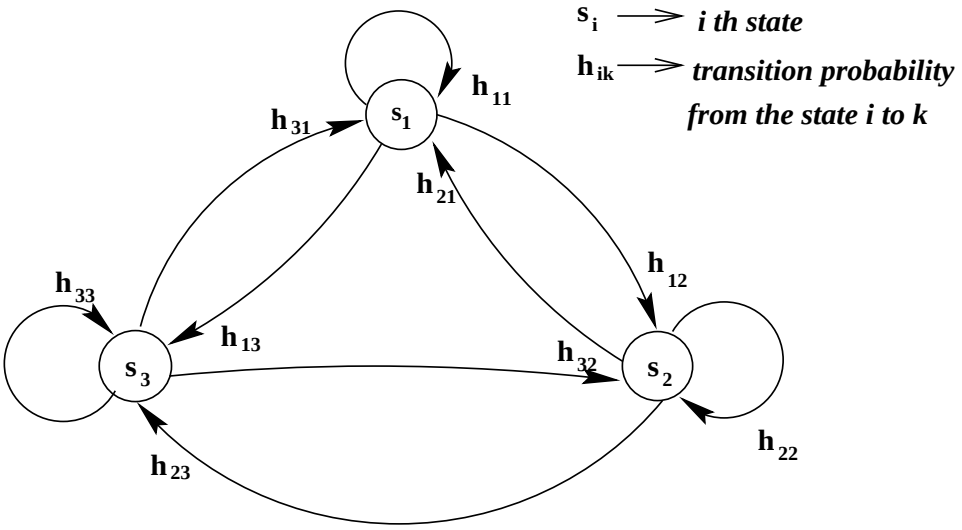

Figure 2.1: A fully connected three state HMM structure.

model where transitions are allowed only from a left to a right state. An HMM is usually characterized by the number of states, the number of mixtures, the initial transition probability, the transition probabilities for one state to another and the mixture coefficients.

Let us present the parameters that are used to characterize an ergodic auto-regressive (AR) HMM [8]:

- N_S, the number of states $[S_1, S_2, \ldots, S_{N_S}]$ in the model.
- N_m, the number of mixtures $[M_1, M_2, \ldots, M_{N_m}]$ per state.
- The set of initial probability distributions, $\pi_{hmm} = \{\pi_i\}$ where

$$\pi_{s_0} = P(s_0 = S_i), \quad 1 \leq i \leq N_S \tag{2.11}$$

where s_0 is the state at time 0.

- The set of the state transition probabilities, $h = \{h_{s_{t-1}s_t}\}$ where

$$h_{s_{t-1}s_t} = P(s_t = S_j | s_{t-1} = S_i), \quad 1 \leq i, j \leq N_S \tag{2.12}$$

  where s_t is the state at time t.

- The set of mixture weights, $c = \{c_{m_t|s_t}\}$ where

$$c_{m_t|s_t} = P(m_t = M_k | s_t = S_j), \quad 1 \leq k \leq N_m, \ 1 \leq j \leq N_S \tag{2.13}$$

  where $c_{m_t|s_t}$ expresses the probability of choosing the mixture m_t given that the process is in state s_t.

- $a = \{a_{k|j}\}$ with $a_{k|j}$ being the AR parameter set of a zero-mean $N_x th$ order Gaussian AR output process corresponding to state and mixture pair (j, k), where $a_{k|j} = \{a_{k|j}(0), a_{k|j}(1), ..., a_{k|j}(N_x), \sigma^2_{k|j}\}$, $a_{k|j}(0) = 1$, $\sigma^2_{k|j}$ being the variance for $i, j = 1, 2, ..., N_S$ and $k = 1, 2, ..., N_m$.

- Let $\lambda = (\pi_{hmm}, h, c, a)$ be the parameter set for a Gaussian AR HMM.

2.4.2 HMMs for Speech Processing

HMMs have long been used as a reliable statistical model for speech as it can model the nonstationary nature of speech by transitions between different states. A large number of states can be used to represent different spectral prototypes of speech. As mentioned earlier a state dependent probability density can be chosen to be a mixture of Gaussian probability densities. An advantage of such representations is that we get finer models of speech data[2]. In the case of speech recognition, a separate left-right model is used to characterize the temporal structure of every speech unit which may be a phoneme or a word[8]. As each model contains the

ordered sequence of stochastic properties corresponding to a particular speech unit, transitions from a higher indexed to a lower indexed state is prohibited. In a left-right model similar speech properties, i.e., similar speech units occurring in different frames of time, are assigned to different states depending on the context. The objective in speech recognition is to find models with maximal separation so that they give as different likelihoods as possible for pattern recognition purposes.

For speech enhancement, we have different objectives and thus the modeling problem is different from that of speech recognition[8]. We need two distinct models for clean speech and for noise to estimate the de-noised speech from the noisy speech. We require that an HMM extracts the general spectral properties of clean speech regardless of the phoneme, word or sentence. Such types of HMMs are necessary to differentiate the speech from the noise. Hence, we need to only model the averaged out or the global characteristics of speech and noise. Unlike speech recognition there are no constraints on the transition probabilities in enhancement models i.e. an ergodic HMM model can be used for enhancement. HMMs can also be used to model a wide variety of noise encountered in practice[2].

Now for speech enhancement, let us apply the generic ergodic HMM parameter set (described earlier) to define HMM parameter sets for clean speech and noise. Thus let $\lambda_x = (\pi_{hmm}, h, c, a)_x$ be the parameter set for a Gaussian AR HMM modeling the clean speech x while $\lambda_v = (\pi_{hmm}, h, c, a)_v$ be the parameter set for a Gaussian AR HMM modeling the noise v.

There are two steps in HMM based enhancement (details in Chapter 3). First HMMs are trained (discussed in the following section) for clean speech and noise. Then the noise model together with the clean speech model is used to filter out the noise from the noisy signal. Two distortion measures commonly used in HMM based speech enhancement are the minimum mean square error (MMSE) [24, 25, 8]

and the maximum a posteriori (MAP) [23] estimation. It has been shown in [2, 8] that MMSE estimation has computational advantages over MAP estimation based enhancement. Thus we shall only discuss the MMSE estimation associated with HMM as proposed in [24].

2.4.3 Problem Formulation for HMM based Enhancement

Let $z \stackrel{\triangle}{=} \{z_t, t = 0, \ldots, T-1\}, z_t \in \Re^K$ be a sample function of the noisy speech, where $\Re^K$ represents K dimensional Euclidean space (frame-length). Let $x \stackrel{\triangle}{=} \{x_t, t = 0, \ldots, T-1\}, x_t \in \Re^K$ and $v \stackrel{\triangle}{=} \{v_t, t = 0, \ldots, T-1\}, v_t \in \Re^K$ respectively represent sample functions of the clean speech and the noise process. z, x and v are related to each other according to equation 2.1, where x and v are statistically independent. Let $p\lambda_x(x)$ and $p\lambda_v(v)$ be the PDF of a Gaussian AR HMM for the clean signal and noise respectively. Let $s \stackrel{\triangle}{=} \{s_t, t = 0, \ldots, T-1\}, s_t \in 1, \ldots, N_S$, be a sequence of states corresponding to x and let $m \stackrel{\triangle}{=} \{m_t, t = 0, \ldots, T-1\}, m_t \in 1, \ldots, N_m$ be a sequence of mixtures corresponding to (s, x).

2.4.4 Training HMMs for Clean Speech and Noise

The PDF of a Gaussian AR HMM for the clean signal $p\lambda_x(x)$ is given by [24, 8],

$$p\lambda_x(x) = \sum_s \sum_m p\lambda_x(s, m, x) \tag{2.14}$$

$$= \sum_s \sum_m \{\prod_{t=0}^{T-1} h_{s_{t-1}s_t} c_{m_t|s_t} b(x_t|m_t, s_t)\} \tag{2.15}$$

where $b(x_t|m_t, s_t)$ is the PDF of the output vector x_t given (m_t, s_t). For N_xth order AR process with zero mean, if $K \gg N_x$, we have [47],

$$b(x_t|m_t, s_t) = \frac{\exp\{-B/(2\sigma_{i,j}^2)\}}{(2\pi\sigma_{k,j}^2)^{K/2}} \tag{2.16}$$

where B is the autocorrelation function defined as,

$$B \stackrel{\triangle}{=} r_t(0)r_{k,j}(0) + 2\sum_{n=1}^{N_x} r_t(n)r_{k,j}(n), \tag{2.17}$$

$r_t(n) = \sum_{l=0}^{K-n-1} x_t(l)x_t(l+n)$ and $r_{k,j}(n) = \sum_{l=0}^{N_x-n-1} h_{k,j}(l)h_{k,j}(l+n)$ are autocorrelation sequences for $1 \leq k \leq N_m$ and $1 \leq i, j \leq N_S$.

As we have defined a complete parameter set $\lambda_x = (\pi_{hmm}, h, c, a)_x$ for an AR HMM process for clean speech, we are now left with the problem of given a training sequence x, how do we obtain a maximum likelihood (ML) estimate of the parameter set $\lambda_x = (\pi_{hmm}, h, c, a)_x$, that is,

$$\max_{\lambda_x} \ln p_{\lambda_x}(x) = \max_{\lambda_x} \ln \sum_s \sum_m p_{\lambda x}(s, m, x) \tag{2.18}$$

This maximization is usually carried out by using the Baum algorithm [24]. An approximate maximum likelihood estimate of the parameter set can also be obtained using segmental k-means method [8] when the double sum in (2.18) is dominated by a single state and mixture sequence and the parameter set can be maximized along that dominant sequence, that is,

$$\max_{s,m,\lambda x} ln \sum_s \sum_m p_{\lambda x}(s, m, x). \tag{2.19}$$

The segmental k-means method being computationally more efficient than the Baum algorithm [8], we have used it in our work for parameter estimation for HMMs. A good initial model is required for the k-means reestimation method as it only computes a local maximum of an objective function (given by (2.19)). The initial model for segmental k-means method is obtained from vector quantization of the training data using the generalized Lloyd algorithm(GLA) with the Itakura-Saito distortion measure[8, 48]. The training procedure for HMMs consists of two main steps, namely, vector quantization (VQ) and segmental k-means (SKM)

method for estimating the parameter sets λ_x and λ_v for the clean speech and noise models respectively.

Vector Quantization (VQ)

The generalized Lloyd algorithm is used to design a $(N_S \times N_m)$ VQ code-book for an HMM with N_S states and N_m mixtures. Codewords are successively split, starting from the centroid of the training data, until an N_S entry code-book is obtained. Each code-word consists of the AR parameter set and the gain term associated with them. In each step, the code-word with the largest residual energy is selected to be split by perturbing by two small values to obtain two new AR models. To ensure the stability of the perturbed models, the reflection coefficients associated with the AR models are first calculated and then multiplied by two numbers close to 1, finally the corresponding AR models are obtained from these perturbed reflection coefficients [8]. After each perturbation GLA (details in [48]) is used to optimize the code-book. This process of splitting and optimization is carried out until desired size N_S is reached.

The mixtures within each state codeword are determined using the same iterative procedure with the AR models initially in the parent partition. Thus, an initial estimate for the AR parameters of N_S state and N_m mixture HMM is obtained. Then the training data is clustered using the estimated code-words and then the initial estimate for (π_{hmm}, h, c, a) parameters is obtained from the relative frequencies at which the initial state, state transition and mixture component are chosen.

Segmental k-Means Algorithm (SKM)

Here we discuss the algorithm for modeling with N number of training sequences of speech data. Given N training sequences of speech data, an approximate maximum likelihood estimate of the parameter set λ_x is obtained using the SKM algorithm [49]. The parameter set is estimated along with the most likely sequence of states and mixture components. The objective function we maximize in this case is [8],

$$\ln p_{\lambda x}(s, m, x) = \sum_{n=1}^{N} \ln p_{\lambda x}(s_n, m_n, x_n). \tag{2.20}$$

The maximization of (2.20) is carried out in two stages. First starting with an initial model λ_x, the optimal state and mixture sequence (s_n^{opt}, m_n^{opt}) for the nth token is obtained using a Viterbi algorithm [21]. Viterbi search gives the optimal state and mixture sequence for each training sequence using the following metric path,

$$\begin{aligned} &\ln \pi_j + \ln c_{k|j} + \ln b(y_{0,n}|m_{0,n} = k, s_{0,n} = j) \quad t = 0 \\ &\ln a_{i,j} + \ln c_{k|j} + \ln b(y_{t,n}|m_{t,n} = k, s_{t,n} = j) \quad 1 \leq t < T_n \end{aligned} \tag{2.21}$$

Once the optimal path is obtained the model parameters are re-estimated in the second stage of maximization. The parameter re-estimation formulae being,

$$\pi_j^{'} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{N_m} P_{0,n}(j,k) \tag{2.22}$$

$$a_{i,j}^{'} = \frac{\sum_{n=1}^{J} \sum_{t=1}^{T_n - 1} \sum_{k=1}^{N_m} P_{t,n}(i,j,k)}{\sum_{j=1}^{N_S} \sum_{n=1}^{N} \sum_{t=1}^{T_n - 1} \sum_{k=1}^{N_m} P_{t,n}(i,j,k)} \tag{2.23}$$

$$c_{k|j}^{'} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n - 1} P_{t,n}(j,k)}{\sum_{k=1}^{N_m} \sum_{n=1}^{N} \sum_{t=1}^{T_n - 1} P_{t,n}(j,k)} \tag{2.24}$$

where $P_{t,n}(j,k)$ is the probability of being in state j and choosing mixture k at time t given the model λ_x and x_n while $P_{t,n}(i,j,k)$ is the probability of transition

from state i at time $t-1$ to state j at time t given the model λ_x and x_n. These probabilities can be calculated using the forward-backward algorithm[23]. This process of alternative maximization is carried out until the convergence of (2.20).

2.4.5 HMM Based Minimum Mean Square Error (MMSE) Enhancement

The HMM based MMSE enhancement system [8, 24] we are going to discuss, uses a multiple state and mixture noise model to accomodate non-stationarity in noise. The system is designed to determine an estimate $\hat{x}_t$ of clean speech x_t, where

$$\begin{aligned} \hat{x}_t &= \mathcal{E}\{x_t|z_0^t\} \\ &= \int x_t p(x_t|z_0^t) \\ &= \frac{\int x_t p_{\lambda s}(x_0^t) p_{\lambda v}(z_0^t - x_0^t) dx_0^t}{\int p_{\lambda s}(x_0^t) p_{\lambda v}(z_0^t - x_0^t) dx_0^t} \end{aligned} \tag{2.25}$$

We shall not go into the detailed derivation of $\hat{x}_t$ which is estimated using the forward algorithm in [24] but rather present the solution as,

$$\hat{x}_t = \sum_{j=1}^{N_S x} \sum_{k=1}^{N_m x} \sum_{\xi=1}^{N_S v} \sum_{\delta=1}^{N_m v} P_t(j,k,\xi,\delta|z_0^t) \mathcal{E}\{x_t|z_t, s_{t_x}=j, m_{t_x}=k, s_{t_v}=\xi, m_{t_v}=\delta\}, \tag{2.26}$$

where $P_t(j,k,\xi,\delta|z_0^t)$ is the posterior probability of speech state j and mixture k, and noise state ξ and mixture δ at time t given the noisy signal z_0^t.

In (2.26) we notice that the MMSE estimator of x_t given z_0^t is a weighted sum of the individual MMSE estimators of the output processes generated by the clean speech HMM, where the weights are the probabilities that the individual estimators are the correct ones for the given noisy signal [8]. The exact evaluation of $\mathcal{E}\{x_t|z_t, s_t=j, m_t=k, n_t=\xi, p_t=\delta\}$ is not trivial. It has been shown by Ephraim

[24] that if the variances of the innovation process of the AR sources are assumed to be circulant, $E\{g(x_t)|z_t, s_t = j, m_t = k, n_t = \xi, p_t = \delta\}$ can be given by the Wiener filter where $x_t = \{X_t(k), k = 0, 1, ..., K-1\}$, $X_t(k)$ being the kth component of the discrete Fourier transform(DFT) of x_t.

There are two major problems in HMM for speech enhancement [8]. First, such a model requires a large number of states to accomodate rapidly varying speech signals. This increases the computational complexity together with the risk of affecting the performance of HMMs for slowly varying speech signals. The second problem with HMMs is that since a constant mean is assumed for the observation probability within each state and different states have different mean values, the continuity of speech features is affected. We shall be discussing the performance of MMSE based HMM speech enhancement in the following chapter.

2.5 The Kalman Filter Based Speech Enhancement

The fact that AR state-space models for speech can be used with the Kalman filter has given good motivation for using the Kalman filter for speech enhancement[5, 30]. As we shall be using, the Kalman filter as an estimator, in next two chapters, we present an extensive derivation of the Kalman filter algorithms[50, 45, 51]. The Kalman filtering problem for a linear dynamic system is formulated in terms of two basic equations: the process equation that describes the dynamics of the system in terms of a state vector and the measurement equation that describes measurement noise incurred in the system.

Let an N_x-dimensional parameter vector $\mathbf{x}(t)$ denote the state of the discrete-time, linear, dynamical system and let $z(t)$ denote the observed data of the system

at time t. The canonical state space model for the AR model in (2.2) is given by [5, 30],

$$\mathbf{x}(t) = \mathbf{F}\mathbf{x}(t-1) + \mathbf{G}w(t) \tag{2.27}$$

where $\mathbf{x}(t)^T = [x(t-N_x+1)\ x(t-N_x+2)\ \ldots,\ x(t)]$ and $\mathbf{x}(t) = \mathbf{0}\ \ for\ \ t \leq 0$, the state-transition matrix $\mathbf{F}$ is given by

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ a_{N_x} & a_{N_x-1} & a_{N_x-2} & \ldots & a_2 & a_1 \end{bmatrix}$$

where the a_{N_x} is the N_xth order AR coefficient, and the process matrix $\mathbf{G}$ is given by,

$$\mathbf{G} = \begin{bmatrix} 0 & 0 & \ldots & 0 & 0 & 1 \end{bmatrix}^T$$

and the observation model for (2.1) is given by,

$$z(t) = \mathbf{H}^T\mathbf{x}(t) + v(t) \tag{2.28}$$

where the observation matrix is given by,

$$\mathbf{H} = \mathbf{G}$$

The noise sequences $\{w(t)\}$ and $\{v(t)\}$ are zero mean Gaussian white noise processes with variances $q_w = \sigma_w^2$ and $q_v = \sigma_v^2$ respectively and are uncorrelated. For all t and k, we can write,

$$\mathcal{E}\{w(t)\} = 0, \quad \mathcal{E}\{w(t)w(k)\} = q_w\delta_{t,k} \tag{2.29}$$

$$\mathcal{E}\{v(t)\} = 0, \quad \mathcal{E}\{v(t)v(k)\} = q_v\delta_{t,k} \tag{2.30}$$

$$\mathcal{E}\{w(t)v(k)\} = 0, \quad \mathcal{E}\{x(t)v(k)\} = 0. \tag{2.31}$$

If $\mathbf{x}(t)$ and $z(t)$ are assumed to be be jointly Gaussian the **Kalman filter** is an estimator which gives optimal estimate of the $\mathbf{x}(t)$ given the noisy data $\{z(t), z(t-1), \ldots\}$. For such a Gaussian distribution, the optimal estimate is the MMSE estimate given by

$$\hat{\mathbf{x}}(t|t) = \mathcal{E}[\mathbf{x}(t)|z(t), z(t-1), \ldots] \tag{2.32}$$

The corresponding state estimation error covariance $\mathbf{\Sigma}_{\mathbf{x}}(t|t)$ is then defined as,

$$\mathbf{\Sigma}_{\mathbf{x}}(t|t) = \mathcal{E}\{\epsilon_{\mathbf{x}}(t|t)\epsilon_{\mathbf{x}}(t|t)^T\} \tag{2.33}$$

where $\epsilon_{\mathbf{x}}(t|t) = \mathbf{x}(t|t) - \hat{\mathbf{x}}(t|t)^T$ is the state estimation error. Similarly, the one step predicted error of $\mathbf{x}(t|t)$ is $\epsilon_{\mathbf{x}}(t|t-1) = \mathbf{x}(t|t) - \hat{\mathbf{x}}(t|t-1)$ and associated error covariance matrix $\mathbf{\Sigma}_{\mathbf{x}}(t|t-1)$ is defined as

$$\mathbf{\Sigma}_{\mathbf{x}}(t|t-1) = \mathcal{E}\{\epsilon_{\mathbf{x}}(t|t-1)\epsilon_{\mathbf{x}}(t|t-1)^T\} \tag{2.34}$$

In solving the Kalman filtering problem, we shall use the innovation approach that takes advantage of a special stochastic process called the innovation process [52, 45] which we shall introduce in the following section.

2.5.1 The Innovations Process

Let $\hat{z}(t|t-1)$ be the MMSE estimate of the observation $z(t)$ at time n given all the past observations upto time $n-1$, that is, given $z(1), z(2), \ldots, z(t-1)$. $\hat{z}(t|t-1)$ is also known as one-step prediction of $z(t)$. We can now define the forward prediction error as,

$$\epsilon_z(t|t-1) = z(t) - \hat{z}(t|t-1) \tag{2.35}$$

According to the principle of orthogonality[45], $\epsilon_z(t|t-1)$ is orthogonal to all past measurements, i.e. to {z(1),...,z(t-1)}. In (2.35) we see that the new information about the measurement $z(t)$ at time t is accommodated in the forward prediction error $\epsilon_z(t|t-1)$. Hence, the name *innovation* for $\eta(t) = \epsilon_z(t|t-1)$. The innovation process is a stochastic process that has the following properties [45]:

- *The innovation $\eta(t)$ associated with the observation $z(t)$ at time n is orthogonal to all past observations*,

$$\mathcal{E}[\eta(t)z(k)] = 0, \quad 1 \leq k \leq t-1 \tag{2.36}$$

- *The innovation process consists of orthogonal random variables*,

$$\mathcal{E}[\eta(t)\eta(k)] = 0, \quad 1 \leq k \leq t-1 \tag{2.37}$$

- *There is one-to-one correspondence between the observations* $\{z(1), z(2), \ldots, z(t-1)\}$ *and the innovation process* $\{\eta(1), \eta(2), \ldots, \eta(t-1)\}$.

2.5.2 State Variable Estimation

According to the measurement model (2.28), there is a linear relationship between the state vector $\mathbf{x(t)}$ and observation $z(t)$. Since there is one to one correspondence between the observations and the innovation process as stated in the previous section, $\mathbf{x(t)}$ must be linearly related to the innovation $\eta(t)$ associated with the observation $z(t)$. Again for a Gaussian time-varying process, the optimal MMSE estimator is linear [53]. Thus, we can express $\hat{\mathbf{x}}(t|t)$, the MMSE estimate, of the state vector $\mathbf{x}(t)$, as linear combinations of the innovation sequence, that is,

$$\hat{\mathbf{x}}(t|t) = \sum_{k=1}^{t} \mathbf{B}_t(k)\eta(k) \tag{2.38}$$

where $\{\mathbf{B}_t(k)\}$ is a set of N_x-dimensional vectors to be determined. According to the principle of orthogonality[45], in order for the cost function to attain its minimum value in the mean square sense, the state estimation error $\epsilon_{\mathbf{x}}(t|t)$ and the observation $z(t)$ are orthogonal. Thus, according to the properties of the innovations, the state estimation error must also be orthogonal to the innovation $\eta(t)$, that is,

$$\begin{aligned} \mathcal{E}\{\epsilon_{\mathbf{x}}(t|t)\eta(k)\} &= \mathcal{E}\{\mathbf{x}(t) - \hat{\mathbf{x}}(t|t))\}\eta(k) \\ &= \mathbf{0} \qquad k = 1, 2, \ldots, t \end{aligned} \tag{2.39}$$

Using (2.38) and (2.37), we rewrite (2.39) as,

$$\mathbf{B}_t(k) = \mathcal{E}\{\mathbf{x}(t-1)\eta(k)\}r_\eta^{-1} \tag{2.40}$$

where $r_\eta = \mathcal{E}\{\eta(k)^2\}$ is the zero-lag autocorrelation for $\eta(k)$. Substituting (2.40) and (2.27) in (2.38) and making use of the fact that $\mathcal{E}\{w(k)\eta(k)\} = 0$ for $0 \leq k \leq t$, we express the MMSE estimate for $\mathbf{x}(t|t)$ as,

$$\hat{\mathbf{x}}(t|t) = \mathbf{F}\hat{\mathbf{x}}(t|t-1) + \mathcal{E}\{\mathbf{x}(t-1)\eta(t)\}r_\eta^{-1}\eta(t) \tag{2.41}$$

Defining,

$$\kappa = \mathcal{E}\{\mathbf{x}(t-1)\eta(t)\}r_\eta^{-1} \tag{2.42}$$

and since according to (2.35) and (2.28), the innovation $\eta(t)$ and $z(t)$ are related by,

$$\eta(t) = z(t) - \mathbf{HF}\hat{\mathbf{x}}(t-1|t-1) \tag{2.43}$$

we define the state vector estimate as,

$$\hat{\mathbf{x}}(t|t) = \mathbf{F}\hat{\mathbf{x}}(t-1|t-1) - \kappa(t)[z(t) - \mathbf{H}^T\mathbf{F}\hat{\mathbf{x}}(t-1|t-1)] \tag{2.44}$$

$$= \mathbf{F}\hat{\mathbf{x}}(t-1|t-1) - \kappa\eta(t) \tag{2.45}$$

From (2.44) we observe that the MMSE estimate of the state of a linear dynamical system can be estimated by adding a correction term $\{\kappa\eta(t)\}$ to the product of the

previous state estimate $\hat{\mathbf{x}}(t-1|t-1)$ and the state transition matrix $\mathbf{F}$. Thus κ is refered to as *Kalman gain.*

2.5.3 Kalman Gain

In this section, we express the Kalman gain κ in a convenient form for computation[45, 50, 53]. We rewrite the expression for the Kalman gain, by substituting for $\mathbf{x}(t-1)$ and the innovation $\eta(t) = \mathbf{H}\epsilon_x(t|t-1) + v(t)$, as,

$$\kappa = \mathcal{E}\{\mathbf{x}(t)\epsilon_x(t|t-1)\}\mathbf{H}r_\eta^{-1} \tag{2.46}$$

As $\epsilon_x(t|t-1)$ and $\mathbf{x}(t-1)$ are orthogonal,

$$\begin{aligned} \kappa &= \mathcal{E}\{\epsilon_x(t|t-1)^T\epsilon_x(t|t-1)\}\mathbf{H}r_\eta^{-1} && (2.47)\\ &= \mathbf{\Sigma_x}(t|t-1)r_\eta^{-1} && (2.48)\\ &= \mathbf{\Sigma_x}(t|t-1)\mathbf{H}[\mathbf{H}^T\mathbf{\Sigma_x}(t|t-1)\mathbf{H} + q_v]^{-1} && (2.49) \end{aligned}$$

where the one step state prediction error covariance $\mathbf{\Sigma_x}(t|t-1)$ is given by,

$$\mathbf{\Sigma_x}(t|t-1) = \mathbf{F}\mathbf{\Sigma_x}(t-1|t-1)\mathbf{F}^T + \mathbf{G}q_w\mathbf{G}^T \tag{2.50}$$

and state estimation error covariance $\mathbf{\Sigma_x}(t|t)$ is given by,

$$\mathbf{\Sigma_x}(t|t) = [\mathbf{I} - \kappa(t)\mathbf{H}^T]\mathbf{\Sigma}(t|t-1) \tag{2.51}$$

2.5.4 Summary of Kalman Filter Algorithms

In this section we summarize Kalman Filter Algorithms derived in the previous subsections[45, 51, 53].

Priors:

$$\hat{\mathbf{x}}(0|0) = \mathbf{0} \tag{2.52}$$

$$\mathbf{\Sigma_x}(0|0) = [0]_{N_x \times N_x} \tag{2.53}$$

Prediction Steps:

$$\hat{\mathbf{x}}(t|t-1) = \mathbf{F}\hat{\mathbf{x}}(t-1|t-1) \tag{2.54}$$

$$\mathbf{\Sigma_x}(t|t-1) = \mathbf{F}\mathbf{\Sigma_x}(t-1|t-1)\mathbf{F}^T + \mathbf{G}q_w\mathbf{G}^T \tag{2.55}$$

Update Steps:

$$\kappa(t) = \mathbf{F}\mathbf{\Sigma_x}(t|t-1)\mathbf{H}[q_v + \mathbf{H}^T\mathbf{\Sigma_x}(t|t-1)\mathbf{H}]^{-1} \tag{2.56}$$

$$\hat{\mathbf{x}}(t|t) = \hat{\mathbf{x}}(t|t-1) + \kappa(t)[z(t) - \mathbf{H}^T\hat{\mathbf{x}}(t|t-1)] \tag{2.57}$$

$$\mathbf{\Sigma}(t|t) = [\mathbf{I} - \kappa(t)\mathbf{H}^T]\mathbf{\Sigma_x}(t|t-1) \tag{2.58}$$

The speech sample estimate $\hat{x}$ at time t is finally found by,

$$\hat{x}(t) = \mathbf{H}^T\hat{\mathbf{x}}(t|t) \tag{2.59}$$

The simplicity of the Kalman filter algorithm makes it an attractive candidate over a more complex HMM based system. The problem with this sort of implementation of the Kalman filter is that we are using the AR model in (2.2) for modeling speech signals. This model is not a good model for representing all speech types. Thus, AR parameters estimated with this model affect the enhancement capability of Kalman filter. Gibson et al. [5] have presented a Kalman filter formulation for colored noise. It was found that the colored noise formulation gave only minor improvement at the cost of increased implementation complexity. Hence we have restricted our direction of research towards using the additive white measurement noise formulation only.

2.6 Speech Production System

In order to have a good model for representing the speech signal, we need to have a good understanding of the process of speech production. In the following subsection, we present a concise description of the anatomy and physiology of speech production [54, 43, 55].

2.6.1 Anatomy and Physiology of the Human Speech Production

The speech production apparatus is comprised of three major anatomical subsystems [54]: the respiratory, the laryngeal and the articulatory subsystem. Figure 2.2 depicts the speech production system. The respiratory subsystem is composed of the lungs, trachea or windpipe, diaphragm and the chest cavity. The Larynx and pharyngeal cavity or throat constitutes the laryngeal subsystems. The articulatory subsystem includes the oral cavity and the nasal cavity. The oral cavity is comprised of velum, tongue, lips, jaw and teeth. In speech processing technical discussions, the vocal tract is referred to the combination of the larynx, the pharyngeal cavity and the oral cavity. The nasal tract begins at the velum and terminates at the nostrils.

The respiratory subsystem behaves like an air pump, supplying the aerodynamic energy for the other two subsystems. In speech processing, the basic aerodynamic parameters are air volume, flow, pressure and resistance [54]. The main contribution of the respiratory subsystem for speech production is that when a speaker inhales air by muscular adjustments causing an increase in volume of the respiratory system, then the lungs release air by a combination of passive recoil and

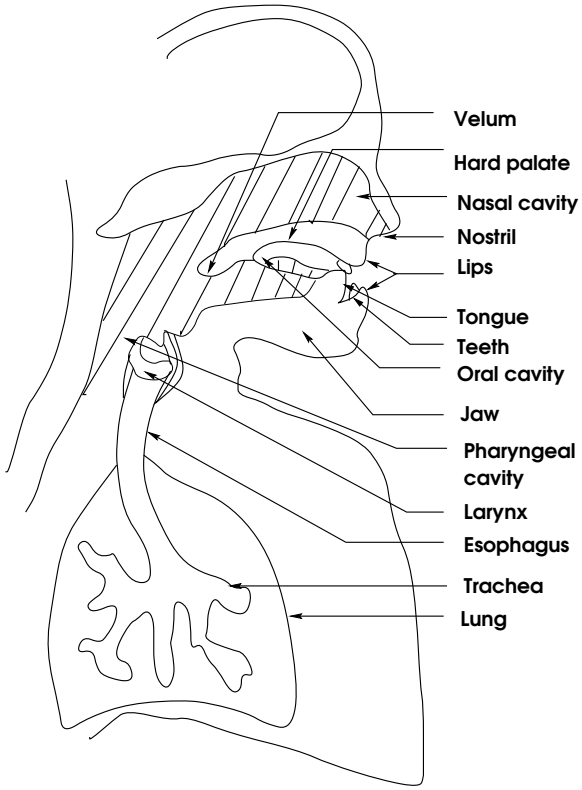

Figure 2.2: An illustration of the human speech production system after [43]

muscular adjustments. Air release depends on the volume of air in the lungs and aerodynamic requirements. The laryngeal subsystem acts as a passage for air flow from the respiratory subsystem to the articulatory subsystem. In the laryngeal subsystem, the larynx consists of various cartilages and muscles. For speech production, of particular importance are a pair of flexible bands of muscle and mucus membrane called vocal folds, stretching from the thyroid cartilage in the front to the arytenoids cartilages at the rear. The vocal folds vibrate to lend a periodic excitation for production of certain speech types which we will discuss in Chapter 3. The vocal folds come together or separate to respectively close or open the laryngeal airway. The opening between the vocal folds is known as the glottis.

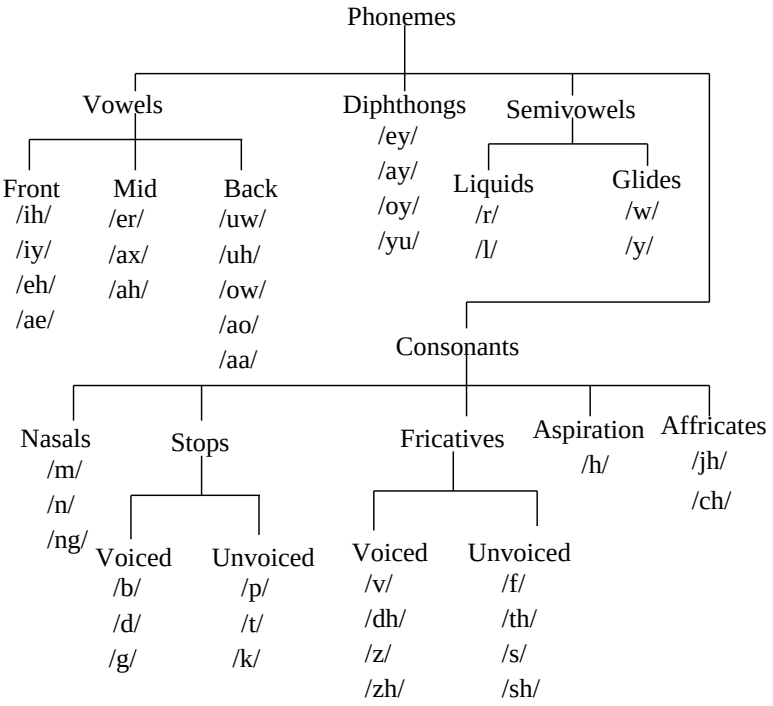

Figure 2.3: Classification of Phonemes in American English

The articulatory subsystem stretches from the top of the larynx up to the lips and nose through which the acoustic energy can escape. The articulators are movable structures that shape the vocal tract, determining its resonant properties. This subsystem also provides an obstruction for some cases or generates noise for certain speech types.

2.7 Phonemes and Phones

Let us first define some of the very basic speech representing units. A sequence of various sound units constitute a speech signal. These sound units are manipulated by the language rules known as linguistics[43]. The sound units, used as basic theo-

retical units for expressing linguistic meaning are called phonemes. Each phoneme has a unique set of articulatory gestures. These articulatory gestures specify the type and location of speech excitation and the position or movement of the vocal tract articulators. In American English, there are 42 phonemes [43], [55]. These phonemes are divided into four broad classes: Vowels, Diphthongs, Semivowels and Consonants as shown in Table 2.3. Consonants include five classes of phonemes: Nasals, Stops, Fricatives, Affricates and Aspiration. A phoneme is considered a continuant if it is produced by a steady-state vocal tract configuration excited by an appropriate source. Vowels, Fricatives, Affricates and Aspiration are continuant phoneme classes. The remaining phoneme classes are produced by varying vocal tract configuration.

As the definition of a phoneme goes, it can be considered as an ideal unit of sound with a set corresponding articulatory gestures. In reality accents, gender, coarticulatory effects etc. all give rise to variability of the same phoneme. Thus, from an acoustical point of view, a phoneme basically represents a class of sounds with similar meaning. The actual sounds units, generated while speaking, are referred to as phones in speech literature [43].

2.8 TIMIT Database

Since we shall be using speech data from the TIMIT database, we present a brief overview of this database in this section. TIMIT is an acoustic-phonetic speech corpus designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of speech processing systems[56]. It is prepared by the National Institute of Standards and Technology (NIST) with sponsorship from the Defense Advanced Research Projects Agency - Information

Science and Technology Office (DARPA-ISTO). TIMIT consists of a total of 6300 sentences, 10 sentences spoken by each of 630 male and female speakers from 8 major dialect regions of the United States. The speech data in TIMIT is divided into two broad groups: train and test for training and testing purposes. Each group is further subdivided into eight dialect groups. There are four files associated with each sentence data: a wave file (.wav), a text file (.txt), a word file (.wrd) and a phone file (.phn). The wave file consists of waveform speech data with a header. The speech waveforms are digitized at the sampling rate of 16 kHz and are stored in binary format. The text file contains the associated orthographic transcriptions of the words in a sentence. The word file is composed of the time-aligned word transcriptions while the phone file consists of the time-aligned phonetic transcription. A more detailed description of the TIMIT phonetic lexicon can be found in [56]. Finally, in the table 2.1 and 2.2 we present the TIMIT phonetic transcription to be used consistently in this thesis.

Phone type	Symbol	Example word	Phonetic transcription
Stops	b	bee	BCL B iy
	d	day	DCL D ey
	g	gay	GCL G ey
	p	pea	PCL P iy
	t	tea	TCL T iy
	k	key	KCL K iy
	dx	muddy, dirty	m ah DX iy, dcl d er DX iy
Affricates	jh	joke	DCL JH ow kcl k
	ch	choke	TCL CH ow kcl k
Fricatives	s	sea	S iy
	sh	she	SH iy
	z	zone	Z ow n
	zh	azure	ae ZH er
	f	fin	F ih n
	th	thin	TH ih n
	v	van	V ae n
	dh	then	DH e n
Nasals	m	mom	M aa M
	n	noon	N uw N
	ng	sing	s ih NG
	em	bottom	b aa tcl t EM
	en	button	b ah q EN
	eng	washington	w aa sh ENG tcl t ax n
	nx	winner	w ih NX axr

Table 2.1: Phonetic transcription used in the TIMIT database for Stops, Affricates, Fricatives and Nasals.

Semivowels	l	lay	L ey
	el	bottle	bcl b aa tcl t EL
	r	ray	R ey
	w	way	W ey
	y	yacht	Y aa tcl t
Aspiration	hh	hay	HH ey
	hv	ahead	ax HV eh dcl d
Vowels	iy	beet	bcl b IY tcl t
	ih	bit	bcl b IH tcl t
	eh	bet	bcl b EH tcl t
	ey	bait	bcl b EY tcl t
	ae	bat	bcl b AE tcl t
	aa	bott	bcl b AA tcl t
	aw	bout	bcl b AW tcl t
	ay	bite	bcl b AY tcl t
	ah	but	bcl b AH tcl t
	ao	bought	bcl b AO tcl t
	oy	boy	bcl b OY
	ow	boat	bcl b OW tcl t
	uh	book	bcl b UH kcl k
	uw	boot	bcl b UW tcl t
	ux	toot	tcl t UX tcl t
	er	bird	bcl b ER dcl d
	ax	about	AX bcl b aw tcl t
	ix	debit	dcl d eh bcl b IX tcl t
	axr	butter	bcl b ah dx AXR
	ax-h	suspect	s AX-H s pcl p eh kcl k tcl t

Table 2.2: Phonetic transcription used in the TIMIT database for Semivowels, Aspiration and Vowels.

Chapter 3

Impulsive AR Models for Speech Enhancement

This chapter introduces and implements AR models with impulsive excitation models for various speech types. Section 3.2 studies and compares the performance of the state-of-the-art HMM model based and AR enhancement systems. Section 3.2 also motivates us to review the production mechanisms for various phoneme classes in Section 3.3. Models for each phoneme class are proposed in Section 3.4. Sections 3.5 and 3.6 are dedicated to the Kalman filter algorithms and AR parameter estimation techniques for the proposed impulse driven AR models. The assertions and the assumptions made by impulsive models are explained in detail in Section 3.7. Experimental results for various phoneme classes are discussed in Section 3.8. Finally, the tables for enhancement results are presented in Section 3.10.

3.1 Introduction

One of our objectives is to determine to what extent we can produce a "good" model for representing a speech signal. But "good" being extremely subjective and qualitative term, necessitates the need for setting some useful criteria for judging a model. For our speech enhancement research we shall be using model residual plots and output signal-to-noise ratios (SNRs) as major deciding factors for judging a model. The quest, for a good model, has motivated us in Section 3.2 to have a careful look at the speech models used by two very popular state of the art enhancement systems: HMM based Wiener filter and AR model based Kalman filter. Because of the simplicity (explained in detail in Section 2.2) of a white noise driven AR model for speech we have chosen to focus our attention towards an AR model used by the Kalman filter. According to linear speech production theory[3], speech signal or pressure wave, measured at a microphone, is produced by the combined effects of the voice source excitation, vocal tract articulation and radiation from the lips or nostrils. An AR model driven by white noise, used for speech enhancement, fails to capture the effects of the excitation source and radiation, especially in the case of the voiced speech. This has motivated us to include a relevant driving term in the conventional AR model.

We intend to develop models for representing each phoneme class by taking into account the production mechanism of that class. We discuss the production mechanism of different phonemes classes in Section 3.3 while in Section 3.4 we introduce the developed models. Because of inclusion of impulses in the speech model we shall not be able to use conventional AR parameter estimation procedures. Section 3.6 explains impulse synchronous AR parameter estimation procedure. We shall be testing the developed models with the performance of the Kalman filter.

In Section 3.5 we briefly mention the Kalman filter algorithms. As one of our objectives is to establish limits to performance for the Kalman filter, we shall be making a number of assertions and assumptions for our proposed models which are discussed in Section 3.7. In Section 3.8 we present and discuss the results. We shall be using output SNRs as objective measures of enhancement while for subjective measures we shall be observing the plots of the time waveforms and residuals of the AR estimation process.

3.2 Review of the State of the Art Enhancement Systems

The main reason for reviewing the two very popular state of the art speech enhancement systems is to present motivations that led into applying impulsive models for various speech types. In this section, we review the results obtained by using an HMM based system[24, 8] and an AR model based Kalman filter[30, 5] (discussed in detail in Sections 2.4 and 2.5 respectively). The speech data is taken from the TIMIT database. A global measure of SNR is used as objective evaluation criterion. The output SNR was calculated by,

$$\text{SNR} = 10 \log \frac{\sum_{t=1}^{J} x^2(t)}{\sum_{t=1}^{J} [x(t) - \hat{x}(t)]^2} \tag{3.1}$$

where J is the length of the speech signal. Subjective evaluation of the results is based on human hearing perception and inspection of spectrograms. We have first listened to the clean speech and noisy speech then followed by the enhanced speech. The Quality of Perception(QOP) was divided into four categories on the scale of score=5, namely- excellent (score=5), good (score=4), fair (score=3) and

poor (below 3) [8]. We have also inspected spectrograms of the clean speech, noisy speech and enhanced speech.

3.2.1 HMM Based Enhancement System Overview

The HMM based enhancement system was used in enhancing speech signals which have been degraded by white noise at signal-to-noise ratio (SNR) values of 5, 10, 15 dB. The speech data used was selected from the sentences in the TIMIT database. One hundred sentences, spoken by 15 different speakers with a sampling frequency of 16kHz, were used for training the clean speech model. Four sentences spoken by 4 different speakers were used for enhancement purpose. The speech material and the speakers used for training were different from those used for testing. Training was done using non-overlapped frames while enhancement was done using 50% overlapping of adjacent frames. The clean signal was modeled with a 5 state 5 mixture HMM while each noise type was modeled with a 3 state 3 mixture HMM.

A block diagram of the implemented system is shown in Figure 3.1. First autocorrelation coefficients, of each frame of the noisy signal, are extracted. These coefficients are then fed into the noise adaptation model. The non-speech intervals of the noisy speech are detected by this model and a Viterbi forward algorithm is performed on noise data using three different types of noise models. Then the likelihood for each noise model is calculated and the model associated with the highest likelihood is selected. Using the selected noise HMM parameters and the clean speech model, the preprocessed noisy speech is fed into the MMSE forward algorithm which generates the weights for the Wiener filters [8]. In the mean time, all Wiener filters for each combination of the state and mixtures pairs in the speech and noise models are calculated. A single weighted filter is constructed for

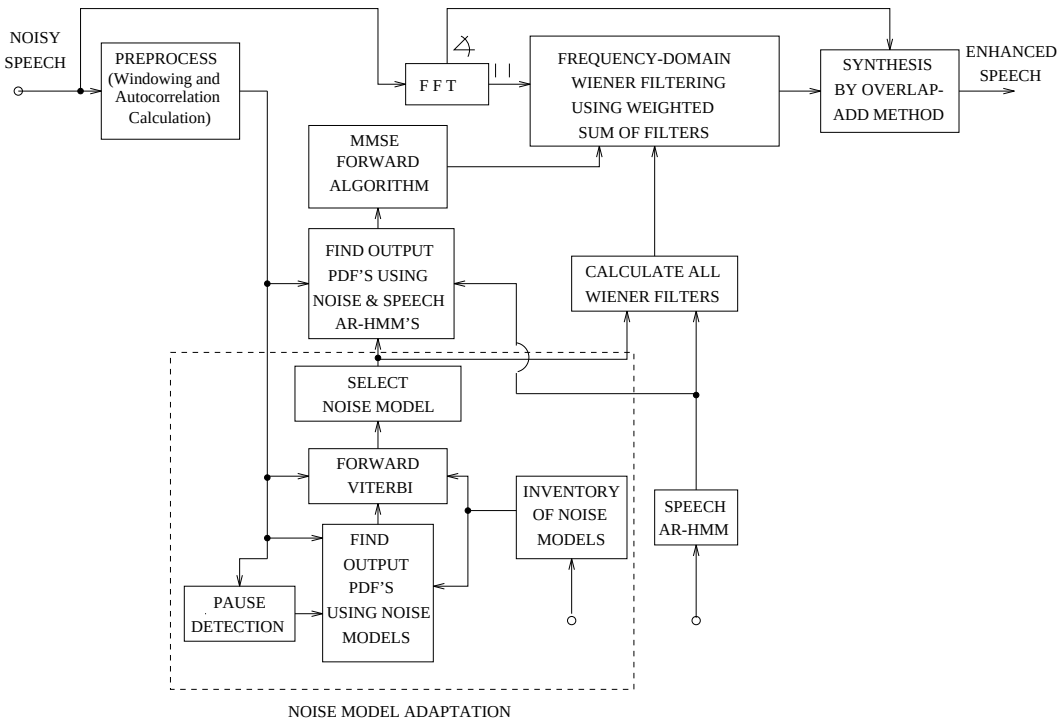

Figure 3.1: HMM based enhancement system, after [8].

each frame of noisy speech using the calculated filter weights and the pre-trained Wiener filters. The filtering of the noisy signal is carried out using the weighted filter. The output is the spectral magnitude of the enhanced speech signal. Using this magnitude together with the noisy speech's phase information, an inverse FFT is performed to obtain the time-domain enhanced speech.

3.2.2 Overview of an AR Model based Kalman Filtering

Kalman filter algorithm, given by (2.54) to (2.59), is used to estimate the clean speech signal from the noisy speech for each frame length of 256 data points. Kalman filtering algorithms require the knowledge of AR coefficients, σ_w^2 and σ_v^2. The AR coefficients of the noisy speech are computed using the covariance method [16, 17]. The residual white noise component $w(i)$ is calculated using (2.2). The variance σ_w^2 of this residual time series is then computed. We have used (3.2) to compute σ_v^2.

For the first frame, the state vector is initialized as $\hat{\mathbf{x}}(0|0) = \mathbf{0}$ and the corresponding error covariance is initialized as $\Sigma_{\mathbf{x}}(0|0) = [0]_{K\times K}$. Then the one-step state prediction estimate and the corresponding error covariance given by (2.54) and (2.55) are estimated respectively. This is followed by update steps through evaluations of the Kalman gain (given by (2.56)), state estimation (2.57) and the corresponding state estimation error covariances (2.58). The speech sample estimate is found by using (2.59). For the following frames the state vector and the corresponding error covariance are initialized using their last values from the previous frame.

Input SNR in dB	Output SNR in dB
5	10.969
10	12.097
15	15.515

Table 3.1: Enhancement results averaged over four sentences for HMM based Wiener filter.

Input SNR	lpc Order		
in dB	8	10	12
5	11.109 dB	11.140 dB	11.183 dB
10	14.542 dB	14.578 dB	14.625 dB
15	18.306 dB	18.343 dB	18.385 dB

Table 3.2: Results averaged over four sentences for AR model based Kalman filter.

3.2.3 Enhancement Results

Four sentences from TIMIT database were corrupted with simulated white noise at the input SNRs of 5 dB, 10 dB and 15 dB. These noisy utterances were enhanced using HMM based and AR model based systems described in the subsections 3.2.1 and 3.2.2. Enhancement results, for HMM based system, averaged over the four sentences are shown in Table 3.1. The QOP was judged by three listeners including the author. Overall QOP was found to be equal to the scale of 3 or fair. There were some interruptions or discontinuities present in the enhanced speech signal. These dropouts were due to filtering of the speech data especially fricatives, stops or affricates which were mistaken as noise by the model.

Averaged results for four test utterances for the AR model based Kalman filter are shown in the Table 3.2 for various AR orders and input SNRs. Increasing AR order provides better modeling of speech signals to some extent. But after a certain AR order the output SNR does not change much which indicates AR models' limitations on modeling the speech signals. The output SNR values indicate considerable amount of improvement over that for HMM. The overall QOP for the enhanced speech was found to be equal to the scale of 3.8-4 which was higher than that for HMM. The enhanced speech signals were more pleasant sounding compared to that for HHMs. There was slight noise present in enhanced speech.

Figures 3.2 and 3.3 respectively show the spectrograms for a part of the clean speech signal, "It was exposed to a high velocity gas jet" and the noisy speech created by corrupting the clean with white noise at the input SNR of 5 dB. In the spectrograms, each small square in the horizontal direction corresponds to the 0.1 sec of time while that in the vertical direction corresponds to frequency of 1.0 kHz. Enhanced speech, from HMM based and AR model based systems, are shown

respectively in Figures 3.4 and 3.5. For HMM based enhancement the dropouts are also quite evident in the spectrogram for enhanced speech in Figure 3.4. In the spectral region upto 0.53 sec, the speech signal corresponding to "It was ex" has been almost wiped out. The spectrum, around $.44 - .53$ sec, corresponding to the stop /k/ (x from "exposed) is barely visible. The fricatives /z/ (i.e. s from "exposed") and /s/ (i.e. c from "velocity") corresponding respectively to the intervals between $0.93 - 1.0$ sec and between $1.72 - 1.86$ sec have been filtered out as these two phonemes were mistaken as noise by HMM.

We observe that Figure 3.5, for AR model based enhancement, in general more resembles the spectrogram for the clean signal in Figure 3.2 compared to that for HMM in Figure 3.4. The spectra, upto 0.53 sec corresponding to "It was ex", has been better preserved than for Figure 3.4. In Figure 3.5, we observe that unlike Figure 3.4, spectra corresponding to the fricatives (/z/ and /s/ around $0.93 - 1.0$ and $1.72 - 1.86$ respectively) and the stops (/k/ and /d/ around $.44 - .53$ and $1.00 - 1.03$ sec respectively) have been well preserved but not exactly same as that for the clean speech in Figure 3.2. In Figure 3.5 we notice some noise present in the high frequency region above 4 kHz. For both Figures 3.4 and 3.5 we notice that periodicities for the voiced speech is not as evident as in the case of the clean speech in Figure 3.2. This may be due to the presence of some observation noise and model residual noise.

As mentioned in the previous chapter, there are two major limitations associated with HMMs for speech enhancement [8]. For accommodating rapidly varying speech signals, it requires a large number of states. This increases the computational complexity (e.g., $N_S x \times N_m x \times N_S v \times N_m v$ in (2.26) for each frame) in the model together with the risk of affecting the performance of HMM for slowly varying speech signals. The second problem with HMMs is that even with the higher

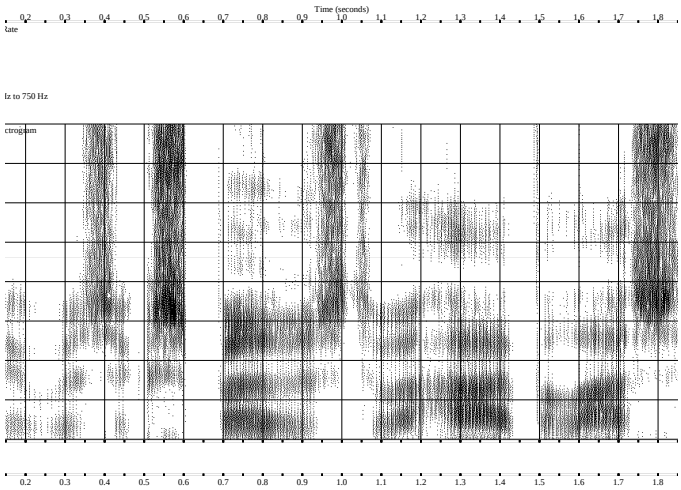

Figure 3.2: Spectrogram of the part of the original clean test speech signal "It was exposed to a high velocity gas jet".

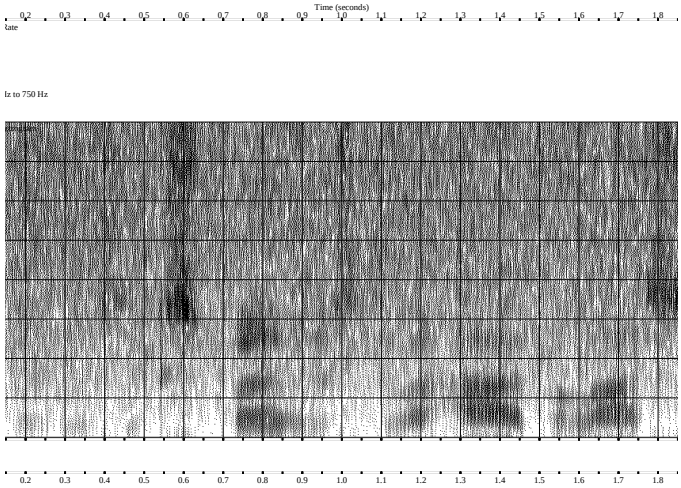

Figure 3.3: Spectrogram of the noisy test utterance, corrupted with white noise at SNR of 5 dB.

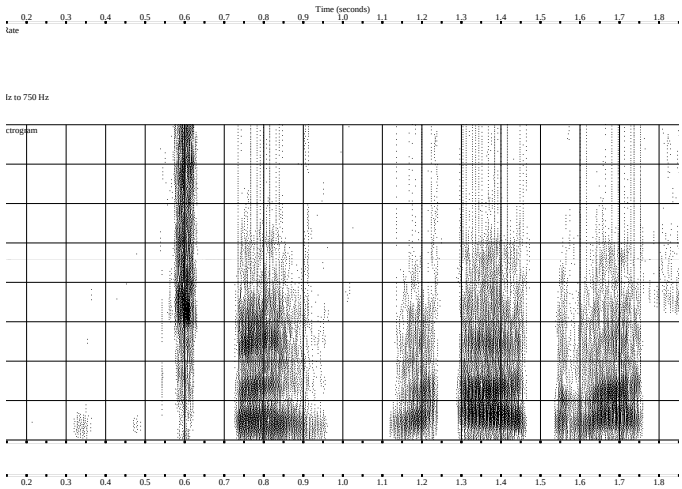

Figure 3.4: Spectrogram of the enhanced speech using HMM based enhancement system.

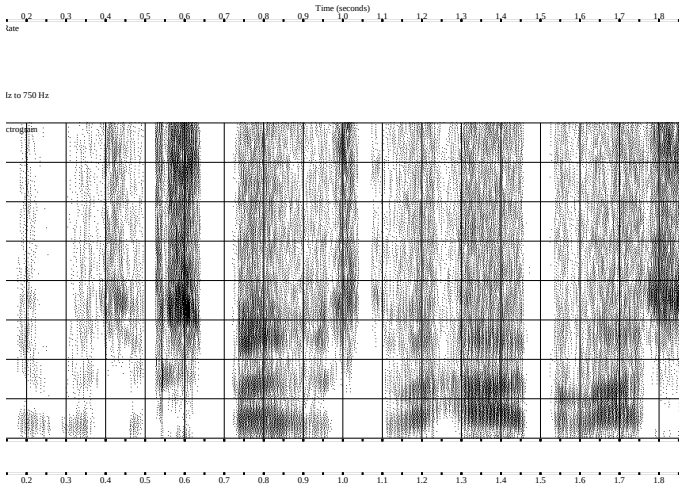

Figure 3.5: Spectrogram of the enhanced speech using Kalman filter based enhancement system.

number of states, the continuity of the speech signal is greatly affected. This is due to fact that discrete number of states and mixtures are used to represent the speech features. Whenever there is a big jump from one state-mixture pair to another, the continuity of the speech signal deteriorates. The flaws in AR model in (2.2) become apparent when the model residuals,

$$x(t) - \sum_{i=1}^{N_x} a_i x(t-i) \tag{3.2}$$

are examined, as shown in Figure 3.6. The model in 2.2 asserts that these residuals should be white (random), an assertion which is flatly contradicted by the figure, since obvious quasi-periodic (deterministic) components are present in each of the four phones shown. The remainder of this thesis investigates more consistent alternatives to $w(t)$ in (2.2).

3.3 Speech Sound Types

Prior to developing new models for various speech classes we shall review various phoneme classes [55, 54, 43] that have been characterized based on the positions and movement of speech articulators, type of excitation, transient properties of their time waveforms and frequency domain properties. For phonemic or phonetic transcription we shall be using the same convention that used in the TIMIT lexicon.

3.3.1 Vowels

Vowels are produced by exciting a steady-state vocal tract configuration with quasi-periodic pulses of air [55]. Quasi-periodic pulses are produced when air from the

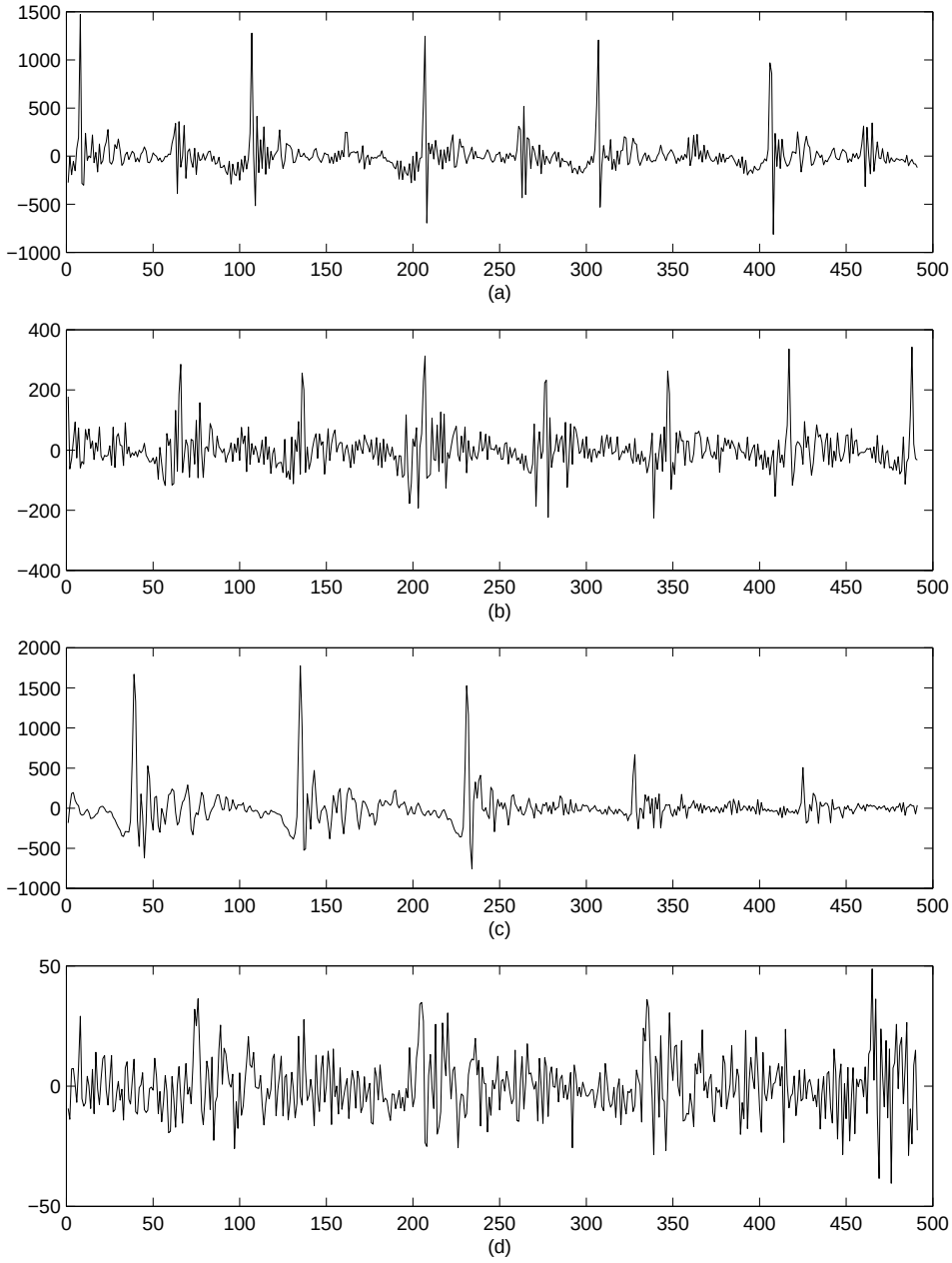

Figure 3.6: Plots of AR residuals for four voiced speech phones: (a) front vowel /ae/, (b) diphthong /ay/, (c) semivowel /r/, (d) nasal /n/. The model (2.2) predicts that each of these signals be white (random) — clearly incorrect.

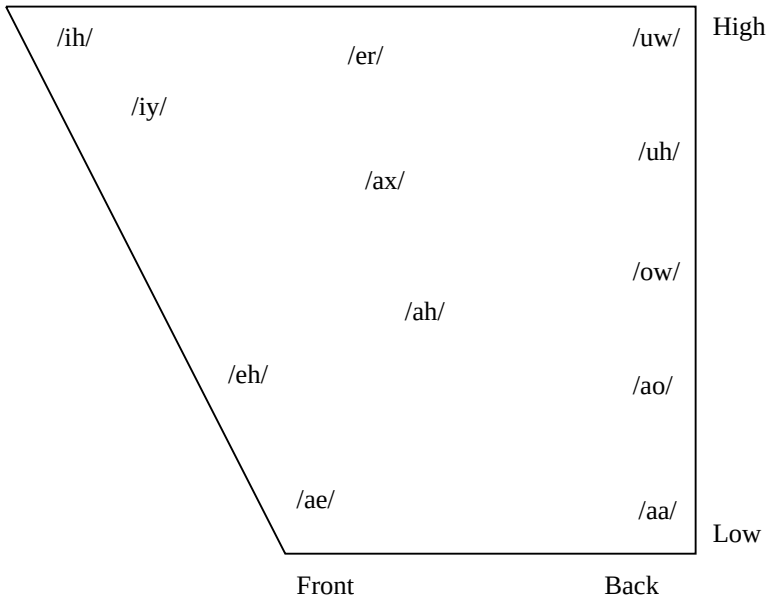

Figure 3.7: Position of the tongue in the oral cavity during the production of the vowels, after [43].

respiratory subsystem is forced through the glottis, the tension of the vocal cords are adjusted so that they begin to vibrate and cause periodic interruption of the subglottal airflow. The variation in the cross-sectional area along the vocal tract determines the resonant frequencies known as formants of different vowels. Three factors that influence formant frequency locations for vowels are: the overall length of the pharyngeal-oral tract, the location of the tract and the narrowness of the constrictions [43]. The first three formant frequencies can be used to roughly characterize vowels. The term *roughly* is applied here as some variability is to be expected among the speakers producing the same vowels. Other factors used for acoustically identifying the vowels are spectra, durations and formant bandwidths. Vowels are classified into three groups by the position of the tongue and the degree of constriction at that position. Figure 3.7 shows Front, Central and Back vowels.

The vowels are differentiated within each group by the degree to which the tongue is raised towards the palate.

3.3.2 Diphthongs

A diphthong is a dynamic sound that starts at or near the articulatory position of one target vowel and moves to or towards the position for another target vowel [55, 54]. The first target vowel is usually longer than the the latter one but the transition between the targets is longer than either of the targets [43]. There exists some discrepancy about drawing sharp distinction between a diphthong and two adjacent distinct vowels. The four universally recognized diphthongs in North American English include /ey/ (as in "bay"), /ay/ (as in "buy"), /aw/ (as in "how") and /oy/ (as in "toy"). Even though a diphthong represents transition from one vowel to another, it is often the case that neither target vowel is actually reached.

3.3.3 Semivowels

Semivowels consisting of four phonemes /w/, /l/, /r/, /y/ are divided into two groups: glides (/w/ & /y/) and liquids (/l/ & /r/). Semivowels have glottal excitation that produces well defined formant structure like vowels but unlike vowels formant structure is gradually changing due to a constriction in the vocal tract. The degree of constriction is smaller than that in vowels but large enough not to produce any turbulence. A glide is defined as a vocalic syllable nucleus consisting of one target position with corresponding formant transitions toward and away from the target [43]. Liquids also have spectral characteristics similar to vowels but are usually weaker than most vowels due to their more constricted vocal tract. The

acoustic properties of semivowels are strongly affected by the context in which they occur.

3.3.4 Nasals

The nasal consonants /m/, /n/ and /ng/ are produced by the glottal excitation of an open nasal cavity and the oral cavity constricted at some point at the front. The velum is lowered to permit the sound propagation through the nasal cavity. The oral cavity, being acoustically coupled to the pharynx and the nasal cavity, serves a resonant cavity by capturing energy at certain natural frequencies. These resonant frequencies of the oral cavity emerge as anti-resonances or zeros of sound transmission [55]. Nasal formants and formants of the adjacent vowels have wider bandwidth or more highly damped compared to those of the vowels. This is caused by the fact that heat conduction and viscous losses are greater as inner surface of the nasal tract has large surface area.

The three nasals have three different areas of constriction along the oral cavity. For /m/ the constriction is at the lips (labial constriction), /n/ has constriction at the back of the teeth with the tongue resting at the gum ridge (also known as alveolar constriction) while for /ng/ the constriction is in the front of the velum (velar constriction).

3.3.5 Fricatives

Fricatives are characterized by the formation of narrow constriction at some location in the vocal tract, by the development of turbulent air stream and by the generation of noise. Fricatives are divided into unvoiced fricatives and voiced fricatives based

on the mode of excitation of the vocal tract. The unvoiced fricatives include /f/, /th/, /s/ and /sh/ while /v/, /dh/, /z/ and /zh/ constitute the voiced fricatives.

Unvoiced fricatives are produced by exciting the vocal tract by a steady air flow that becomes turbulent in a region of constriction. The constriction divides the vocal tract into two cavities. The cavity preceding the constriction then becomes a noise source due to turbulence. The speech sounds are radiated from the front cavity whereas the back cavity traps energy as in the case of nasals introducing anti-resonances into the speech output [55]. The location of the constriction determines the uttered fricative. For /f/ the constriction is labiodental (upper jaw teeth on lower lip), /th/ has interdental (tongue behind front teeth) constriction, for /s/ it is alveolar and /sh/ has palatal (tongue resting on hard or soft palate) constriction.

Voiced fricatives have both turbulent noise source at the constriction and quasi periodic glottal excitation of the vocal tract. Because of these two types of excitations, their spectra may show both periodicity (to some extent) and frication. Voiced fricatives /v/, /dh/, /z/ and /zh/ are the counterparts of unvoiced fricatives /f/, /th/, /s/ and /sh/ respectively, as far as the location of the constriction is concerned.

3.3.6 Stops

Stops are, also as in the case of fricatives, classified into unvoiced and voiced stops. Stops are noncontinuant speech signals produced by the total closure of the vocal tract during which a pressure builds up and sudden release of this pressure. This type of excitation is also known as plosive. The closure can be referred to as bilabial (/p/ and /b/), alveolar (/t/ and /d/) and velar (/k/ and /g/).

Unvoiced stops /p/, /t/ and /k/ are produced by abrupt release of air pressure

that builds up during the vocal tract occlusion. The air release, marked by a short interval of frication, is followed by a steady air flow from the glottis known as aspiration. The frication and aspiration are together known as stop release. The interval preceding the stop release is known as stop gap or closure.

Voiced stops /b/, /d/ and /g/ not only have plosive excitation but also a glottal excitation that continues throughout the closure and release. During the closure some amount of low frequency energy may be radiated through the walls of the throat [43] as the vocal cords keep vibrating. This is indicated by a voice bar in the frequency region in the spectral analysis.

3.3.7 Affricates

Affricates are non-continuant sounds having a palatal place of articulation. Affricates /jh/ and /ch/ are produced by the transition from a stop to a fricative. As in stops, affricates are produced with the total closure of the vocal tract. Similar to fricatives, affricates have a period of frication. But the frication interval tends to be shorter than that for fricatives [54]. The unvoiced affricate /ch/ is produced by a transition from unvoiced stop /t/ to unvoiced fricative /sh/ while the voiced affricate /jh/ is created by a transition from voiced stop /d/ to voiced fricative /zh/.

3.4 Models for Phoneme Classes

While reviewing the Kalman filter based enhancement system we have seen that a single AR model has been used to represent the speech signal. As each phoneme class has a different production mechanism, it is more appropriate to use different

models for various phonemes instead of using a single model for a whole utterance that is composed of various phoneme classes. In this section, we present models for various phoneme classes considering the type of excitation each class is associated with and closely observing conventional white noise excited AR residual plots.

3.4.1 Model for Voiced Speech

According to the acoustic theory of speech production[57], speech involves a source function and a vocal-tract filtering process. The output of the filtering process is speech pressure signal which is related to the volume velocity at the lips through a radiation term [16, 58, 57]. In speech synthesis, the combined effects of the excitation source, vocal-tract filter and radiation is modeled by an AR process where AR coefficients account for the filtering action of the vocal tract, the radiation and the excitation. The obvious flaw with the conventional autoregressive model in (2.2), for speech enhancement, is that the vocal tract is modeled as being driven by white noise, whereas vowels, diphthongs, semivowels and nasals all have quasi-periodic glottal pulse excitation of the vocal tract. Quasi-periodic pulses are produced when air is forced through the glottis, causing the vocal cords to vibrate and periodically interrupt the subglottal airflow. It is befitting to introduce a forcing term that models the glottal excitation in AR model for voiced speech.

We can begin to account for a quasi-periodic glottal excitation by modifying the AR forcing function to obtain

$$x(t) = \sum_{i=1}^{N_x} a_i x(t-i) + \quad w(t) + a_{_{N_x+1}} u_{_I}(t) \tag{3.3}$$

where $a_{_{N_x+1}}$ is the amplitude of the driving term, and where $u_{_I}(t)$ is a train of

impulses:

$$u_I(t) = \sum_j \delta(t - t_j) \tag{3.4}$$

where the times t_j mark the times of the glottal pulses. The impulse train $u_I(t)$ in (3.3) is simulating the effects of the voiced excitation source. Such an impulsive source function is extremely simplified approximation of the complex source function involved in speech production[16, 57]. The main reason for using such a tentative model (impulsive source function) is that we intend to investigate the effect of inclusion of the appropriate forcing function in an AR model for speech enhancement. Impulsive models, if successful over the conventional white noise driven AR model, may be replaced by more sophisticated source models used in speech synthesis.

The state-space model for speech given by (2.27) needs to be modified for the deterministic driving term $u_I(t)$. The state space model for voiced speech represented by (3.3) is given by,

$$\mathbf{x}(t) = \mathbf{F}\mathbf{x}(t-1) + \mathbf{G}w(t) + \mathbf{D}_I u_I(t) \tag{3.5}$$

where the input distribution matrix $\mathbf{D}_I \in R^{N_x}$ is defined as,

$$\mathbf{D}_I = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 & a_{N_x+1} \end{bmatrix}^T.$$

Transition and process matrices denoted by $\mathbf{F}$ and $\mathbf{G}$ are the same as those defined for (2.27). Inclusion of the driving term not only changes the conventional AR parameter estimation procedures discussed in Section 2.2 but also the assertions made by the model. We shall be discussing such issues in Sections 3.6 and3.7 respectively. The observation model however remains the same as (2.28), i.e.

$$z(t) = \mathbf{H}^T\mathbf{x}(t) + v(t) \tag{3.6}$$

as we have the same assumptions for additive Gaussian white measurement noise $v(t)$.

3.4.2 Models for Fricatives

Unvoiced fricatives have turbulent noiselike excitation also known as unvoiced excitation due to airflow through a narrow constriction. We need a model for frication noise source for representing such phonemes. The white noise term $w(t)$ in the model given by the AR model in (2.2) is adequate for representing the effect of the frication source[16]. The process state space model for unvoiced fricatives is given by,

$$\mathbf{x}(t) = \mathbf{F}\mathbf{x}(t-1) + \mathbf{G}w(t) \tag{3.7}$$

The observation state space model is the same as in (3.6).

Voiced fricatives can be represented by (3.3) as they have both unvoiced and voiced excitation. The state space models are the same as given by (3.5) and (3.6).

3.4.3 Models for Stops and Affricates

The stops have the following acoustic sequence,

$$< closure >< burst >< frication >< aspiration(for\ unvoiced\ stops) > \tag{3.8}$$

The stops have plosive excitation which is caused by a buildup of air pressure behind a completely closed part of the vocal tract ensued by a sudden release of this air pressure. For the unvoiced stops, the stop release can be modeled by a white term $w(t)$ and onset of the burst after the stop closure can tentatively be modeled by an

impulsive driving term $u_{Istop}(t)$. The model for unvoiced stops is given by

$$x(t) = \sum_{i=1}^{N_x} a_i x(t-i) + \quad w(t) + a_{N_x+1} u_{Istop}(t) \tag{3.9}$$

where a_{N_x+1} is the amplitude of the driving term $u_{Istop}(t)$, and $u_{Istop}(t)$ is a single impulse marking the onset of the burst i.e.,

$$u_{Istop}(t) = \delta(t - t_j) \tag{3.10}$$

where t_j is the time at which the burst occurs. The state space model for (refI11s) is then given by,

$$\mathbf{x}(t) = \mathbf{F}\mathbf{x}(t-1) + \mathbf{G}w(t) + \mathbf{D}_{Istop} u_{Istop}(t) \tag{3.11}$$

where $\mathbf{D}_{Istop} = \mathbf{D}_I$ is the input distribution matrix for $u_{Istop}(t)$.

The voiced stops have both plosive and voiced excitations. The voiced excitation is tentatively modeled by $u_I(t)$ which is a train of impulses separated by the pitch periods while the plosive excitation can be modeled in the same way as the unvoiced stops by a white noise term $w(t)$ and a single impulse $u_{Ivstop}(t) = \delta(t - t_j)$ marking the onset of the burst at t_j. The model for the voiced stops is given by,

$$x(t) = \sum_{i=1}^{N_x} a_i x(t-i) + \quad w(t) + a_{N_x+1} u_I(t) + a_{N_x+2} u_{Ivstop}(t) \tag{3.12}$$

where a_{N_x+1} and a_{N+2} are respectively the amplitudes of the driving terms $u_I(t)$ and $u_{Ivstop}(t)$. The state space model for (3.12) is then given by,

$$\mathbf{x}(t) = \mathbf{F}\mathbf{x}(t-1) + \mathbf{G}w(t) + \mathbf{D}_I u_I(t) + \mathbf{D}_{Ivstop} u_{Ivstop}(t) \tag{3.13}$$

where $\mathbf{D}_{Ivstop} \in R^{N_x}$ is the input distribution matrix for $u_{Ivstop}(t)$ defined as,

$$\mathbf{D}_{Ivstop} = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 & a_{N_x+2} \end{bmatrix}^T .$$

As mentioned earlier Affricates are non-continuant sounds produced by a transition from a stop to a fricative. Hence affricate /jh/ is represented by the same model as for voiced stops while /ch/ shares the model for unvoiced stops.

3.5 Kalman Filter Algorithms for Impulsive State Space Models

We shall be using the Kalman filter algorithms for filtering additive white noise $v(t)$ from noisy speech $z(t)$. The measurement model is given by (3.6). We have defined separate models for various classes of phonemes. We shall use these process models for Kalman filter algorithms. The **Priors** for the Kalman filter remains the same as those for white-noise excitation AR model i.e.,

$$\hat{\mathbf{x}}(0|0) = \mathbf{0} \tag{3.14}$$

$$\mathbf{\Sigma_x}(0|0) = [0]_{N_x \times N_x} \tag{3.15}$$

In the **Prediction Steps**, due to the inclusion of the driving terms in the process models for the voiced speech speech sounds, the voiced fricatives and the stops the one-step state prediction $\hat{\mathbf{x}}(t|t-1)$ will have various forms according to the corresponding process model. For the model in (3.5) for the voiced speech (e.g. vowels, semivowels, diphthongs,nasals) and voiced fricatives, the one step state prediction is given by,

$$\hat{\mathbf{x}}(t|t-1) = \mathbf{F}\hat{\mathbf{x}}(t-1|t-1) + \mathbf{D}u_{_I}(t) \tag{3.16}$$

For the unvoiced stop model in (3.11), the state prediction is given by,

$$\hat{\mathbf{x}}(t|t-1) = \mathbf{F}\hat{\mathbf{x}}(t-1|t-1) + \mathbf{D}_{_{Istop}}u_{_{Istop}}(t) \tag{3.17}$$

The one step state prediction for voiced stop model in (3.13) is given by,

$$\hat{\mathbf{x}}(t|t-1) = \mathbf{F}\hat{\mathbf{x}}(t-1|t-1) + \mathbf{D}_{_I}u_{_I}(t) + \mathbf{D}_{_{Ivstop}}u_{_{Iv\,stop}}(t) \tag{3.18}$$

Finally the for the unvoiced fricatives, the state prediction remains the same as in (2.54) i.e.,

$$\hat{\mathbf{x}}(t|t-1) = \mathbf{F}\hat{\mathbf{x}}(t-1|t-1) \tag{3.19}$$

The one-step state prediction error is given by

$$\mathbf{\Sigma_x}(t|t-1) = \mathbf{F}\mathbf{\Sigma_x}(t-1|t-1)\mathbf{F}^T + \mathbf{G}q_w\mathbf{G}^T \tag{3.20}$$

The **Update Steps** remain the same as white noise excited AR based Kalman filter,

$$\kappa(t) = \mathbf{F}\mathbf{\Sigma_x}(t|t-1)\mathbf{H}[q_v + \mathbf{H}^T\mathbf{\Sigma_x}(t|t-1)\mathbf{H}]^{-1} \tag{3.21}$$

$$\hat{\mathbf{x}}(t|t) = \hat{\mathbf{x}}(t|t-1) + \kappa(t)[z(t) - \mathbf{H}^T\hat{\mathbf{x}}(t|t-1)] \tag{3.22}$$

$$\mathbf{\Sigma}(t|t) = [\mathbf{I} - \kappa(t)\mathbf{H}^T]\mathbf{\Sigma_x}(t|t-1) \tag{3.23}$$

The speech sample estimate $\hat{x}$ at time t is given by,

$$\hat{x}(t) = \mathbf{H}^T\hat{\mathbf{x}}(t|t) \tag{3.24}$$

3.6 Parameter Estimation for Impulse Driven AR Model

The inclusion of the weighted excitation term in (3.3), (3.9) and (3.12) implies that the conventional covariance LP analysis[16], which applies to (2.2), needs to be modified. The principle of covariance LP analysis is just parameter estimation to minimize a least-squares criterion

$$C_K = \sum_{t=0}^{K-1} \epsilon(t)^2 \tag{3.25}$$

where K is length of the speech segment (frame) being processed, and where the error is given by the model residual for (3.3)

$$\epsilon(t) = x(t) - \sum_{i=1}^{N_x} \hat{a}_k x(t-i) + \hat{a}_{N_x+1} u_I(t) \tag{3.26}$$

The optimal parameters are found by finding the roots of the squared error (3.25),

$$\frac{\partial C_K}{\partial \hat{a}_j} = 0, \quad 1 \leq j \leq N_x \quad \& \quad \frac{\partial C_K}{\partial \hat{a}_{N_x+1}} = 0 \tag{3.27}$$

leading to a set of linear equations:

$$\begin{bmatrix} \Phi(i,k) & \Psi(i,0) \\ \Psi^T(i,0) & R_u \end{bmatrix} \begin{bmatrix} \hat{\mathbf{a}} \\ \hat{a}_{N_x+1} \end{bmatrix} = \begin{bmatrix} \Phi(i,0) \\ \Psi(0,0) \end{bmatrix} \tag{3.28}$$

which is easily solved, using the Cholesky decomposition, for the unknowns $\hat{\mathbf{a}} = [\hat{a}_1, \ldots, \hat{a}_N]^T$ and $\hat{a}_{N+1}$. The terms in the square matrix are the correlation terms: $\Phi(i,k)$ the cross-correlation matrix of clean speech given by,

$$\Phi(i,k) = \sum_{t=0}^{K-1} x(t-i)x(t-k) \tag{3.29}$$

$$\Psi(i,k) = \sum_{t=0}^{K-1} x(t-i)u_I(t-k) \tag{3.30}$$

the cross-correlation between clean speech and the excitation, and

$$R_u = \sum_{t=0}^{K-1} u_I(t)^2 \tag{3.31}$$

the energy (zero-lag autocorrelation) of the excitation u_I. In the same way AR parameters for (3.9) and (3.12) can be estimated.

3.7 Model Assertions and Parameters Assumptions

As main objectives were to come up with appropriate models for various speech types and study enhancement limits of the AR model based Kalman filter, we have made a number of model assertions and explicit assumptions. The first assertion

we have made for our models is that phoneme boundaries are known. Phonemes, used for testing our speech models, were extracted from (20) sentences spoken by 10 female and 10 male speakers from the TIMIT data base. Phoneme boundaries given in the TIMIT base were initially used to accomplish this separation, followed by the inspection of spectrograms and temporal plots of each sentence to corroborate the exact phoneme boundaries. We have used clean speech for estimating AR parameters and noise covariances for the Kalman filter.

For our impulsive models we have explicitly assumed that pitch locations for the voiced speech and impulse locations for the stop bursts are known. Even though a number of automatic pitch detection algorithms [55, 43] are available, there always exists a room for error in the results obtained by using these algorithms. The times are approximated manually from the residual signal (3.2) in which the pulses are conspicuous, followed by an automated local peak-finder to guarantee accurate positioning.

3.8 Experimental Results

Each speech signal, representing a single phoneme, is segmented into frames of $K = 256$ data points. The Kalman filter was used as the estimation algorithm, using one of the four different models in (2.2), (3.3), (3.9) and (3.12). The speech signals were corrupted with additive white noise to an SNR of 5 dB; for each signal the identical noise process was added, so that output SNR results are meaningfully comparable.

3.8.1 Voiced Speech

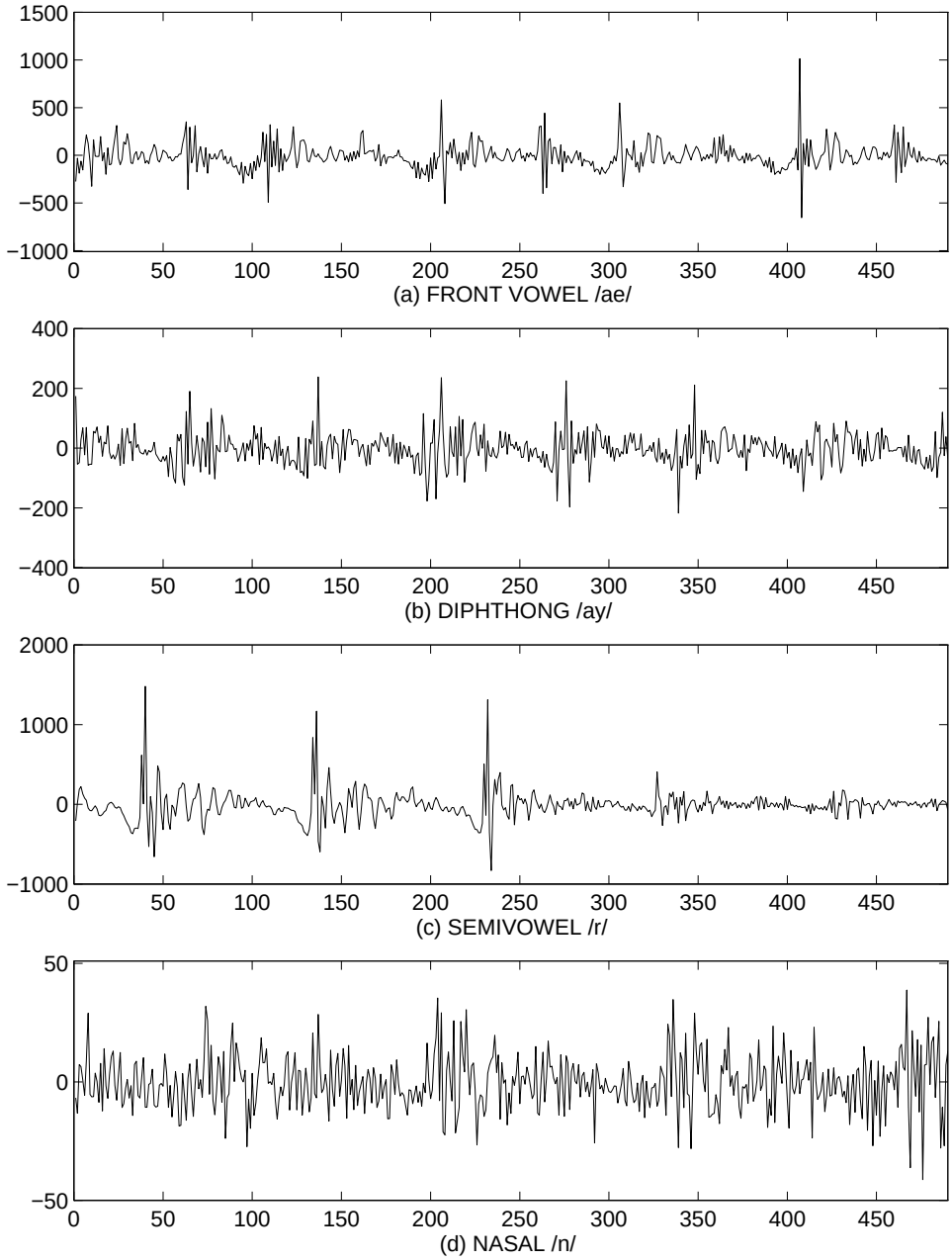

Figure 3.8: AR residuals for the impulsive model 3.3 for the voiced phonemes of Figure 3.6. The residual peaks have become narrower and shorter compared to those in Figure 3.6 but still are noticable.

Phone class	Output SNR in dB white noise AR model	Output SNR in dB impulsive AR model
Front vowels	8.492	9.129
Mid vowels	8.893	9.5614
Back vowels	9.515	10.158
Semivowels	9.087	9.568
Nasals	9.042	9.480
Diphthongs	9.206	9.910

Table 3.3: Averaged enhancement results for voiced speech for the input SNR of 5 dB and the lpc order of 10.

A total of 80 phone tokens for the voiced speech vowels, diphthongs, semivowels and nasals were tested with the impulse excited AR model in (3.3). For the voiced speech we have used lpc order of 10 in general. Figure 3.8 shows the impulse-AR residuals given by (3.26) for the same four phonemes of Figure 3.6. In general the residual pulses in Figure 3.8 are thinner or narrower and shorter than those in Figure 3.6, but still conspicuously present. Impulsive driving term partially fails to model the voice source excitation of the vocal tract. The reason being the peaks in the residual usually do not consist of a single impulse rather have a very narrow triangular shape with more than one adjacent peaks (usually two or three). We also observe a distinct deterministic shape between the sharp spikes. Such facts strongly indicate that the effects of the glottal source may simply not be a train of impulses but rather may be a quasi-periodic pulse. This is confirmed by the literature on voice source models and glottal pulse models (discussed in detail in Chapter 4). Table 3.3 compares the averaged output SNRs for voiced speech tokens.

Input SNR	Front Vowels Change in Output SNR in dB for the lpc Order				Mid Vowels Change in Output SNR in dB for the lpc Order			
in dB	8	10	12	14	8	10	12	14
0	0.816	0.962	1.076	1.261	0.875	0.954	1.033	1.232
5	0.569	0.689	0.790	0.920	0.615	0.674	0.739	0.907
10	0.378	0.473	0.550	0.643	0.425	0.471	0.524	0.659
15	0.233	0.301	0.352	0.415	0.280	0.314	0.353	0.454
Input SNR	Back Vowels Change in Output SNR in dB for the lpc Order				Diphthongs Change in Output SNR in dB for the lpc Order			
in dB	8	10	12	14	8	10	12	14
0	0.899	1.013	1.070	1.251	0.907	1.062	1.168	1.318
5	0.622	0.699	0.736	0.882	0.655	0.765	0.859	0.995
10	0.434	0.488	0.514	0.628	0.458	0.537	0.609	0.721
15	0.304	0.343	0.366	0.455	0.299	0.357	0.405	0.489
Input SNR	Semivowels Change in Output SNR in dB for the lpc Order				Nasals Change in Output SNR in dB for the lpc Order			
in dB	8	10	12	14	8	10	12	14
0	0.704	0.778	0.852	0.932	0.594	0.608	0.639	0.644
5	0.507	0.563	0.622	0.688	0.431	0.438	0.460	0.466
10	0.352	0.394	0.443	0.496	0.307	0.313	0.330	0.336
15	0.238	0.269	0.313	0.353	0.204	0.211	0.225	0.230

Table 3.4: Averaged improvements in output SNR for the white noise AR model and impulsive AR model for voiced speech classes.

lpc order	Phone class	Output SNR in dB white noise AR model	Output SNR in dB impulsive AR model
12	Fricatives	7.0488	not applicable
12	Unvoiced fricatives	9.537	10.0493
10	Voiced stops	7.59	8.088
10	Unvoiced stops	7.174	7.314
10	Affricates	8.417	8.524
12	Unvoiced fricatives	9.537	10.0493

Table 3.5: Averaged enhancement results for the consonants for the input SNR of 5 dB.

We observe that improvement in output SNRs is consistent for the impulse driven AR model over the white noise driven AR model.

Table 3.4 shows averaged improvements in output SNRs for impulse excited AR model over the conventional white noise driven AR model for the lpc orders of 8, 10, 12 and 14 and for the input SNRs of 0 dB, 5 dB, 10 dB and 15 dB. For all the voiced speech types the improvement is maximum for the input SNR of 0 dB. The improvement in output SNRs also increase linearly with lpc order. The diphthongs show highest improvement in output SNRs compared to other voiced speech types. The highest improvement for diphthongs is 1.318 dB in the output SNR for the lpc order of 8. The main reason behind this is that diphthongs are very long phonemes with large number of pitch periods. Hence, the effect of the impulsive driving term is stronger compared to other phonemes.

3.8.2 Consonants

Table 3.5 shows averaged results for 30 consonant phones. In general impulsive models yield higher output SNRs compared to those for the conventional white noise excited AR model. Analysis of the results for individual classes is presented as follows.

Unvoiced fricatives use the conventional white noise excited AR model. We have used lpc order of 12 in this case. Table 3.12 in Appendix:A shows the results for unvoiced fricatives indicating consistent improvements in output SNRs from the input SNR of 5 dB.

Voiced fricatives have been represented by the same impulsive model in (3.3) as in the voiced speech case. One problem encountered in applying such a model was in identifying the pitch locations for some of the voiced fricatives as they do not show marked periodicity as in the case of other voiced speech types such as vowels or diphthongs. Thus using the white noise excited AR model is recommended for the unvoiced fricatives. The results for voiced fricatives are shown in Table 3.13 in Appendix:A. As expected, an AR model with impulsive excitation gives better enhancement than conventional white noise excited AR model. The residuals do not show marked periodicity in this case.

Table 3.14 in Appendix:A presents enhancement results for unvoiced stops, voiced stops and affricates. In general impulsive model works better than white noise driven AR model. But in some of the cases for the stops (e.g. /b/ in Table 3.14) we have found AR models do not work well at all in the sense the output SNRs for both white noise driven AR and impulse excited AR models were less than the input SNR of 5 dB. This can be explained by the fact that stops have complex acoustic sequence as given by (3.8) and trying to model such a sequence

with an all-pole model may not be pertinent. Modeling of the stops may require further investigation. Same arguments, as in case of the stops, can be made for modeling the affricates which also possess too complex acoustic sequences for AR models.

3.9 Conclusions

This chapter has demonstrated the application of the AR models with an impulsive excitation term for speech enhancement. The conventional white noise excited AR model for speech fails to account for the excitation source especially in the case of the voiced speech. As various speech classes have different forms of source excitations we have aptly proposed impulse driven AR models with different driving terms for various classes of phonemes. We have represented voiced speech types such as vowels, diphthongs, semivowels, nasals and voiced fricatives by an AR model driven by impulsive train time modulated by the pitch periods and white noise. Unvoiced fricatives were modeled by traditional white noise driven AR model. For unvoiced stops and unvoiced affricates we have used an AR model driven by a single impulse at the instant of the burst onset and white noise. For voiced stops and voiced affricates, we have proposed an AR model both with an impulsive train for voicing, a single burst impulse and white noise. In each case, especially in case of the voiced speech, the Kalman filter with impulse excited AR models clearly outperformed that with a traditional white noise AR model.

One major flaw of the impulsive models is that they are too simplistic for simulating the complex speech excitations. We have come to this conclusion by inspecting impulse driven AR residuals. Especially in the case of voiced speech we have observed the marked presence of quasi-periodic pulses in the residuals. This

drawback of the impulsive models is also confirmed by Speech Synthesis literature where a number of more sophisticated source models[40, 36, 37, 38, 59, 41] replacing impulsive models have been proposed for producing natural sounding synthetic speech. Such facts present greatly motivate us towards using more complex model for representing effects of the voice source in Chapter 4.

3.10 Appendix A: Details of Enhancement Results

Phone	Phone context	SNR in dB white noise AR model	SNR in dB Impulsive AR model
/iy/	*iy*69*si*682_*economically_y*	8.455	9.106
/iy/	*iy*32*si*1386_*the_e*	10.399	10.576
/ih/	*ih*13*sx*378_*near_i*	8.587	9.108
/ih/	*ih*49*si*1386_*unit_i*	9.216	10.078
/ih/	*ih*4*sx*396_*fish_i*	9.077	9.263
/ix/	*ix*29*si*1386_*imagination_i*2	7.250	8.102
/ix/	*ix*17*si*1386_*negotiation_io*	8.589	9.137
/ix/	*ix*31*si*682_*only_y*	8.097	8.675
/eh/	*eh*9*si*682_*they_e*	8.542	9.241
/eh/	*eh*42*si*682_*area_a*1	8.713	9.917
/eh/	*eh*58*si*682_*economically_e*	7.635	8.287
/ae/	*ae*26*sx*96_*imagination_a*1	8.062	9.310
/ae/	*ae*11*sx*396_*began_a*	8.542	8.815
/ae/	*ae*6*sx*86_*hispanic_a*	7.731	8.192

Table 3.6: Enhancement Results for the Front Vowels for the input SNR of 5 dB and lpc order of 10. Phone context *iy*69*si*682_*economically_y* implies that /iy/ is the 69*th* phone from the sentence *si*682 and is taken from the word *economically* corresponding to the letter *y*.

Phone	Phone context	SNR in dB white noise AR model	SNR in dB Impulsive AR model
/ah/	*ah*12*sx*119_*was_a*	9.791	10.729
/ah/	*ah*38*sx*378_*museum_u*2	8.259	8.811
/ah/	*ah*21*sx*96_*ones_o*	8.292	9.171
/ah/	*ah*24*sx*86_*colorful_o*1	9.142	9.587
/ah/	*ah*13*si*1051_*are_ah*	8.929	9.850
/ax/	*ax*29*sx*119_*apology_o*1	8.917	9.400
/ax/	*ax*2*sx*396_*the_e*	8.716	9.164
/ax/	*ax*41*sx*396_*the_e*	8.066	8.375
/ax/	*ax*6*si*682_*often_e*	8.938	9.700
/axr/	*axr*17*sx*86_*are_ar*	8.349	8.891
/axr/	*axr*19*sx*210_*never_er*	9.331	9.849
/er/	*er*34*sx*396_*surface_ur*	9.278	9.663
/er/	*er*21*sx*210_*worked_er*	9.393	10.137

Table 3.7: Enhancement Results for the Mid Vowels for the input SNR of 5 dB and the lpc order of 10.

Phone	Phone context	SNR in dB white noise AR model	SNR in dB Impulsive AR model
/uh/	*uh*28*sx*86_*colorful_u*	9.833	10.096
/ux/	*ux*14*sx*86_*costume_u*	8.172	8.735
/ux/	*ux*15*sa*1_*suit_o*	7.938	8.419
/uh/	*uh*15*sx*396_*to_o*	8.468	8.819
/ow/	*ow*12*si*1386_*negotiations_o*1	9.755	10.502
/ow/	*ow*28*si*682_*only_o*	10.497	11.218
/ow/	*ow*9*sx*119_*misquote_uo*	9.791	10.729
/ow/	*ow*2*sa*2_*dont_o*	9.219	10.110
/ao/	*ao*29*sx*396_*Of_o*	9.478	10.043
/ao/	*ao*2*si*682_*often_o*	10.130	11.124
/ao/	*ao*25*sa*1_*wash_a*	9.308	9.763
/ao/	*ao*25*si*1051_*supporters_o*	9.618	10.105
/ao/	*ao*25*si*1051_*northern_o*	8.970	9.501
/aa/	*aa*32*sx*119_*apology_o*1	9.224	10.515
/aa/	*aa*10*sx*96_*parties_a*	8.398	9.236
/aa/	*aa*35*si*1386_*bargain_a*1	9.121	10.069
/aa/	*aa*11*sa*1_*dark_a*	8.532	8.920
/aa/	*aa*13*sx*210_*cart_a*	9.402	10.039

Table 3.8: Enhancement Results for the Back Vowels for the input SNR of 5 dB and the lpc order of 10.

Phone	Phone context	SNR in dB white noise AR model	SNR in dB Impulsive AR model
/y/	$y64si1386\_union\_i$	9.425	10.062
/y/	$y34sa1_{y}ear_{y}$	7.764	8.329
/y/	$y7sa1_{y}our_{y}$	8.293	8.515
/w/	$w51si1386\_with\_w$	9.967	10.614
/w/	$w24sa1\_wash\_w$	9.363	9.602
/w/	$w20sx96\_ones\_o$	9.781	10.680
/r/	$r10si682\_are\_r$	9.101	9.105
/r/	$r20sx396\_frantically\_r$	8.940	9.115
/r/	$r11sx96\_parties\_r$	8.398	9.376
/l/	$l30si682\_only\_l$	8.243	8.739
/l/	$l16sx396\_leap\_l$	8.715	8.849
/l/	$l26sx378\_archeological\_l1$	8.941	9.707
/el/	$el33sx378\_archeological\_l2$	9.368	9.435

Table 3.9: Enhancement Results for the Semivowels for the input SNR of 5 dB and the lpc order of 10.

Phone	Phone context	SNR in dB white noise AR model	SNR in dB Impulsive AR model
/m/	*m*1*sx*96_*masquerade_m*	10.664	10.860
/m/	*m*11*sx*378_*jim_m*	8.885	9.409
/m/	*m*64*si*682_*economically_m*	8.716	9.457
/m/	*m*44*sx*396_*small_m*	9.010	9.079
/n/	*n*12*sx*378_*near_n*	9.395	9.633
/n/	*n*22*sx*96_*imagination_n*	9.983	10.271
/n/	*n*30*sx*396_*on_n*	9.202	9.688
/ng/	*ng*56*si*682_*declining_ng*	6.868	7.855
/ng/	*ng*57*si*1386_*single_ng*	9.663	10.227
/ng/	*ng*42*si*811_*coagulating_ng*	8.042	8.322

Table 3.10: Enhancement Results for the Nasals for the input SNR of 5 dB and the lpc order of 10.

Phone	Phone context	SNR in dB white noise AR model	SNR in dB Impulsive AR model
/yu/	*yu*18*sx*378_*new_ew*	9.106	9.694
/ey/	*ey*7*sx*96_*masquerade_a*2	8.106	9.143
/ey/	*ey*47*sx*396_*lake*	8.3113	8.452
/ey/	*ey*12*si*682_*able*$_a$	9.612	10.643
/ay/	*ey*45*si*1386_a_a	9.492	10.449
/ay/	*ay*53*si*682_*declining*$_i$1	9.136	10.070
/ay/	*ay*6*si*1739_*time*6_*i*	8.781	9.590
/ay/	*ay*21*sx*86_*quite_ui*	8.806	9.521
/oy/	*oy*21*sa*2_*oily_oi*	9.376	10.105
/oy/	*oy*7*sx*196_*oysters_oy*	10.696	11.077
/oy/	*oy*7*sx*210_*toy_oy*	9.850	10.270

Table 3.11: Enhancement Results for the Diphthongs for the input SNR of 5 dB and the lpc order of 10.

Phone	Phone context	SNR in dB white noise AR model
/f/	*f3si682_often_f*	6.206
/f/	*f3sx396_fish_f*	6.429
/s/	*s3sx96.dat_masquerade_s*	7.365
/s/	*s6sx378_saw_s*	7.886
/s/	*sh32sx96_imagination_sh*	7.358

Table 3.12: Enhancement Results for Unvoiced Fricatives for the input SNR of 5 dB and the lpc order of 12.

Phone	Phone context	SNR in dB white noise AR model	SNR in dB Impulsive AR model
/v/	*v18sx210_never_v*	8.154	8.777
/dh/	*dh1sx396_the_th*	8.458	*
/dh/	*dh1sx119_the_th*	10.856	11.217
/dh/	*dh53si1386_with_th*	9.601	10.154
/z/	*z36sx378_museum_s*	7.631	*
/z/	*z44si1386_as_s*	6.931	*

Table 3.13: Enhancement Results for the Voiced Fricatives for the input SNR of 5 dB and the lpc order of 10. $*$ indicates that pitch periods could not be identified.

Phone	Phone context	SNR in dB white noise AR model	SNR in dB Impulsive AR model
/p/	*p*10*sx*96_*parties_p*	5.349	5.437
/p/	*p*5*sx*86_*hispanic_p*	6.821	6.969
/t/	*t*17*sx*96_*tax_t*	9.792	9.837
/k/	*k*60*si*682_*economically_c*	6.461	6.609
/k/	*k*23*sx*378_*archeological_ch*	7.450	7.718
/b/	*b*33*si*682_*onlybecause_b*	4.404	5.795
/b/	*b*7*sx*396_*began_b*	3.343	3.376
/b/	*b*14*si*682_*able_b*	9.373	9.794
/b/	*b*20*si*811_*terrible_b*	11.819	11.718
/d/	*d*4*sx*196_*howdo_d*	8.87	9.20
/d/	*d*26*si*1386_*industry_d*	5.902	7.154
/g/	*g*10*sx*396_*began_g*	8.586	9.2133
/g/	*g*20*si*682_*toget_g*	6.302	6.547
/g/	*g*38*si*1386_*bargains_g*	9.767	9.997
/j/	*jh*9*sx*378_*jim_j*	8.520	8.537
/j/	*jh*4*sx*378_*just_j*	8.921	9.051
/ch/	*ch*5*sx*378_*church_ch*	7.812	7.985

Table 3.14: Enhancement Results for the Stops and the Affricates for the input SNR of 5 dB and the lpc order of 10.

Chapter 4

LF Model for Enhancement of Voiced Speech

This chapter proposes and implements an LF model (proposed by Fant, Liljencrants and Lin[1]) based AR model for voiced speech enhancement. Section 4.1 discusses the drawbacks of white noise and impulse driven AR models and motivates the application of an LF voice source model commonly used in speech synthesis and analysis, for speech enhancement. Some of the popular glottal and source models used in speech synthesis are reviewed briefly by Subsection 4.2.1 while Subsection 4.2.2 is dedicated to details of an LF model. For speech enhancement, the parameter estimation problem is different from that for speech synthesis or analysis. The parameter estimation problem associated with an LF model for speech enhancement is discussed in detail in Section 4.3. Section 4.4 discusses the results obtained by an LF model based enhancement and draws a comparison among the three different types of excitation models, i.e., a white noise model, an impulsive model and an LF model . Conclusions about an LF based AR model in speech

enhancement is presented in Section 4.5. Section 4.6 contains the tables for LF model based enhancement results.

4.1 Introduction

An AR model driven by white noise, traditionally used for speech enhancement does not take account of the effects of excitation sources for some of the phoneme classes especially those with voiced excitation. In Chapter 3, we have proposed a number of impulse driven AR models for various phoneme classes based on the corresponding excitation types. Impulse excited AR models consist of impulsive deterministic terms which *also* are the very simple *tentative* models for capturing the effects of the excitation source. For voiced speech, the effects of glottal excitation was simulated by a train of impulses spaced by pitch periods. For unvoiced stops and unvoiced affricates, plosive excitation was modeled by a single impulse marking the instant of the onset of the burst and white noise. For voiced stops and voiced affricates, a mixed excitation of a plosive driving term and a quasi-periodic train of impulses were proposed. For voiced fricatives a mixed excitation of white noise and a quasi-periodic train of impulses separated by pitch periods was proposed. Impulsive models, despite their simplicity, yielded considerable improvements in the output SNRs.

In the case of the voiced speech classes such as vowels, semivowels, diphthongs and nasals, residuals for an AR model with impulsive driving term as shown in Figure 3.8 show considerable periodicity even though residual impulses have become narrower and shorter compared to those in white noise excited AR residuals in Figure 3.6. We also observe a continuous curve between the quasi periodic spikes. Such facts strongly suggest that an impulsive model is too simple a model for speech.

To be sure, in speech synthesis and analysis, the modeling of the voice source has been well studied[60, 39, 36, 37, 38, 41, 40]. In fact, AR residuals in Figure 3.8 indicate a quasi-periodic shape which resembles that of the voice source used in speech synthesis. Such facts have strongly motivated us to believe that models for voice source pulses have good potential for speech enhancement. In the following section, we review some of the significant glottal pulse and voice source models proposed and implemented in speech synthesis and analysis. We also present the reasons for selecting an LF model for speech enhancement followed by a detailed discussion of the model.

4.2 Voice Source Models

An impulsive model is a highly simplified approximation of the human voice. Indeed, impulsive-driven systems were found to make poor speech synthesizers, so the synthesis field has proposed a number of more complex glottal pulse models[40, 59, 36, 37, 38, 41] for producing more natural sounding speech. In speech synthesis literature the volume velocity of the air flow is referred to as glottal pulses and the derivative of the glottal pulses are known as voice source pulses. The next subsection briefly reviews some of the voice source models used in speech synthesis and analysis.

4.2.1 Review of Voice Source Models

Rosenberg [40] has proposed a number of glottal pulse models with adjustable amplitude, width and skew. These glottal pulse models were used to study their effects on the quality of vowels. Of all these models, one pulse shape shown in

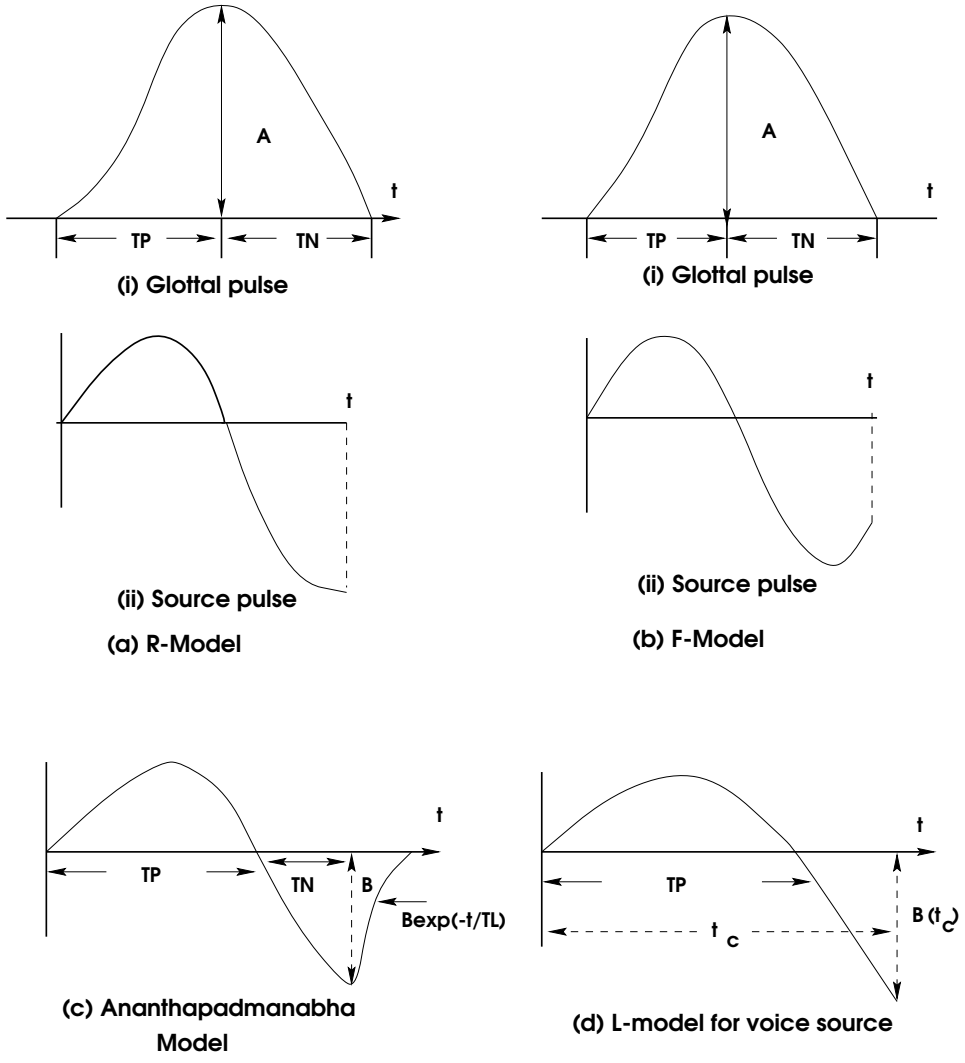

Figure 4.1: Models for glottal and voice source pulses.

Figure 4.1(a) consists of two trigonometric segments with a slope discontinuity at the closure. This model is referred to as the Rosenberg model and has had significant effect on speech synthesis researchers at the time because of its capability of producing better quality synthetic speech compared to other models[59]. The Rosenberg model for glottal flow is specified by three parameters: TP the portion of the pulse with positive slope, TN the portion of the pulse with negative slope and A, the amplitude of the glottal pulse. The Rosenberg source pulse model, in Figure 4.1a(ii), has abrupt discontinuity at the glottal closure and shape of the source pulse model in the vicinity of the closure is sinusoidal. For these reasons, the Rosenberg model is not suitable for fitting the AR residual spikes in Figure 3.6.

Fant has proposed another three parameter model referred to as the F-model by introducing an independent control of the discontinuity at the closing phase of the source pulse[37] as shown in Figure 4.1(b). The three parameter models though economical, failed to capture the wide variations of the glottal pulse shapes. Another major flaw, of the three parameter models, was that abrupt discontinuity at the glottal closure which does not allow for an incomplete closure of the vocal folds or for a residual phase to proceed towards the closure after the discontinuity[1]. For similar reasons as in the case of the Rosenberg model, the F-model is not feasible for application in speech enhancement .

Ananthapadmanabha introduced a five parameter model of the voice source, rather than the glottal pulse, as shown in Figure 4.1(c) which models various variations of the source pulse with a terminal return phase[59, 41]. The return phase in this case is modeled as a parabolic function which tracks abduction states of the vocal folds. The five parameter model does not have the disadvantage of the discontinuity at the closure. But as the pulse shape at the closure does not have a sharp peak as in the case of the AR residuals, it makes the model less desirable as

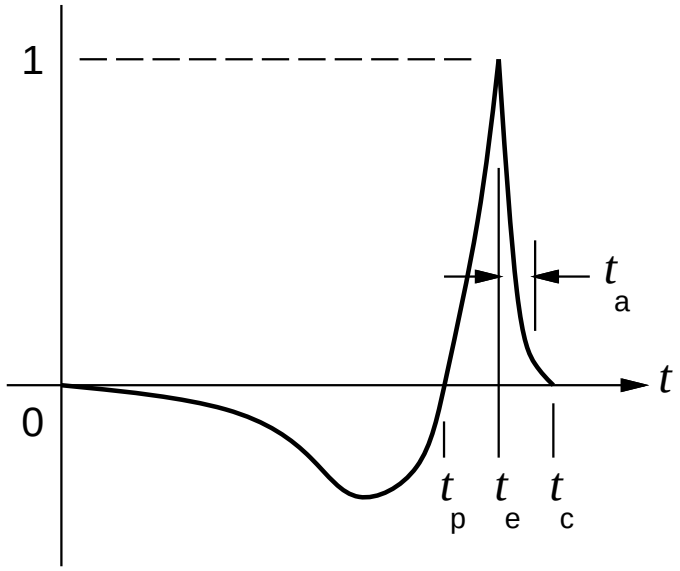

Figure 4.2: The LF deterministic excitation model.

a candidate for representing the driving term in an AR model.

The most popular model for voice source is referred to as the **LF model** as shown in Figure 4.2. It was developed in two stages. In the first stage, Liljencrants proposed a three parameter model of voice source based on the F-model [1, 59, 41] as shown in Figure 4.1(d). This model is referred to as the L-model. The L-model model has the advantage of flow continuity whereas the F-model consists of two portions one with positive slope and the other with negative slope. The main flaw in an L-model is the abrupt flow termination. In the latter stage the L-model was modified by Fant, Liljencrants and Lin[1] by introducing a gradual flow termination modeled by an exponential function. This new modified model is known as an LF model. The reason behind the *popularity* of an LF model is that it provides an overall fit to commonly encountered voice source pulse shapes in speech synthesis and analysis with a minimum numbers of parameters and is flexible in its ability to match extreme cases of phone variabilities[1]. Of all the voice source models

reviewed so far, we observe a strong similarity in the shape of the AR residual pulses in Figure 3.6 and that of the LF model. In fact AR residual spikes resemble the shape of the LF model around the instant of the glottal closure. These facts have motivated us to choose an LF model for representing the effects of the voice source in an AR model for the voiced speech. In the following subsection we discuss the LF model in detail.

4.2.2 LF Model

The four parameter LF model[1] proposed by Fant, Liljencrants and Lin has been widely used practically in speech synthesis and theoretically in speech analysis[39]. The LF excitation model, sketched in Figure 4.2, is the derivative of the LF glottal pulse function, and is parameterized in terms of

- t_c – the fundamental period,
- t_p – the instant of maximum flow,
- t_e – the instant of maximum glottal closing, and
- t_a – exponential recovery time constant.

These four parameters are related to each other by a condition that net flow gain within a fundamental period must be zero.

The LF model is then given by

$$u_{LF}(t) = \begin{cases} e^{\alpha t} \sin \omega_g t & t \leq t_e \\ \frac{-1}{\beta t_a}[e^{-\beta(t-t_e)} - e^{-\beta(t_c-t_e)}] & t_e \leq t \leq t_c \end{cases} \tag{4.1}$$

where α, β satisfy the transcendental equations

$$1 - e^{-\beta(t_c-t_e)} = \beta t_a$$

$$e^{\alpha t_e} \sin(\pi t_e/t_p) = -1,$$

leading to the revised AR model

$$x(t) = \sum_{i=1}^{N} a_i x(t-i) + \quad w(t) + a_{N_x+1} u_{LF}(t). \tag{4.2}$$

The AR model in (4.2) has glottal excitation modeled by a train pulses modeled by LF model given by (4.1) and time modulated by pitch periods. There is a difference in the way the speech models are used in speech synthesis and enhancement. For speech synthesis, we have a prior model using which we generate a sample path. For speech estimation, we observe a noisy version of a speech signal and try to fit it to a prior model. Section 4.3 discusses and proposes an optimization algorithm for LF parameter estimation.

4.3 Parameter Estimation for LF Model

The main challenge with using model $u_{LF}(t)$ in (4.2) is the need to estimate the seven parameters $t_c, t_e, t_p, t_a, \alpha, \beta$ and a_{N_x+1}. Only a_{N_x+1} enters the problem linearly, so it is solved using least-squares as described in Section 3.6. Since AR residual peaks coincide with the maximum glottal closure[39], the point of maximum glottal closing t_e is set to coincide with the impulsive points or residual peak positions t_j as described in Subsection 3.4.1, leaving five remaining parameters to be found by nonlinearly optimizing the mean-squared error C_K in (3.25) and the output SNR via coordinate optimization.

We have developed a technique for automatic fitting of the five LF parameters t_c, t_e, t_a, α and β to the AR residuals using coordinate optimization. The Optimization procedure is carried out in two stages as illustrated in Figure 4.3. We have obtained pitch locations using AR residuals as described in Section 3.7. Good initial estimates of the parameters is crucial for our optimization algorithm. In the

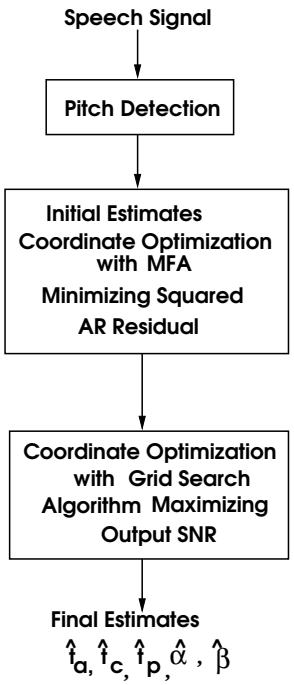

Figure 4.3: Flow chart for LF model parameter estimation. After locating the residual peaks, the initial estimates are obtained using the Minimum Finder Algorithm. Then the initial estimates are fed into the Grid Search Algorithm which gives the final estimates.

first stage of optimization, initial estimates of the parameters are obtained using Minimum Finder Algorithm(MFA) developed by Brent[61]. The MFA combines golden section search and successive parabolic search algorithms(details in [61]) to find a local minimum of a function in a given interval on which it (the function) is defined. We have used the function *fminu* in Matlab based on MFA for obtaining the initial estimates of the LF parameters. The MFA requires the specifications of an upper bound, a lower bound and a termination tolerance for each of the parameters to be estimated and uses the mean-squared AR residual given by,

$$\frac{1}{K}\sum_{t=0}^{K-1}\epsilon^2(t) = \frac{1}{K}\sum_{t=0}^{K-1}[x(t) - \sum_{i=1}^{N_x}\hat{a}_k x(t-i) + \hat{a}_{_{N_x+1}}u_{_{LF}}(t)]^2 \tag{4.3}$$

as cost function to be minimized. The termination tolerance gives the desired length of the final interval on which the cost function is to be minimized. We have set the termination tolerance to a value of 10^{-4} (found empirically which is also the default value used by Matlab) for all the parameters. The bounds on each parameter were estimated by exhaustive testing for a wide range of parameter values for minimizing the cost function in (4.3). It was found that it was necessary to use multiple bounds on the parameters. For optimization, the MFA searches along one parameter coordinate while keeping rest of the parameters constant. Then it updates the estimated parameter and continues the search procedure in other coordinates until all the parameters have been estimated. The order in which the LF parameters were estimated was t_c, t_e, t_a, α and finally β.

Using initial estimates of the LF parameters, another optimization procedure known as grid search algorithm(GSA) via coordinate descent is applied to obtain the final estimates of the parameters. The GSA is illustrated in Figure 4.4. For each coordinate the GSA starts at i the initial estimate for that coordinate found by MFA while keeping other parameters constant at their initial estimate. Search

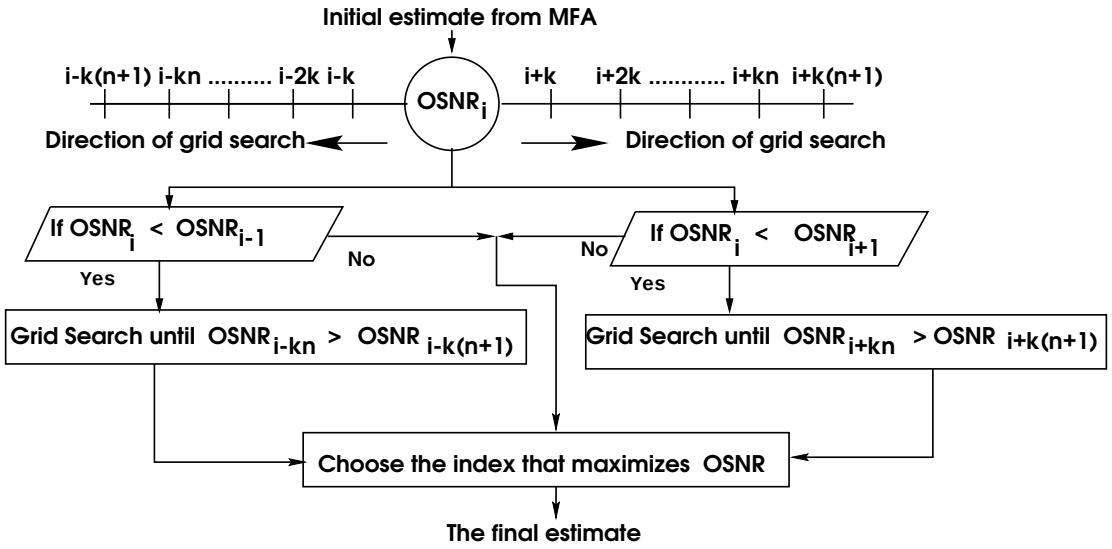

Figure 4.4: Illustration of the Grid Search Algorithm(GSA). $OSNR$ is the output SNR, i is the initial estimate of the parameter at which GSA starts, k is the size of grid and n is the number steps from i at which a maximum is reached. Search may continue in either or both directions until maxima are found. Algorithm selects the index that gives the maximum OSNR.

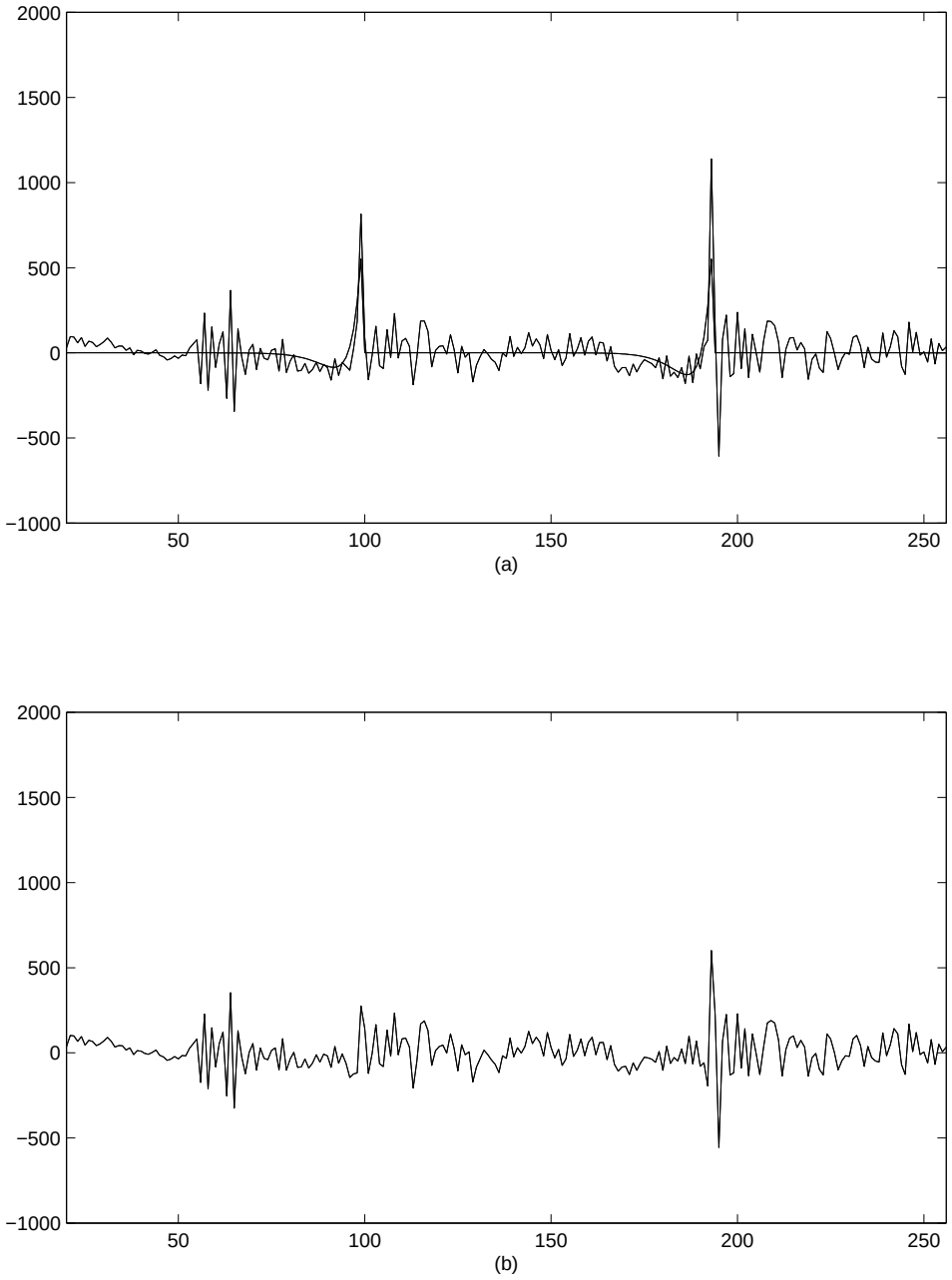

Figure 4.5: AR residuals for Front Vowel /ae/ for one frame: (a) White noise driven AR estimation error with fitted LF model and (b) LF model driven AR estimation error.

for the maximum $OSNR$ (output SNR) starts with either incrementing or decrementing i by the grid size k for that particular parameter and may continue until the maximum $OSNR$ is reached. In certain cases where $OSNR_i$ happens to be in a valley between two peaks, GSA may search in both directions as shown in Figure 4.4. Finally GSA selects the index that gives the maximum $OSNR$.

As mentioned earlier, we have used multiple bound sets for estimating the parameters. The optimization algorithm starts with each set of bounds, finds the initial estimates using the MFA and the final estimates using the GSA for that bound set. The parameter-estimate set, which gives the maximum output SNR, is finally selected.

Figure 4.5(a) shows AR residuals for Front Vowel /ae/ for one frame (256 samples) fitted with LF pulse model by our optimization algorithm. LF model driven residuals shown in Figure 4.5(b) show that the first spike has completely been eliminated and the second spike has been reduced in amplitude to almost 50% compared to that shown in Figure 4.5(a). The main reason behind the second spike not being completely eliminated is that the LF parameters were estimated by fitting the LF pulse train with the AR residuals over the entire length of the speech signal. This was done to keep the computational complexity as low as possible.

4.4 Results

A total of 50 voiced speech phones were taken from the TIMIT database for an LF model based speech enhancement. In order to assess enhancement limits we learn the model parameters separately for each phone as in Chapter 2. Model assertions and parameter assumptions described in Section 3.7 also apply in an LF model based AR model. A frame length of 256 speech samples was used. Noisy signals

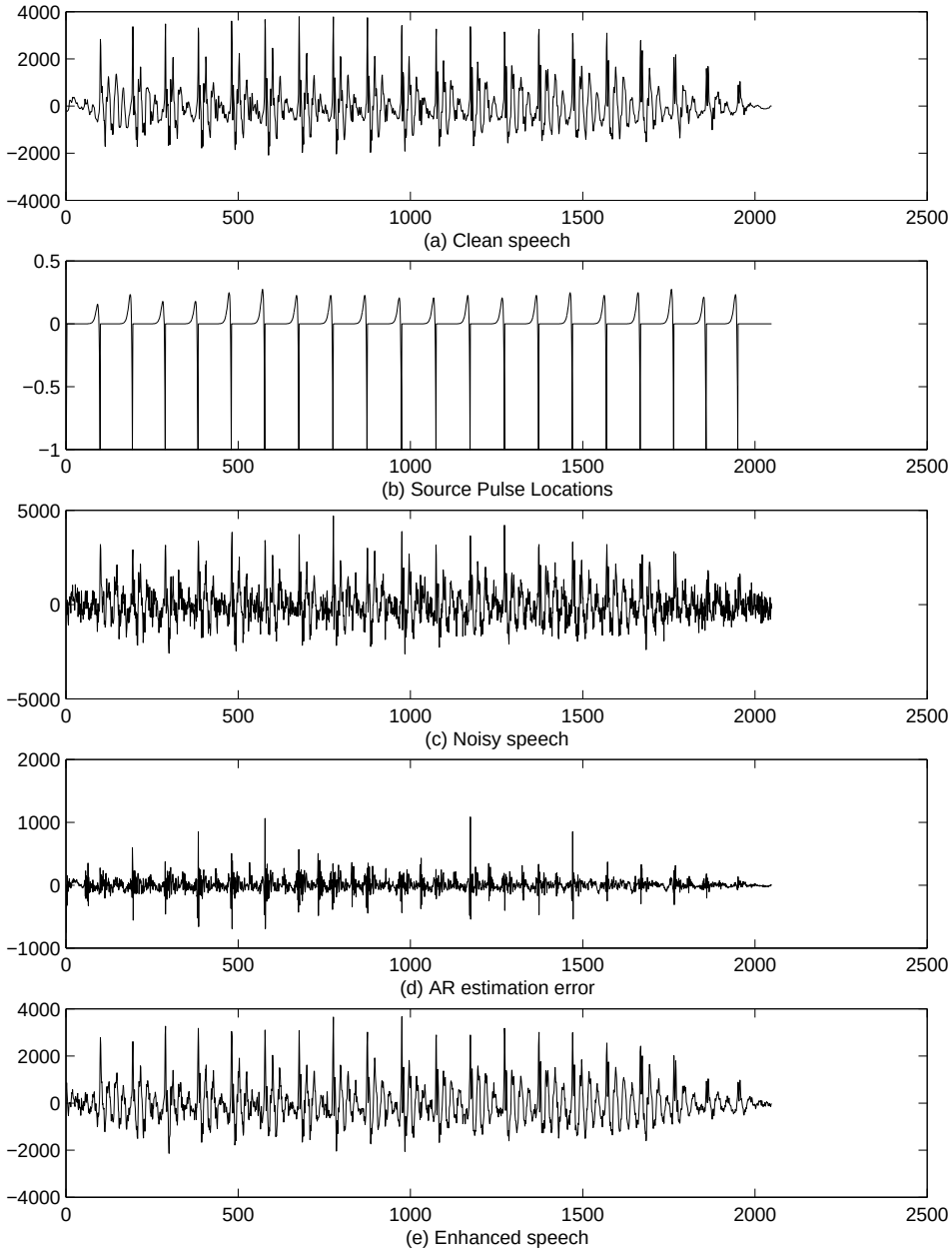

Figure 4.6: Result for Lf based AR model for Front Vowel /ae/ : (a) Clean speech, (b) LF pulse locations, (c) Noisy speech with the input SNR of 5 dB, (d) AR Residual and (e) Enhanced Speech with the output SNR of 10.04 dB.

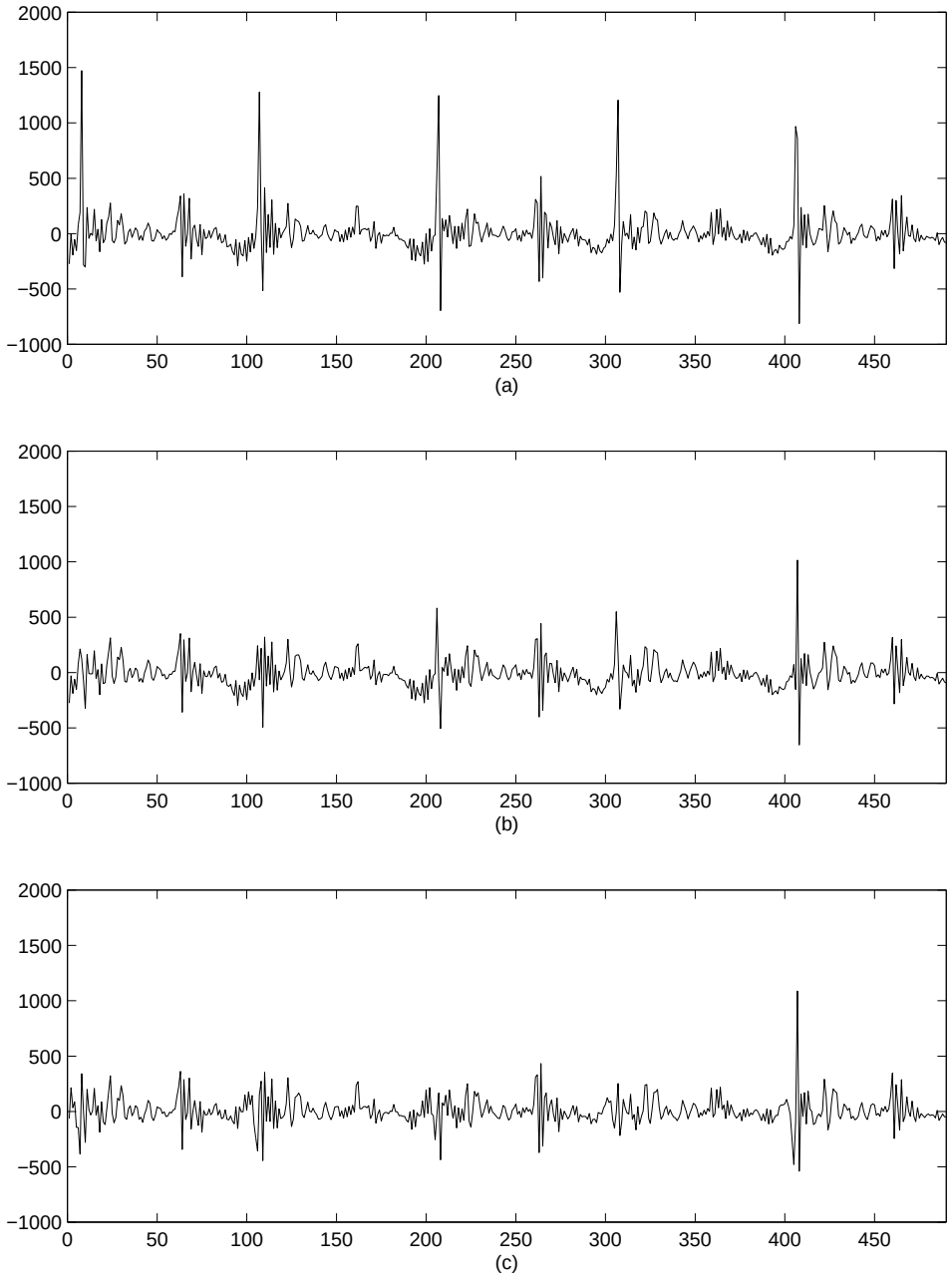

Figure 4.7: AR residuals for Front Vowel /ae/: (a) White noise driven AR model, (b) Impulse driven AR model and (c) LF model driven AR model.

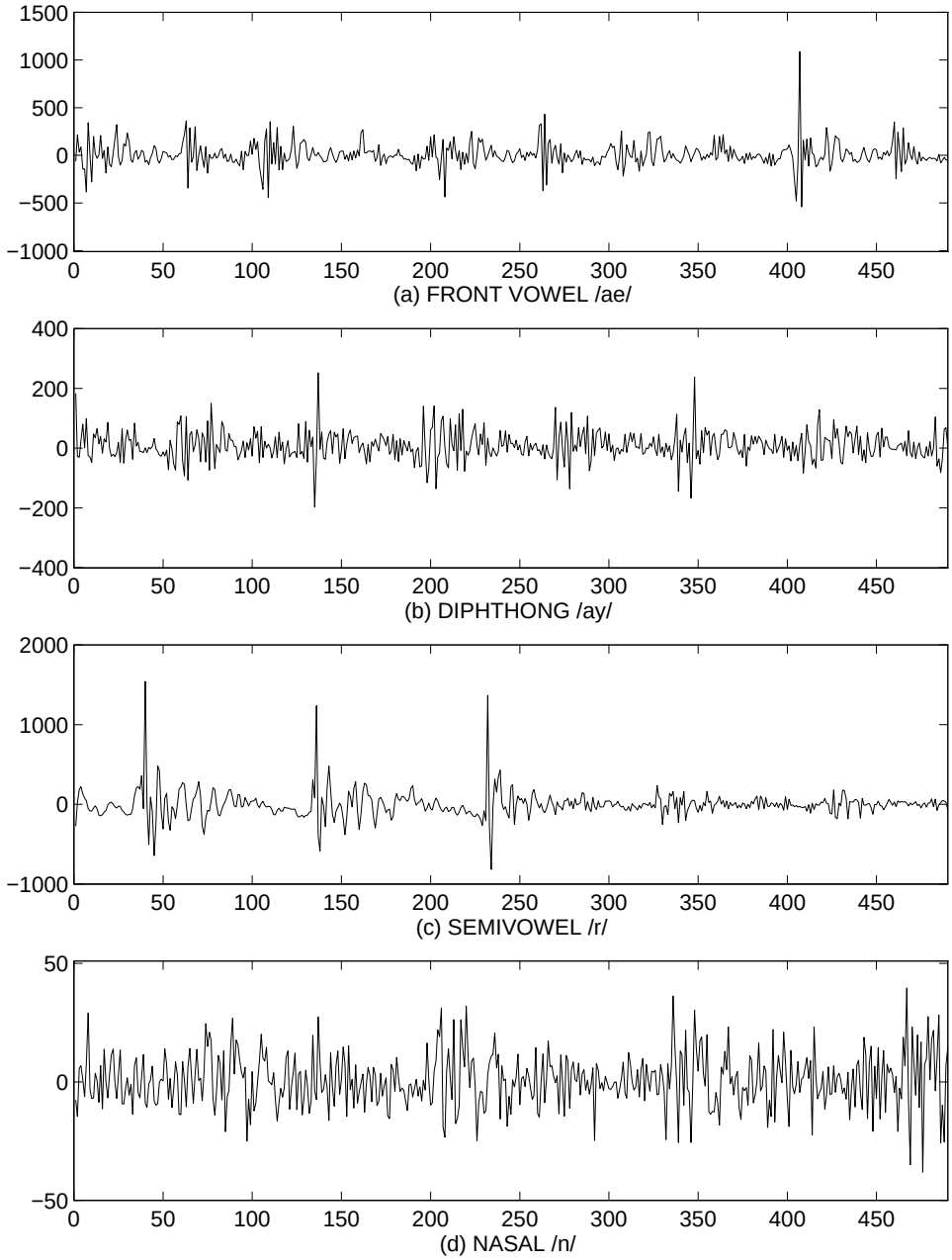

Figure 4.8: AR residuals for the LF model (4.2) for the voiced phones of Figures 3.6, 3.8. The residual spikes either have been reduced or eliminated compared to their white noise driven and impulse driven counterparts.

were created by adding white noise at an SNR of 5 dB. The Kalman filter algorithm described in Section 3.5 was applied.

Figure 4.6 shows enhancement results for the front vowel /ae/. The clean speech /ae/ is shown in Figure 4.6(a). The LF pulse sequence, generated by the optimization algorithm, is shown in Figure 4.6(b). The noisy speech, in Figure 4.6(c), is generated by adding white to the clean speech in Figure 4.6(a) at the the input SNR of 5 dB. Figures 4.6(d) and 4.6(e) respectively show the AR residuals and the enhanced speech for an AR model with LF model based excitation. The LF model based enhancement not only shows considerable reduction in the periodicities in the AR residual but also substantial improvement in the output SNR (10.04 dB) compared to those for white noise (output SNR of 8.06 dB) and impulsive model (output SNR of 9.31 dB) based excitations. Figure 4.7 shows AR residuals for two frames (512 samples) for the front vowel /ae/ in Figure 4.6. We observe in Figure 4.6(b) that the impulsive excitation fails to eliminate all the periodic spikes which clearly demonstrate that effects of glottal excitation can not be modeled efficiently by a quasi-periodic train of impulses. The LF model based AR residual in Figure 4.6(c) alleviates the effects of the glottal excitation to a considerable extent. In fact all except one spike (around the sample 410) have been completely eliminated. Even the lone spike has its negative amplitude considerably reduced.

Figure 4.8 shows the AR-LF residuals, paralleling the earlier results of Figures 3.6 and 3.8. In moving from the purely impulsive to the LF model, the top two panels (front vowel and diphthong), in particular, show a reduction and thinning of residual spikes and exhibit less deterministic structure. A close examination of the figures reveals a substantial limitation in u_I which begins to be addressed in u_{LF}: an impulse $\delta(t)$ is exactly one sample wide, whereas the width of the residual spikes in Figure 3.6 and of the peak in u_{LF} are clearly sampling-rate dependent, and

are frequently, although not always, more than one sample in width. A similar issue can be raised in terms of sampling origin: a single glottal burst may, depending on the sampling origin, be captured as a single impulse or as two smaller impulses. An impulse-train u_I cannot properly address this issue, whereas u_{LF} is a continuous signal and lends itself naturally to resampling. The third panel (Figure 3.6(c)) showing the residuals for the semivowel /r/ still exhibits a periodic component to considerable extent. This shows the shortcoming of our optimization algorithm which fails to find a good fit of the LF pulse in such a case. The two main reasons, that can be associated with the poor parameter estimation, are usage of improper bound set for initial estimates and optimization over the entire length of the speech signal. The fourth panel (Figure 3.6(d)) for the nasal /n/ shows periodicity to extremely small extent. As for the nasals the effects of the voice source cancels out by the zeros in the nasal cavity, they do not possess conspicuous periodic trends in the residuals.

To assess the models more objectively, Tables 4.2, 4.3, 4.4, 4.5 and 4.6 and 4.7 in Appendix:B present results obtained for the front vowels, the mid vowels, the back vowels, the diphthongs, the semivowels and the nasals respectively. Table 4.1 summarizes the SNR improvement for each of the three proposed models, tested on fifty different voiced phones. Most importantly, consistent and nontrivial improvements in SNR are realized, first by the impulsive model, then additionally by the LF model, for *all* voiced phones tested. The LF model based AR model achieves an average improvement of 1.271 dB in output SNR over that for white noise driven AR model. Due to the assertions and the assumptions made by the model, the output SNRs also indicate the limits to performance of the Kalman filter. Among all the phoneme classes, the front vowels yield the highest improvement of 1.987 dB over the conventional AR model. Some of the diphthongs in Table 4.5 have

Phone class	Output SNR in dB white noise AR AR model	Output SNR in dB impulsive AR model	Output SNR in dB LF model-based AR model
Front vowels	8.666	9.398	10.653
Mid vowels	9.181	9.677	10.068
Back vowels	9.245	9.835	10.640
Semivowels	9.121	9.633	10.216
Nasals	8.777	9.293	9.651
Diphthongs	9.332	10.128	10.490
Mean over all phoneme classes	9.115	9.625	10.386

Table 4.1: Averaged enhancement results for voiced speech phones for input SNR of 5 dB and lpc order of 10.

very small improvements in the output SNRs with the LF model over its impulsive counterpart. One of the reasons may be due to the fact that we are optimizing LF parameters by fitting a long train of LF pulses with a long train of quasi periodic AR residual spikes. Another reason may be associated the problem associated with the sampling of the LF model. The parameter t_e of the LF model was made to coincide with pitch periods i.e. with the peaks of the AR residual. While sampling the LF model we must have missed adjacent peaks which contributes to low output SNRs.

4.5 Conclusions

This chapter has clearly established the applicability of an AR model excited by an LF model for voiced speech enhancement purposes. The effects of the voice source is modeled as an LF pulse train time modulated by pitch periods of the voiced speech. The main challenge in using an LF voice model lies in its accurate parameter estimation. The instant of maximum glottal closure, t_e is made to coincide with the pitch location. For the remaining five of LF parameters we have proposed a two step optimization algorithm which finds the best fit of LF voice source pulses to AR residuals. In the first stage, initial estimates are found using a Minimum Finder Algorithm (MFA) proposed by Brent[61]. The initial estimates are then used to compute final estimates using the grid search algorithm (GSA) via coordinate descent. We have obtained very promising results with the LF model based AR model for voiced speech. In comparing white noise-driven, impulsive and LF model based AR model, the LF model based enhancement gave the best results. Among the voiced speech groups, the front vowels showed the average highest improvements (close to 2 dB) in output SNRs over the conventional white noise excited AR model.

One very important point worth mentioning is that AR-LF residuals are non-white i.e. exhibit periodicity to some extent. This may be partly due to the fact that LF parameter optimization is carried out over the entire speech duration which may result in sampling of the residual spikes at the wrong instances. Hence LF parameter optimization over a single pitch period at a time would alleviate the presence of deterministic spikes in the AR residuals at the cost of increased computational complexity. Another reason behind the presence of spikes in the AR residuals may be due to the effects of secondary excitations after the glottal closure[62, 63]. These facts present various directions for future research.

4.6 Appendix B: Details of Enhancement Results

Phone	Phone context	SNR in dB white noise AR model	SNR in dB impulsive AR model	SNR in dB Lf model based AR model
/iy/	*iy*32*si*1386_*the_e*	10.399	10.576	10.946
/iy/	*iy*16*sx*378_*the_e*	8.174	8.639	8.907
/ih/	*ih*49*si*1386_*unit_i*	9.216	10.078	10.680
/ix/	*ix*17*si*1386_*negotiation_io*	8.589	9.137	9.550
/eh/	*eh*9*si*682_*they_e*	8.542	9.241	9.530
/eh/	*eh*42*si*682_*area_a*1	8.713	9.917	10.443
/eh/	*eh*58*si*682_*economically_e*	7.635	8.287	8.659
/ae/	*ae*26*sx*96_*imagination_a*1	8.062	9.310	10.042

Table 4.2: Enhancement results for the Front Vowels for input SNR of 5 dB and lpc order of 10.

Phone	Phone context	SNR in dB white noise AR model	SNR in dB impulsive AR model	SNR in dB Lf model based AR model
/ah/	$ah12sx119\_was\_a$	9.791	10.729	11.062
/ah/	$ah38sx378\_museum\_u2$	8.259	8.811	9.034
/ah/	$ah24sx86\_colorful\_o1$	9.142	9.587	9.760
/ah/	$ah13si1051\_are\_ah$	8.929	9.850	10.534
/ax/	$ax29sx119\_apology\_o1$	8.917	9.400	9.937
/ax/	$ax6si682\_often\_e$	8.938	9.700	10.249
/er/	$er34sx396\_surface\_ur$	9.278	9.663	9.905

Table 4.3: Enhancement results for the Mid Vowels for the input SNR of 5 dB and the lpc order of 10.

Phone	Phone context	SNR in dB white noise AR model	SNR in dB impulsive AR model	SNR in dB Lf model based AR model
/uh/	*uh*28*sx*86_*colorful_u*	9.833	10.096	10.342
/ux/	*ux*14*sx*86_*costume_u*	8.172	8.735	9.904
/ux/	*ux*15*sa*1_*suit_o*	7.938	8.419	9.899
/ow/	*ow*28*si*682_*only_o*	10.497	11.218	11.679
/ow/	*ow*9*sx*119_*misquote_uo*	9.791	10.729	11.513
/ao/	*ao*29*sx*396_*Of_o*	9.478	10.043	11.570
/ao/	*ao*25*si*1051_*supporters_o*	9.618	10.105	10.508
/ao/	*ao*2*si*1051_*northern_o*	8.970	9.501	10.215
/aa/	*aa*10*sx*96_*parties_a*	8.398	9.236	9.651
/aa/	*aa*35*si*1386_*bargain_a*1	9.121	10.069	11.460
/aa/	*aa*13*sx*210_*cart_a*	9.402	10.039	10.307

Table 4.4: Enhancement results for the Back Vowels for the input SNR of 5 dB and the lpc order of 10.

Phone	Phone context	SNR in dB white noise AR model	SNR in dB impulsive AR model	SNR in dB Lf model based AR model
/yu/	*yu*18*sx*378_*new_ew*	9.106	9.694	10.035
/ey/	*ey*7*sx*96_*masquerade_a*2	8.106	9.143	9.584
/ey/	*ey*12*si*682_*able*$_a$	9.612	10.643	10.980
/ay/	*ay*53*si*682_*declining*$_i$1	9.136	10.070	10.426
/ay/	*ay*6*si*1739_*time*6_*i*	8.781	9.590	10.560
/oy/	*oy*21*sa*2_*oily_oi*	9.376	10.105	10.413
/oy/	*oy*7*sx*196_*oysters_oy*	10.696	11.077	11.482
/oy/	*oy*7*sx*210_*toy_oy*	9.850	10.270	10.440

Table 4.5: Enhancement results for the Diphthongs for the input SNR of 5 dB and the lpc order of 10.

Phone	Phone context	SNR in dB white noise AR model	SNR in dB impulsive AR model	SNR in dB Lf model based AR model
/y/	$y7sa1_{y}our_{y}$	8.293	8.515	8.714
/w/	$w51si1386\_with\_w$	9.967	10.614	11.247
/w/	$w20sx96\_ones\_o$	9.781	10.680	11.086
/r/	$r10si682\_are\_r$	9.101	9.105	10.385
/r/	$r11sx96\_parties\_r$	8.398	9.376	9.774
/l/	$l26sx378\_archeological\_l1$	8.941	9.707	10.580
/el/	$el33sx378\_archeological\_l2$	9.368	9.435	9.728

Table 4.6: Enhancement results for the Semivowels for the input SNR of 5 dB and the lpc order of 10.

Phone	Phone context	SNR in dB white noise AR model	SNR in dB impulsive AR model	SNR in dB Lf model based AR model
/m/	$m11sx378\_jim\_m$	8.885	9.409	9.618
/m/	$m64si682\_economically\_m$	8.716	9.457	9.844
/m/	$m44sx396\_small\_m$	9.010	9.079	9.867
/n/	$n22sx96\_imagination\_n$	9.983	10.271	10.632
/n/	$n30sx396\_on\_n$	9.202	9.688	9.946
/ng/	$ng56si682\_declining\_ng$	6.868	7.855	8.001

Table 4.7: Enhancement results for the Nasals for the input SNR of 5 dB and the lpc order of 10.

Chapter 5

Contributions and Future Research

This chapter reviews the contributions of this thesis and discusses possible avenues for future research.

5.1 Thesis Contributions

The main objectives of this thesis were to find an appropriate model for representing speech for enhancement purposes and to establish the limits to performances for enhancement systems using such a model. The main concentration of this thesis has been on modifying the white noise driven AR model which does not include the effects of the excitation source especially in the case of the voiced speech. AR model based Kalman filter has been used to estimate de-noised speech from noisy speech.

Chapter 3 demonstrated applicability of impulsive AR models for the voiced speech the stops and affricates. Impulsive AR models include deterministic impulsive driving terms which are tentative models for the effects of the excitation. The effects of the glottal excitation is simulated by a train of impulses separated by pitch periods. The unvoiced stops and the unvoiced affricates have plosive excitation modeled by a single impulse at the onset of the burst and white noise. Excitation for the voiced stops and the voiced affricates is modeled by both an impulsive train time spaced by pitch periods, a single impulse at the onset of the burst and white noise. Impulsive AR models always yielded higher output SNRs compared to that for white noise excited AR model. This chapter also discusses the flaws of impulsive models thereby motivating need for more sophisticated model for source excitation. The properties, of an AR model driven by impulses, are summarized as following:

- IMPULSE EXCITED AR MODELS
  - Simple tentative models for excitation sources.
  - Linear parameter estimation.
  - Outperform the white noise excited counterparts.
  - Residuals show considerable periodicity.

Chapter 4 contributed a deeper understanding of the modeling of the voice source excitation for voiced speech enhancement. This chapter establishes the feasibility of LF models for voice source in AR models for speech enhancement. The main challenge for using an LF model is the parameter estimation problem. An optimization algorithm, which finds the best fit for the LF pulse sequence with AR residuals, was proposed. This algorithm computes initial estimates using a minimum finder

algorithm via coordinate descent. The initial estimates are then used by a grid search algorithm to give final estimates via coordinate descent. Proposed Lf model based AR model and optimization algorithm was used for enhancing noisy voiced speech phones. An extensive comparative study, of conventional AR model driven by white noise with impulsive and LF model based AR models, infers that AR model excited by LF model outperforms its counterparts. The characteristics of an LF model based AR model for speech enhancement are summarized as following:

- AN LF MODEL BASED AR MODEL
  - More sophisticated model for voice source.
  - Outperform the both white noise and impulse excited models.
  - Nonlinear parameter estimation.

5.2 Future Research

Some of the interesting directions for future research are listed in the following subsections.

5.2.1 Parameter Estimation from Noisy Speech

As one our objectives has been to study the limits to performance for the Kalman filter based enhancement, we have used clean speech to estimate the AR parameters, process noise and measurement noise covariances. This assumption was necessary to as optimum Kalman filtering requires the accurate knowledge of the noise covariances and AR parameters. In reality often is the case when only the noisy speech is available for processing. A number of methods have been proposed for identifying

the noise covariances from the noisy speech[64, 65, 66]. Various methods utilizing EM algorithm have also been used for estimating AR parameters and the noise covariances from the noisy speech[67, 68, 34]. One useful extension of our work would be to estimate the impulsive and LF model based Kalman filter parameters from the noisy speech using existing methods.

5.2.2 Parameter Estimation for LF Model

LF parameter estimation problem has been addressed long since but in speech synthesis perspective[39, 69, 70]. We have solved the parameter estimation problem for LF model using a coordinate optimization algorithm. Our algorithm was found to be sensitive to the upper and lower bounds for each parameter which are required to be specified for MFA. Alternative estimation procedures for enhancement can be realized using available optimization algorithms[71, 72, 73]. One possible alternative method can be developed using steepest descent algorithm that can optimize in multiple dimension.

5.2.3 Automated Pitch Detection

Manual pitch detection was necessary for studying the limits to performance of the Kalman filter. Pitch detection was done manually from the residual signal (3.2) in which the pulses are conspicuous, followed by an automated local peak-finder to guarantee accurate positioning. In order to apply the proposed models to robust continuous speech enhancement it is necessary to automate the pitch detection process. Pitch detection problem has been well studied in speech analysis[74]. Using one of the available pitch detection algorithms may open up a window of opportunities for our proposed models in real life speech enhancement applications.

5.2.4 Various Types of Measurement Noise

We have used the assumption that clean speech is corrupted by additive white noise as given by 2.1. The measurement white noise used for our experiments has been artificially simulated. It may be a good challenge to use noise encountered in real life which may be white, colored or non-stationary for the model given by 2.1.

5.2.5 Subjective Measure of Enhanced Speech

For evaluating enhanced speech we have used output SNR as objective measure and while as subjective measure we have inspected the temporal plots of clean, noisy, enhanced speech and AR residuals. It remains to be determined how much improvement has been made when hearing is used as subjective measure.

5.2.6 Further Investigation of the Driving Term

One problem with Impulsive and LF model based enhancement is that we require *a priori* knowledge of the pitch period. In the case of voiced fricatives, voiced stops or voiced affricate sometime it was very difficult to identify the pitch periods from AR residuals as such speech types do not show marked periodicity due to pole-zero cancelation unlike the vowels, semivowels, diphthongs or nasals. This discrepancy leaves a vast room for investigating production mechanism of the source and modeling of voiced consonants for enhancement.

In this thesis for voiced speech we have assumed glottal excitation occurring at the glottal closure[39]. The LF model based AR residuals show the presence a number of quasi-periodic negative and positive spikes. Thus even when the speech is clearly periodic it may be too simplistic to assume only one form of driving term

in an entire pitch period[63]. In fact there is some evidence in speech synthesis that apart from the main excitation at the glottal closure there may be secondary excitations after the glottal closure and at the glottal opening at the opening phase[62]. Such facts present good motivations for introducing multiple excitations during within a single pitch period for voiced speech.

Another interesting extension of our work would to derive the excitation waveform directly from the speech waveform. This has been done in a number of speech synthesis and analysis literature in order to produce natural sounding speech [63, 75, 76, 77, 78, 58, 79]. Such methods may rectify the modeling errors introduced by the Impulsive or the LF models.

Bibliography

[1] G.Fant, J. Liljencrants, and Q. Lin. A Four Parameter model of Glottal Flow. *STL-QPSR 4, KTH*, pages 1–12, 1985.

[2] Y.Ephraim. Statistical-Model-Based Speech Enhancement Systems. *Proceedings of the IEEE*, 80(10):1526–1555, Oct. 1992.

[3] G.Fant. *Acoustic Theory of Speech Production.* Mouton's Co., Hague, 1960.

[4] B.S.Atal, V.Cuperman, and A.Gersho. *Advances in Speech Coding.* Kluwer Academic Publishers, 1991.

[5] J.D. Gibson, B. Koo, and S.D. Gray. Filtering of Colored Noise for Speech Enhancement and Coding. *IEEE Transactions on Audio Signal Processing*, 39(8):1732–1741, Aug 1991.

[6] S.V.Vaseghi and B.P. Milner. Noise Compensation Methods for Hidden Markov Model Speech Recognition in Adverse Environments. *IEEE Transactions on Speech and Audio Signal Processing*, 5(1):11–21, Jan. 1997.

[7] B.T. Logan and A.J. Robinson. Enhancement and Recognition of Noisy Speech Within An Autoregressive Hidden Markov Model Framework Using Noise Estimates from the Noisy Signal. *Proceedings of IEEE ICASSP*, 2:843–846, 1997.

[8] H.Sameti. *Model-Based Approaches to Speech Enhancement: Stationary State and Nonstationary-State HMMs*. PhD thesis, University of Waterloo, Department of Electrical Engineering, 1994.

[9] J.S.Lim. *Speech Enhancement*. Prentice-Hall Inc., 1983.

[10] J.S.Lim. Evaluation of a Correlation Subtraction Method for Enhancing Speech Degraded by Additive White Noise. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(5):471–472, Oct. 1978.

[11] S.F.Boll. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 2(27):113–120, April 1979.

[12] J.S.Lim and A.V.Oppenheim. Enhancement and Bandwidth Compression of Noisy Speech. *Proceedings of IEEE*, 67(12):1586–1604, Dec 1979.

[13] R.J.McAulay and M.L.Malpass. Speech Enhancement Using a Soft-decision Noise Suppression Filter. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28:137–145, April 1980.

[14] P.Scalart and J. Vieira Filho. Speech Enhancement Based on a Priori Signal to Noise Estimation. *Proceedings of IEEE ICASSP*, 2:629–632, 1996.

[15] Y.Ephraim and H.L.V.Trees. A Signal Subspace Approach for Speech Enhancement. *IEEE Transactions on Speech and Audio Signal Processing*, 3(4):251–266, July 1995.

[16] B.S. Atal and S.L. Hanauer. Speech Analysis and Synthesis by Linear Prediction of the Speech Wave. *The Journal of the Acoustical Society of America*, 50(2(Part 2)):637–655, 1971.

[17] John Makhoul. Linear Prediction: A Tutorial Review. *Proc.IEEE*, 63:561–580, April 1975.

[18] John D. Markel and A.H. Gray. On Autocorrelation Equations as Applied to Speech Analysis. *IEEE Transactions on Audio and Electroacoustics*, AU-21(2):69–79, April 1973.

[19] J.S.Lim and A.V.Oppenheim. All-Pole Modeling of Degraded Speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(3):June,197–210, 1978.

[20] J.H.L.Hansen and M.A.Clements. Contrained Iterative Speech Enhancement with Application to Speech Recognition. *IEEE Transactions on Signal Processing*, 39(4):795–805, April 1991.

[21] L.R.Rabiner. Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–285, Feb. 1989.

[22] A.B.Poritz. Hidden Markov Models: A Guided Tour. *Proceedings of IEEE ICASSP*, 1:7–13, 1988.

[23] Y.Ephraim, D.Malah, and B.Juang. On the Application of Hidden Markov Models for Enhancing Noisy Speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(12):1849–1856, Dec. 1989.

[24] Y.Ephraim. A Minimum Mean Square Error Approach for Speech Enhancement. *Proceedings of the IEEE ICASSP*, pages 829–832, 1990.

[25] Y.Ephraim. On Minimum Mean Square Error Speech Enhancement. *Proceedings of the IEEE ICASSP*, pages 997–1000, 1991.

[26] H.Sameti, H.Sheikhzadeh, L.Deng, and R.L.Brennan. HMM-based Strategies for Enhancement of Speech Embedded in Non-Stationary Noise. *IEEE Transactions on Speech and Audio Signal Processing*, 6(5):445–455, Setp. 1998.

[27] L.Deng, M.Aksmanovic, X. Sun, and C.F.J.Wu. Speech Recognition Using Hidden Markov Models with Polynomial Regression Functions as Nonstationary States. *IEEE Transactions on Speech and Audio Signal Processing*, 2(4):507–520, Oct. 1994.

[28] L.Deng. Integrated Optimization of Dynamic Feature Parameters for Hidden Markov Modeling of Speech. *IEEE Signal Processing Letters*, 1(4), April 1994.

[29] C.W. Seymour and M Niranjan. An HMM-based Cepstral-Domain Speech Enhancement System. *Proceedings of International Conference on Spoken Language Processing*, 3:1595–1598, 1994.

[30] K.K. Paliwal and A. Basu. A Speech Enhancement Method Based on Kalman Filtering. *Proceedings of IEEE ICASSP*, pages 177–180, 1987.

[31] P.Sörqvist and P.Händel. Kalman Filtering for Low Distortion Speech Enhancement in Mobile Communication. *Proceedings of the IEEE ICASSP*, 2:1219–1222, 1997.

[32] S.Gannot, D.Burshtein, and E.Weinstein. Iterative-Batch and Sequential Algorithms for Single Microphone Speech Enhancement. *Proceedings of the IEEE ICASSP*, 2:1215–1218, 1997.

[33] Byung-Gook Lee and K. Y. Lee. An EM-based Approach for Paramater Enhancement with an Application to Speech Signals. *Signal Processing*, 46(1):1–14, Sept. 1995.

[34] W.Du and P.Driessen. Speech Enhancement Based on Kalman Filtering and EM Algorithm. *Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 142–145, May 1991.

[35] K.Y.Lee and K.Shirai. Efficient Recursive Estimation for Speech Enhancement in Colored Noise. *IEEE Signal Processing Letters*, 3(7):196–199, July 1996.

[36] G.Fant. Glottal Source and Excitation Analysis. *STL-QPSR-1, KTH*, pages 70–85, 1979.

[37] G.Fant. Voice Source Analysis - A Progress Report. *STL-QPSR 3-4, KTH*, pages 31–54, 1979.

[38] G.Fant. The Voice Source Analysis - Acoustic Modeling. *STL-QPSR 4, KTH*, pages 28–47, 1982.

[39] D.G.Childers and C.K.Lee. Vocal Quality Factors: Analysis, Synthesis and Perception. *Journal of Acoustical Society of America*, 90(5):2394–2410, 1991.

[40] A.E. Rosenberg. Effect of Glottal Pulse Shape on the Quality of Natural Vowels. *Journal of Acoustical Society of America*, 49(2(part 2)):583–590, 1971.

[41] T.V. Ananthapadmanabha and G Fant. Calculation of True Glottal Flow and Its Components. *Speech Communication*, 1(3-4):167–184, Dec. 1982.

[42] K.W.Hipel and A.I.McLeod. *Time Series Modelling of Water Resources and Environmental Studies*. Elsevier, Amsterdam, The Netherland, 1992.

[43] J.R.Deller, J.G.Proakis, and J.H.L.Hansen. *Discrete-time Processing of Speech Signals*. Macmillan Publishing Company, NY, 1993.

[44] S.V.Vaseghi. *Advanced Signal Processing and Digital Noise Reduction.* Wiley and Teubner, 1996.

[45] S.Haykin. *Adaptive Filter Theory.* Prentice Hall Information and System Sciences, second edition, 1991.

[46] Sheldon M.Ross. *Introduction to Probability Models.* Academic Press, 1989.

[47] B.Juang and L.R.Rabiner. Mixture Autoregressive Hidden Markov Models for Speech Signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-38(9):1639–1641, Sept. 1990.

[48] A.Buzo, A.H.Gray, and J.D.Markel. Speech Coding Based on Vector Quantization. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28(5):562–574, Oct. 1980.

[49] H.Sameti. Segmental k-Means for Mixture Hidden Markov Models. Project Report, University of Waterloo, Dept. of Electrical Engineering, Aug. 1992.

[50] Jerry M. Mendel. *Lessons in Estimation Theory for Signal Processing, Communications, and Control.* Prentice Hall Inc, 1995.

[51] G.C.Goodwin and K.S.Sin. *Adaptive Filtering Prediction and Control.* Prentice-Hall Information and System Sciences Series, 1984.

[52] Thomas Khailath. A View of Three Decades of Linear Filtering Theory. *IEEE Transactions on Information Theory*, IT-20(2):146–181, March 1974.

[53] F.L.Lewis. *Optimal Estimation.* Wiley-Interscience, 1986.

[54] R.D. Kent and C. Read. *The Acoustic Analysis of Speech.* Singular Publishing Group,Inc, 1992.

[55] L.R.Rabiner and R.W.Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[56] National Institute of Standards and Technology (NIST). *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Oct. 1990.

[57] G.Fant. The Source Filter Concept in Voice Production. *STL-QPSR 1, KTH*, pages 21–37, 1981.

[58] D.Y.Wang, J.D.Markel, and A.H.Gray. Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(4):350–355, Aug. 1979.

[59] T.V. Ananthapadmanabha. Acoustic Analysis of Voice Source Dynamics. *STL-QPSR 2-3, KTH*, pages 1–24, 1984.

[60] I.R.Titze. A Four Parameter Model of the Glottis and Vocal Fold Contact Area. *Speech Communication*, 8(3):191–201, Sept. 1989.

[61] G.E.Forsythe, M.A.Malcolm, and C.B.Moler. *Computer Methods for Mathematical Computations*. Prentice-Hall, 1977.

[62] J.H.Holmes. Formant Excitation Before and After Glottal Closure. *Proccedings of IEEE ICASSP*, pages 39–42, 1976.

[63] B.S.Atal and J.R.Remde. A New Model of LPC Excitation for Producing Natural-sounding Speech at Low Bit Rates. *Proccedings of IEEE ICASSP*, 2:614–617, 1980.

[64] T.T.Lee. A Direct Approach to Identify the Noise Covariances of Kalman Filter. *IEEE Transactions on Automatic Control*, 25(4):841–842, Aug. 1980.

[65] R.K.Mehra. On the Identification of Variances and Adaptive Kalman Filtering. *IEEE Transactions on Automatic Control*, 15:175–184, Aug. 1970.

[66] S.S.Godbole. Kalman Filtering with No a priori Information about Noise-White Noise Case: Identification of Covariances. *IEEE Transactions on Automatic Control*, 19:561–563, Oct. 1974.

[67] R.H.Shumway and D.S. Stoffer. An Approach to Time Series Smoothing and Forecasting Using the EM Algorithm. *Journal of Time Series Analysis*, 3(4):253–263, 1982.

[68] V.Digalakis, J.R.Rohlicek, and M.Ostendorf. ML Estimation of a Stochastic Linear System with the EM Algorithm and its Application to Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 1(4):431–442, Oct. 1993.

[69] W.Ding and H.Kasuya. Simultaneous Estimation of Vocal Tract and Voice Source Parameters with Application to Speech Synthesis. *International Conference on Spoken Language Processing*, 1:159–162, 94.

[70] H.Strik, B.Cranen, and L.Boves. Fitting an LF-model to Inverse Filter Signals. *Eurospeech*, 1:103–106, 93.

[71] D.Jacobs. *The State of the Art in Numerical Analysis.* Academic Press INC, 1976.

[72] A.V.Ballakrishnan. *Techniques of Optimization.* 1971, Academic Press INC.

[73] E.J. Beltrami. *An Algorithmic Approach for Nonlinear Analysis and Optimization.* Academic Press INC, 1970.

[74] W.Hess. *Pitch Determination of Speech Signals.* Springer-Verlag, 1983.

[75] H.Wakita. Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms. *IEEE Transactions on Audio and Electroacoustics*, 21(5):417–427, Oct. 1973.

[76] P.Rubin, T.Baer, and P.Mermelstein. An Articulatory Synnthesizer for Perceptual Research. *Journal of Acoustical Society of America*, 70(2):321–328, Aug. 1981.

[77] W.Zhu and H.Kasuya. A New Speech Synthesis System Based on the ARX Speech Production Model. *International Conference on Spoken Language Processing*, pages 1413–1416, 1996.

[78] D.H.Klatt. Software for a Cascade-Parallel Formant Synthesizer. *Journal of Acoustical Society of America*, 67(3):971–995, March 1980.

[79] P.Alku and E.Vilkman. Estimation of the Glottal Pulseform Based on Discrete All-pole Modeling. *International Conference on Spoken Language Processing*, 3:S27:10.1–4, 1994.