THE PITCH MODE MODULATION MODEL WITH APPLICATIONS IN SPEECH PROCESSING

BY MICHAEL A. RAMALHO

A dissertation submitted to the
Graduate School—New Brunswick
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
Graduate Program in Electrical Engineering

Written under the direction of
Professor Richard J. Mammone
and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 1994

© 1994
Michael A. Ramalho
ALL RIGHTS RESERVED

ABSTRACT OF THE DISSERTATION

THE PITCH MODE MODULATION MODEL WITH APPLICATIONS IN SPEECH PROCESSING

by Michael A. Ramalho

Dissertation Director: Professor Richard J. Mammone

The techniques currently used for the enhancement of speech critically depend upon some form of statistical stationarity, either in the speech signal, the noise signal, or both, in order to accomplish the enhancement. Virtually all speech processing techniques utilize a speech model to reduce the amount of information necessary to characterize the speech signal. Although the speech signal is known to be highly redundant, it is also non-stationary. This non-stationarity requires that the parameters of these models be extracted from short duration signal segments, where the stationarity assumption in the models is not seriously violated. Unfortunately, the use of short speech frames makes the estimation of the model parameters difficult and sometimes obscures the very redundancy the model was based on. The use of a longer frame size is desirable for many signal processing techniques that require increased frequency domain resolution.

The Pitch Mode Modulation Model (PMMM) developed in this dissertation is a *long-term* speech model which accommodates non-stationarity in normal speech and, therefore, can process longer speech segments. The PMMM processing will be seen to *non-linearly* time warp the speech signal such that all quasi-harmonic speech components become periodic. The PMMM processing is capable of associating speech signal components in the presence of arbitrary noise, thereby focusing the processing on known speech energy. The PMMM processing does not require *a priori* pitch information; instead its parameters, including a robust pitch estimate, are estimated accurately in the presence of noise.

The primary application of the PMMM is speech enhancement for Speaker Identification (SI)

systems. Computer simulations are presented which demonstrate that the PMMM based enhancement significantly reduces the sensitivity of the SI system to mismatches between training and testing data.

Other applications of the PMMM, such as time-scale and pitch modification, co-channel speaker/sound separation, harmonic coding, and super-resolution pitch estimation, are also discussed. Comparison of the PMMM to other speech models, including those that attempt to model speech non-stationarity, is performed. In addition, the form of the PMMM speech components is also shown to be in direct agreement with recent auditory and cochlear modeling research.

Preface

In many speech communications settings, the presence of background noise causes the quality or intelligibility of speech to degrade. The human auditory system is able to tolerate (i.e., neurologically enhance) speech degraded by noise sources much better than present day speech enhancement systems. Early work in speech enhancement recognized that the auditory system appears to rely heavily on binaural processing of stereophonic data for this task. Recent auditory research suggests that the grouping of sounds may involve neurological integration of these sounds for 100 to 200 ms, much longer than present speech processing averaging techniques. Unfortunately, the precise manner in which this binaural processing is done is not as well understood as the speech production mechanics. Speech enhancement techniques, therefore, have focused on known speech models to perform the enhancement. In addition, since the majority of applications that require speech enhancement only have one input source, most research has considered only the single-input case.

The speech models that these enhancement techniques depend upon use some form of statistical stationarity, either in the speech signal, the noise signal, or both, in order to accomplish the enhancement. Although the speech signal is known to be highly redundant, it is also non-stationary. This non-stationarity requires that the parameters of these models be extracted from short signal segments (typically 10 to 30 ms), where the stationarity assumption in the models is not seriously violated. Unfortunately, the use of short speech frames makes the estimation of the model parameters difficult and sometimes obscures the very redundancy the model was based on. This is particularly true for noisy speech. Additionally, because of the highly disparate statistical nature of noise signals and the non-stationarity of speech signals, most enhancement techniques are *short-term* techniques that focus on only a specific form of noise impairment.

What is needed are *long-term* speech modeling techniques that accommodate the non-stationarity found in normal voiced and mixed voice speech. Additionally, since noise sources are statistically varied and usually non-stationary, it would be beneficial for the techniques to focus their processing energy on the redundancy in speech. Such techniques should be able to associate known speech components, due to their redundancy, and disassociate these components

from noise.

The Pitch Mode Modulation Model (PMMM) is a *long-term* speech model which accommodates non-stationarity in normal speech and, therefore, can process longer speech segments. The PMMM processing will be seen to *non-linearly* time warp the speech signal such that all quasi-harmonic speech components become periodic. The PMMM processing is capable of associating speech signal components in the presence of arbitrary noise, thereby focusing the processing on known speech energy.

Chapter 1 of this thesis describes classical speech models, the nature of the speech and noise signals, and the prior art in speech enhancement research as necessary prolegomena to the introduction of the PMMM in Chapter 2. In Chapter 2 the PMMM is shown to be a generalization of the Fourier Series representation for quasi-periodic signals and an example of its ability to model long speech frames is presented.

In Chapter 3 the PMMM is compared and contrasted to other relevant speech models and its uniqueness to model speech non-stationarity is described. Chapter 4 presents PMMM ties to recent auditory fusion and cochlea modeling and shows that the PMMM speech parameterization is in direct agreement with these findings.

In Chapter 5 the primary applications of speech enhancement for speaker identification and PMMM based time-scale and pitch modification of speech are presented. In Chapter 6 other potential applications for the PMMM, including low bit rate speech coding, are presented.

Chapter 7 presents a summary of the innovations of this research and proposes areas of continuing research.

Acknowledgements

My pre-eminent acknowledgment is for my advisor, Dr. Richard J. Mammone, whose belief in my abilities from the day he signed my Ph.D. re-admission form to the present, has been an inspiration for me. His oversight and sincere care for *all* his graduate students is to be admired. Thank you.

A significant debt of gratitude is due to the other members of my Thesis Committee for their efforts in guiding my research and in the review of this manuscript. These people are: Dr. James L. Flanagan, Dr. Joseph Wilder and Dr. Daniel Kahn.

I would like to acknowledge my employer, Bell Communications Research, Inc., for allowing me to pursue this work under their Support for Doctoral Education program. Specifically, I thank the following people whose personal effort made it possible for me to accomplish this work: Dr. Gary J. Handler, Mr. Richard T. Aloia, Mr. Michael A. Schwartz, Dr. Louis J. Scerbo, Dr. James D. McKenna, Dr. Daniel Kahn, Dr. Charles N. Judice, Dr. L. S. (Stagg) Newman and Mr. David S. Burpee.

I gratefully acknowledge Rome Laboratories, Contract No. F30602-91-C-0120, for their support of this work. Additionally, this research was made possible, in part, through the support of the New Jersey Commission on Science and Technology and the Computer Aids for Industrial Productivity (CAIP) Center at Rutgers University.

Dedication

To my wife, Diana, and my daughters, Jessica and Ashley, all of whom have sacrificed of themselves so that I may complete this endeavor. I consider my Ph.D. accomplishments as much theirs as mine.

I also dedicate this work to my parents, James and Mary Ramalho, who taught me diligence and the value of honest work.

Table of Contents

Abstract ii

Preface iv

Acknowledgements vi

Dedication vii

List of Figures ix

List of Abbreviations x

1 Introduction 1

1.1 Problem Description 1

1.2 Classical Speech Models 2

1.3 Nature of Speech Signals and their Spectra 4

1.4 Nature of Interfering Sounds 5

1.5 The Problem of Speech Enhancement 6

1.5.1 Short-Term Spectral Domain Enhancement 7

1.5.2 Speech Modeling and Wiener Filtering Based Enhancement 8

1.5.2.1 Iterative Wiener Filtering 9

1.5.2.2 Enhancement via All-Pole Modeling 10

1.5.2.3 Constrained Iterative Enhancement 13

1.5.3 Adaptive Noise Canceling Enhancement 14

1.5.4 Fundamental Frequency Tracking Based Enhancement 16

1.5.4.1 Single-Channel ANC 16

1.5.4.2 Adaptive Comb Filtering 17

1.5.4.3 Harmonic Selection 19

1.6 Need for Long-Term Model for Speech 20

1.7 Shortcomings of Present Speech Enhancement and Speech Modeling 21

1.8 Thesis Outline 22

2 The Pitch Mode Modulation Model 24
2.1 Modulation Model Description 24
2.2 Formant Mode Modulation Model 24
2.3 Pitch Mode Modulation Model: Fundamentals 25
2.3.1 Beta Function Extraction Example 29
2.4 Pitch Mode Modulation Model: Detailed Analysis 31
2.4.1 Spectral Implications of $\beta(t)$ Modulation 31
2.4.2 Analysis for Sinusoidal $\beta(t)$ 33
2.4.3 The Goal of PMMM Preprocessing 39
2.5 PMMM Parameter Estimation Algorithm 40
2.5.1 Beta Function Extraction Methodology 40
2.5.2 Time-Warping Function $\beta_i(t)$ Clustering 45
2.5.3 PMMM-Based Fundamental Frequency Estimation 49
2.5.4 Iterative Parameter Estimation 53
2.5.5 Iterative Parameter Estimation Algorithm: Detailed Analysis 54
2.6 PMMM: Ability to Model Long-Term Speech Non-Stationarity 57
2.7 PMMM: Exploiting Redundancy in Voiced Speech 61
2.8 PMMM Processing Gain: A Spread Spectrum Analogy 63
2.9 Summary PMMM Attributes 66

3 PMMM Comparison to Other Speech Models 68
3.1 Vocoders Based on Classical Speech Models 69
3.1.1 Classical Spectral Envelope Analysis Methods 70
3.1.2 Classical Fundamental Frequency Estimation 71
3.1.3 Classical LPC Based Vocoders 73
3.1.3.1 Residual Excited LPC Vocoder 73
3.1.3.2 Multipulse LPC Coders 74
3.1.3.3 Code-Excited Linear Prediction 75
3.2 Spectrum Based Harmonic Coding 76
3.2.1 Sinusoidal/Harmonic Models of McAulay and Quatieri 76
3.2.2 Multi-Band Excitation Vocoder of Griffin and Lim 77
3.3 Non-Stationary Speech Models 78
3.3.1 Non-Stationary Model of Almeida and Tribolet 79

3.3.2 Long-Term Model of Stettiner, et al. 80
3.3.3 DTW-Based Comb Filtering of Graf and Hubing 82
3.4 Summary of PMMM Processing and Modeling Advantages 83

4 PMMM Ties to Recent Speech Perception Research 85
4.1 Auditory Fusion Principles 85
4.2 Cochlear Modeling and Time Synchrony Analysis – Implications for Harmonic Association 87
4.3 PMMM Spread Spectrum Analogy Revisited 90
4.4 Comments on PMMM Ties to Auditory Research 91

5 PMMM Applications and Numerical Results 93
5.1 Speaker Identification in Adverse Environments 93
5.1.1 PMMM Speech Enhancement Technique 94
5.1.2 Typical Speaker Identification System Training Results 95
5.1.3 PMMM-Based Enhancement Speaker Identification Results 97
5.2 PMMM-Based Speech Synthesis 99
5.2.1 Time-Scale and Pitch Modification of Speech 100

6 Other Potential Applications of the PMMM 103
6.1 Co-Channel Speaker or Sound Separation 103
6.2 Low Bit-Rate Harmonic Speech Coding 105
6.3 Robust Technique for "Super-Resolution" Pitch Extraction 107
6.4 Vocal Physiology Abnormality Determination 108

7 Summary, Innovations, and Future Research 110
7.1 The Problem and Solution 110
7.2 Innovations of this Research 111
7.3 Proposed Areas of Continuing Research 112

References 114

Vita 126

List of Figures

1.1 Representations of Speech Signals 3

1.2 General Discrete-Time Speech Model (after Rabiner 1978) 4

1.3 Idealized Spectra for Voiced Speech: Glottal Spectrum $G(z)$, Spectrum of $V(z)R(z)$, and Output Spectrum $H(z)$ 5

1.4 Classification of Speech Enhancement Methods 7

1.5 Generalized Spectral Subtraction Methodology 8

1.6 LPC Based Speech Model 11

1.7 Adaptive Noise Canceling Block Diagram 15

1.8 Single Channel ANC Block Diagram 16

2.1 The Mapping from $\tau = g(t) = [t + \beta(t)]$ 28

2.2 STFT Magnitude of vowel /æ/ in "bat" (4096 samples @ 16kHz sampling) . . . 30

2.3 $\beta_i(t)$ Estimates for $i = 1, 2, ...10$ 31

2.4 Illustrative Fourier Magnitude for $s(t)$ with $\beta(t)$ Modulation 32

2.5 $J_n(\psi)$ as a Function of Modulation Index ψ 36

2.6 Illustrative Fourier Magnitude for $s(t)$ and $s(\tau)$ 39

2.7 First eight pitch harmonics of vowel /æ/ in "bat" 43

2.8 Three beta function estimates for first pitch harmonic 44

2.9 Three beta function estimates for the second pitch harmonic 44

2.10 Three beta function estimates where there is not a pitch harmonic 45

2.11 Pseudocode for $\beta(k)$ Clustering Algorithm 47

2.12 Beta clustering example for 2048 sample segment of "How are you?" 49

2.13 Harmonic Number Association Pseudocode for $\bar{\omega}_i$ 52

2.14 PMMM Iterative Parameter Estimation Block Diagram 54

2.15 Iterative Parameter Estimation Pseudocode 55

2.16 "How are you": Spectra of Original and De-Warped Spectra (De-Warped plotted 40 dB down) 58

2.17 "How are you?": Original (grey) and De-Warped (black) Spectra 59

2.18 PMMM Spread Spectrum Analogy 64
3.1 LPC Based Speech Model 73
3.2 Multipulse LPC Excitation Methodology 75
4.1 Auditory Fusion Principle 86
5.1 Training on Clean Speech (No enhancement) 96
5.2 Training on 25 dB White Noise (No enhancement) 96
5.3 Training on Enhanced Clean Speech 97
5.4 Training on 25 dB White Noise Enhanced Speech 98
5.5 Time Scale Modification of Speech 101

List of Abbreviations

ACF	Adaptive Comb Filter
AM	Amplitude Modulation
AMPEX	Auditory Model-based Pitch EXtractor
AMDF	Average Magnitude Difference Function
ANC	Adaptive Noise Canceling
AS	Analytic Signal
BPF	Band-Pass Filter
bps	bits per second
CELP	Code Excited Linear Prediction
CNS	Central Nervous System
dB	Decibel
DTW	Dynamic Time Warping
EM	Estimate Maximize
FFT	Fast Fourier Transform
FM	Frequency Modulation
FMMM	Formant Mode Modulation Model
Hz	Hertz
ICASSP	IEEE International Conference on Acoustics, Speech, and Signal Processing
IFFT	Inverse Fast Fourier Transform
Kbps	Kilobits per second
kHz	Kilohertz
KLT	Karhunen-Loève Transform
LIN	Lateral Inhibition Networks
LMS	Least Mean Squared Error
LP	Linear Prediction
LPC	Linear Predictive Coding or Linear Predictive Coefficients
LSP	Line Spectral Pairs
LVQ	Learning Vector Quantizer

MAP	Maximum A Posteriori
MBE	Multi-Band Excitation
MLPC	Multipulse LPC
MSE	Mean Squared Error
PCA	Principal Component Analysis
PDS	Power Density Spectrum
PM	Phase Modulation
PMMM	Pitch Mode Modulation Model
RELPC	Residual Excited LPC
SHS	Sub-Harmonic Summation
SI	Speaker Identification
SIFT	Simple Inverse Filtering Tracking
SNR	Signal-to-Noise Ratio
STFT	Short-Term Fourier Transform
STFM	Short-Term Fourier Magnitude
stPDS	short-term Power Density Spectrum
TDHS	Time Domain Harmonic Scaling
V/UV	Voiced/Unvoiced
WSS	Wide-Sense Stationary

Chapter 1

Introduction

1.1 Problem Description

Virtually all applications in speech processing use either an explicit speech model or statistical techniques to exploit redundancy contained in the speech signal. In most speech processing techniques, speech models are used in order to reduce the amount of information necessary to characterize the speech signal. The speech models developed often partition the problem in a physiology-based partitioning of a source/filter model. The source models the vocal tract excitation (i.e., air movement from the lungs) and the filter the combined vocal tract and mouth/lips/nostrils radiation responses.

Unfortunately, although the speech signal is highly redundant, it is also non-stationary. This non-stationarity, in turn, requires that the parameters of these models be extracted from short signal segments, whose short duration make the estimation of the parameters difficult and sometimes obscures the very redundancy the model was based on. This is particularly true for noisy speech.

Likewise, virtually all speech enhancement techniques critically depend upon some form of stationarity, either in the speech signal, the noise signal, or both. Noise types are quite varied in nature; they can be impulsive, continuous, periodic or quasi-periodic, broadband or narrowband, or lack amplitude stationarity. Due to the highly disparate statistical nature of these noise signals and the non-stationarity of speech, most enhancement techniques focus on only a specific form of impairment.

The Pitch Mode Modulation Model (PMMM) is a *long-term* speech model which accommodates non-stationarity in normal speech and, therefore, can be used to process longer speech segments. The PMMM processing will be seen to *non-linearly* time warp the speech signal such that all quasi-harmonic speech components become periodic. The PMMM processing is capable of associating speech signal components in the presence of arbitrary noise, thereby focusing the processing on known speech energy. The PMMM processing does not require *a priori* pitch information and its parameters, including a robust pitch estimate, are estimated accurately in

the presence of noise. Comparison of the PMMM to other speech models, including those that model speech non-stationarity, will be performed. The form of the PMMM speech components will also be seen to be in direct agreement with recent auditory and cochlear modeling research.

The primary application of the PMMM for this work is speech enhancement for Speaker Identification (SI) systems and results using PMMM based speech enhancement for this application are presented. However, as the PMMM is a new speech model, it can be exploited in the same manner as other speech models have been exploited. Other applications, such as time-scale and pitch modification, co-channel speaker/sound separation, harmonic coding, and super-resolution pitch estimation, are also discussed.

In this chapter, prior art in speech modeling and speech enhancement is presented as necessary prolegomena to the introduction of the PMMM in Chapter 2. Also presented in this chapter is a brief discussion of the nature of speech and noise short-term spectra. An outline of the remaining portions of the thesis is presented in Section 1.8.

1.2 Classical Speech Models

Models of speech production have been developed for many purposes. Some examples are the applications to noise reduction/speech enhancement, speaker separation, pitch or time-scale modification of speech, low bit rate speech coding, high-quality speech synthesis, and others. The representation used is often dictated by the application. Most representations can be classified into two broad groups of either waveform or parametric type. Waveform representations, as the name implies, are concerned with preserving, in a certain sense, the "wave shape" or appearance of the signal while parametric representations are concerned with representing the speech signal as the output of an underlying speech model. These representations are illustrated in Figure 1.1. The primary advantage of introducing a speech model is to transform the highly redundant speech waveform to a model in which the parameters have a lower bandwidth.

Speech signals are often modeled as the output of a slowly time-varying linear filter that has been "driven" by an excitation waveform [48, 151, 34]. This model is motivated and based upon the acoustic theory of human sound production, and it is assumed that the parameters of this model will change slowly in time. The excitation for voiced sounds is a quasi-periodic waveform which models the quasi-periodic puffs of air produced by the glottis. The excitation for unvoiced sounds is typically random noise which models the turbulence at a constriction in the vocal tract. The excitation source then drives a filter which models the vocal tract and is modeled by a transfer function $V(z)$. The output of this filter is input to another filter that models the

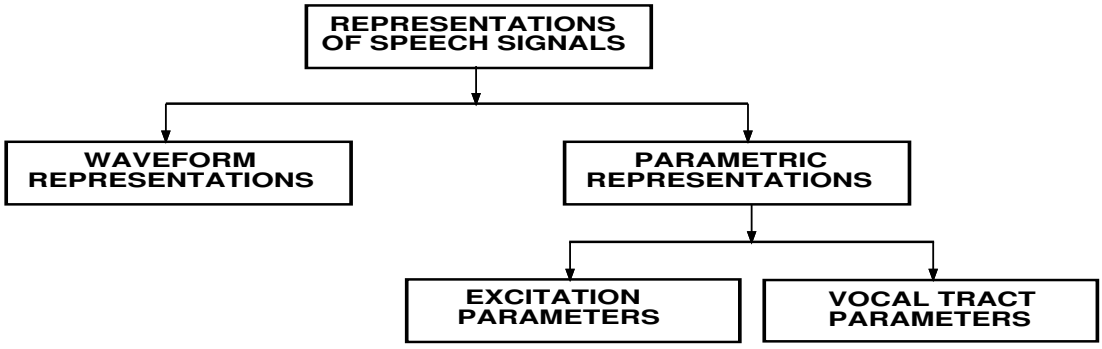

Figure 1.1: Representations of Speech Signals

radiation of the speech out of the mouth and nostrils and is modeled by a transfer function $R(z)$. Owing to the speech production mechanism, the parametric-type speech model usually models the excitation and the vocal tract separately. Classical methods for independently estimating the excitation and vocal tract parameters are discussed and compared to the PMMM later in Chapter 3.

For voiced speech, the excitation waveform can be further modeled as the output of a Glottal Pulse Model, $G(z)$, driven by an impulse train with pitch period, T. A model that incorporates the above attributes is presented in Figure 1.2, where a voiced/unvoiced switch is used to switch the excitation for voiced and unvoiced sounds. A_V and A_N are slowly varying gains for the respective voiced and unvoiced signal inputs to the vocal tract filter.

It should be emphasized that the use of transfer functions implicitly assumes that we can represent the speech signal on a "short-time" basis, whereby the model parameters are assumed to be constant over time intervals typically 10-20 milliseconds long. As a result most speech analysis-synthesis systems assume short term stationarity and use relatively short analysis frames for which this assumption is not seriously violated. For many application areas, the above model is adequate and has been successfully used for many years. This model, and small modifications to it, has formed the basis of most general purpose speech synthesis systems. Vocoders (i.e., "voice coders") designed from this model produce highly intelligible speech. However, although highly-intelligible, the speech produced has a machine-like, unnatural sound quality to it. Further comments on the speech quality produced by such a model are deferred to Section 3.1.3, where we compare it to the Pitch Mode Modulation Model.

This thesis then argues for the need for a long-term model of speech production that the Pitch Mode Modulation Model provides.

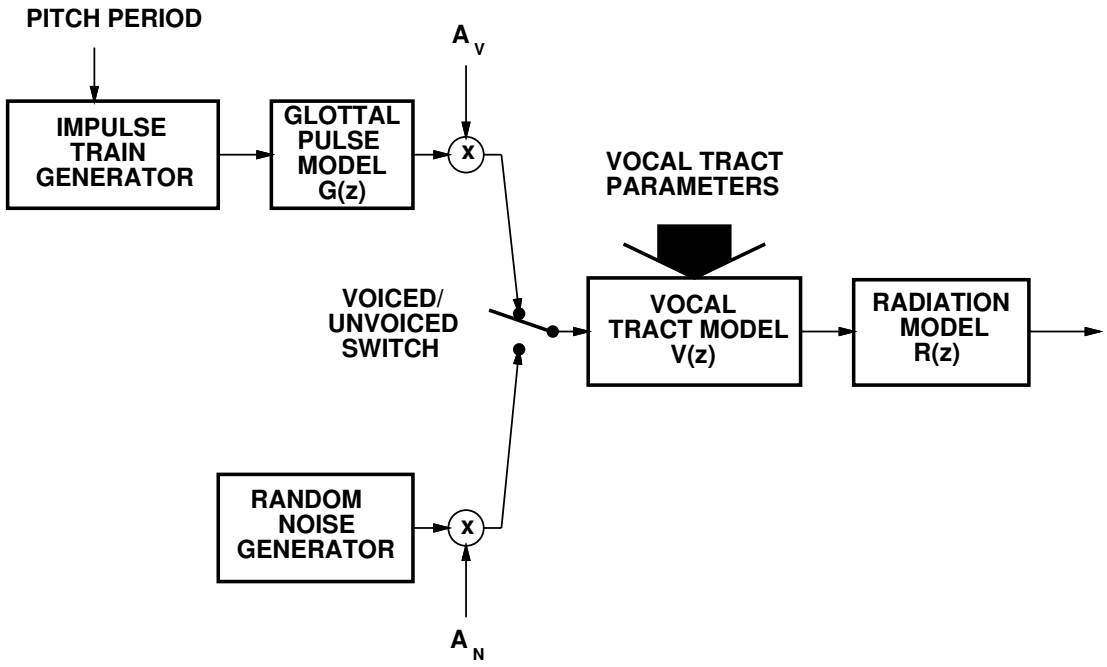

Figure 1.2: General Discrete-Time Speech Model (after Rabiner 1978)

1.3 Nature of Speech Signals and their Spectra

It is instructive, however, to continue discussion of the model of the previous section to obtain insight for the spectra of voiced speech signals. The spectrum output of the glottal filter, $G(z)$, for a perfectly periodic pulse train in time is, in the frequency domain, a perfectly periodic pulse train in frequency, but "shaped" by the glottal filter model $G(z)$. The spectral lines of this spectrum are separated by $\frac{1}{T}$, or the "fundamental" frequency of the input (often denoted as F_0 in speech literature). This spectrum is modified or "shaped" by the vocal tract and mouth radiation responses ($V(z)R(z)$). In other words, the spectrum of the output speech appears as a line spectrum modified by the transfer function $H(z) = G(z)V(z)R(z)$ at the multiples of F_0. This is illustrated in Figure 1.3

Unfortunately, the spectrum of speech seldom appears like this ideal, and the spectral lines due to the harmonics become quite blurred. Several reasons for this are: non-periodicity of the glottal pulses, vocal tract movement and windowing effects. For example, the windowing smears the line spectrum in inverse proportion to the duration of the window [133]. The short-term stationarity assumptions in the aforementioned speech models require the use of short-time windows, obscuring the line spectrum for purely periodic speech from the spectrum alone. The use of short windows also makes the observability of any non-periodicity of the input speech

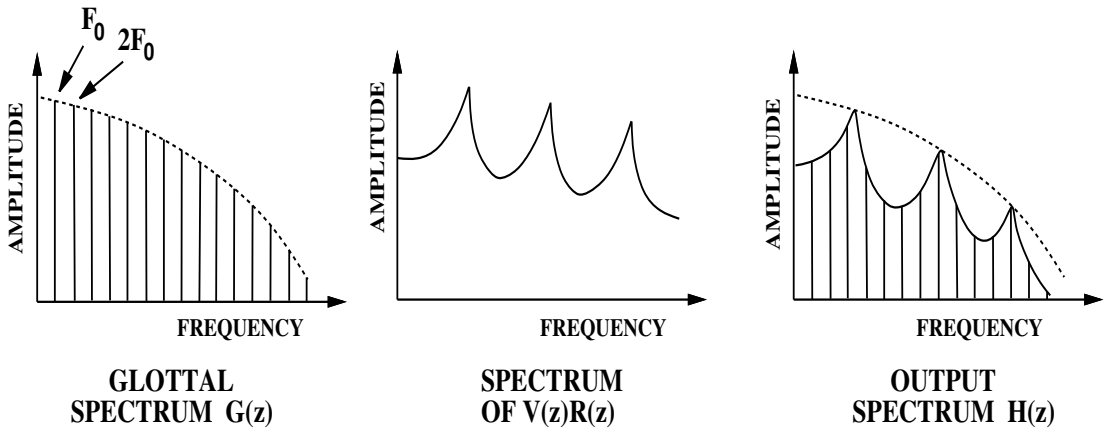

Figure 1.3: Idealized Spectra for Voiced Speech: Glottal Spectrum $G(z)$, Spectrum of $V(z)R(z)$, and Output Spectrum $H(z)$

difficult, if not impossible.

1.4 Nature of Interfering Sounds

The types and nature of interfering sounds are quite problematic to many speech recognition and speaker identification systems [152, 34]. The noise may be impulsive, continuous, periodic, or quasi-periodic. The amplitude (or envelope) of the noise may also be time-varying. A different assumption concerning the noise type and *its* stationarity leads to different suppression techniques to mitigate it [130].

Transmission or other background noise is typically continuous and broadband (e.g. quantization noise) while other noises are abrupt, strong and/or narrowband. Humming noise from machinery or from AC power lines is continuous, stationary and present only at a few frequencies. Humming noise, once identified, can be removed by narrowband filtering.

The speech of an interfering speaker is troublesome in that it occupies the same overall frequency range as that of a desired speaker and is typically as non-stationary as that of the wanted speaker. Most of the time, the two speakers will be speaking with a different fundamental frequency, or pitch. Some enhancement methods attempt to identify the strong frequencies of either the desired speaker or the unwanted source and attempt to separate their spectral components to the extent that the components do not overlap. We later investigate this phenomenon in Sections 1.5.4 and 2.6 and show that, in general, these techniques are ineffective at separating the signals at the higher frequencies due to pitch non-stationarity.

The Pitch Mode Modulation Model circumvents many of the difficulties in identifying the

type of noise impairment by focusing its processing power instead on known speech properties. It is shown that a new formulation for the redundancy contained in quasi-periodic speech waveform is not only useful to associate the wanted signal components, but is also able to identify some common noise types.

1.5 The Problem of Speech Enhancement

In many speech communication settings, the presence of background interference causes the quality or intelligibility of speech to degrade. Simply put, a noisy environment reduces the listener's ability to understand what is being said or to correctly identify the person who is speaking. The purposes of speech enhancement algorithms are to reduce background noise, improve speech quality, or suppress channel or interference from other speakers. Background noise refers to any additive broadband source (e.g. white Gaussian noise, most machinery noise sources, general background noise).

Because the range of possible enhancement applications and the noise types are varied, most speech enhancement techniques focus on a specific application area and subset of noise types. The success of an enhancement algorithm depends on the actual application and specific assumptions concerning the goals. For example, an enhancement algorithm to eliminate tones is effectively useless against noise from a competing speaker. Most speech enhancement techniques focus on reducing the effects of a broadband *stationary* noise source introduced at the source.

Speech enhancement techniques have been reported in the literature dating substantially back to 1976 [92] and can be classified in a number of ways. These are depicted in Figure 1.4. A broad grouping is concerned with the manner in which speech is modeled. The two broad categories here are those which use stochastic models for speech and those which are based on the perception aspects of speech. Systems based on the former category rely on a given mathematical criterion and usually impose stationarity arguments from stochastic process theory. Due to the non-stationarity of speech, these methods rely, almost exclusively, on *short-term* models of speech. Systems based on the second category rely on criteria important to human perception of sounds (e.g. auditory masking or critical band theory [9, 51]).

In addition, one could also classify the methods as those that have a single-channel or a dual-channel (or multichannel) input. It has been acknowledged that human listeners are able to tolerate (i.e., neurologically enhance) speech degraded by arbitrary noise sources better than signal-processing-based enhancement algorithms. Many of the early works recognized that although the human auditory system is adept at resolving the speech of one talker against

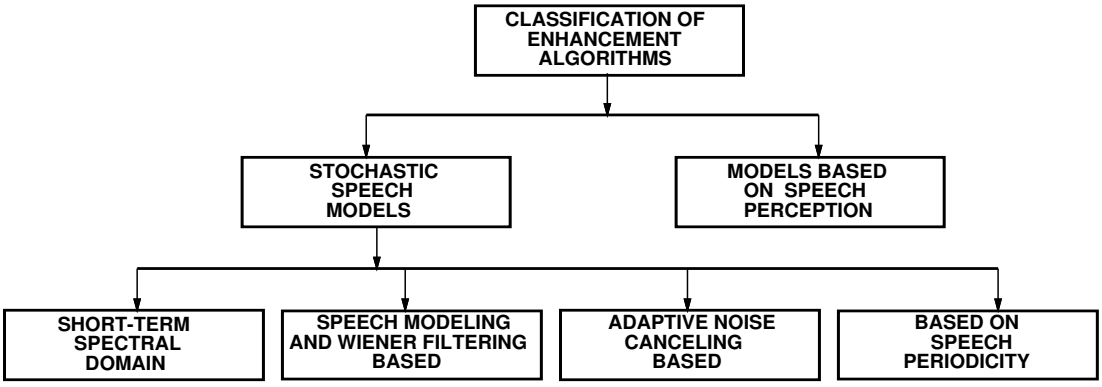

Figure 1.4: Classification of Speech Enhancement Methods

arbitrary background noise (including other talkers), that this task appears to rely heavily on binaural processing of stereophonic data [92, 87, 167, 117]. Unfortunately, the precise manner in which the human auditory system performs this binaural processing is not as well understood as the speech production mechanics. These techniques, therefore, focus on known speech models to perform the enhancement. Also, since the majority of the applications that require speech enhancement only have one input source, most research has only considered the single-input case. An exception to this is the "adaptive noise cancelling" technique to be discussed shortly. The common assumptions in speech enhancement research are that: (1) the noise is additive, (2) the noise and speech signals are uncorrelated, and (3) only a single input ("co-channel") channel is available.

For the enhancement techniques that focus on mathematical models, there are four general classes. All of the four classes are shown in Figure 1.4 and are described in the sections below.

1.5.1 Short-Term Spectral Domain Enhancement

The first class concentrates on properties gleaned from the short-term spectral domain. The fundamental idea is to enhance a noisy speech signal by subtracting an estimated noise bias. The earliest of these techniques is called Spectral Subtraction and is attributed to Boll [12, 13]. A generalized version of this technique is diagrammatically illustrated in Figure 1.5. In spectral subtraction, the enhancement is performed over frames by obtaining the short-term magnitude and phase of the noisy speech, subtracting an estimate of the noise magnitude spectrum, and inverse transforming this spectrum using the phase of the original degraded speech. In this case, a equals 1 in Figure 1.5, and the noise bias is an estimate of the short-term spectral envelope of the noise. As a generalization of this technique, other values of a can be used with

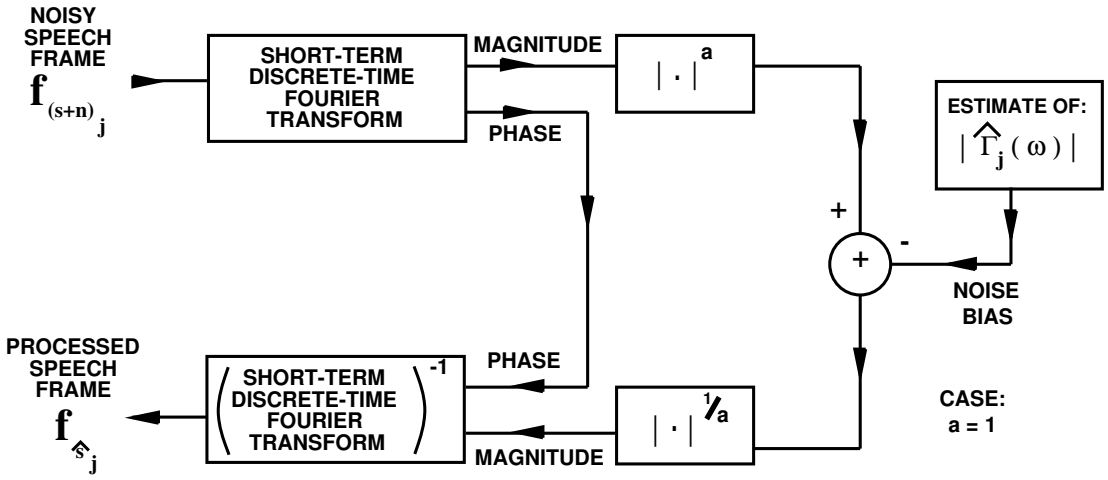

Figure 1.5: Generalized Spectral Subtraction Methodology

a corresponding change in the noise bias.

There are three noteworthy points to be made with regard to these Spectral Subtraction Methods. The first is that the background noise degrades both the spectral magnitude *and* the phase. It is reasonable to question the performance of such a technique that does not address the phase issue. Wang, in [200], addresses the relative unimportance of phase for Speech Enhancement. Secondly, these methods have, as a fundamental underpinning, that both the speech *and* the background noise are stationary. Lastly, it is not guaranteed that the speech magnitude spectrum will be greater than the noise estimate in any particular frequency region. Since negative spectral magnitudes are not allowed, simply setting the magnitude to zero for these cases results in a "musical" tone artifact. Different systems have different ways to handle this problem. Extensions of this technique have been performed by McAulay and Malpass [107], Peterson and Boll [141], Curtis and Niederjohn [32], Preuss [144], and Un and Choi [195].

1.5.2 Speech Modeling and Wiener Filtering Based Enhancement

The second class is based on speech modeling and involve either Wiener Filtering or an iterative scheme to estimate the parameters of an assumed speech model (e.g. all pole modeling). These techniques, as with the spectral subtraction techniques, use stationary arguments to estimate the parameters in a Mean Squared Error (MSE) sense. In a single-channel application, the noise statistics must be obtained either during silent frames or via an iterative procedure that imposes constraints on the speech and an underlying speech model. Below, we discuss three

types of enhancement based on these models, Iterative Wiener Filtering, Enhancement via All-Pole Modeling and Constrained Iterative Enhancement.

1.5.2.1 Iterative Wiener Filtering

Short-term Wiener filtering is an approach in which a frequency weighting for an optimum filter is first estimated from the noisy speech. The best linear estimator for the uncorrupted speech $s(n)$, in the MSE sense, is obtained by filtering the corrupted speech $y(n)$ with a noncausal Weiner filter. These models consider $\underline{s}$, $\underline{d}$ and $\underline{y}$ to be stochastic processes representing speech, noise and noisy speech, respectively. The process $\underline{d}$ is assumed to be uncorrelated to $\underline{s}$. The random variables $\underline{s}(n)$, $\underline{d}(n)$, and $\underline{y}(n)$ represent random variables from the respective processes, and $s(n)$, $d(n)$ and $y(n)$ denote realizations. As the noise $d(n)$ is considered additive, the model is simply:

$$y(n) = s(n) + d(n). \tag{1.1}$$

The goal is to formulate a linear filter which will produce an estimate of $s(n)$, say $\hat{s}(n)$, from $y(n)$ which is optimal in the MSE sense. It is well-known [135, 57, 187, 133] that such a filter would have a frequency domain transfer function of:

$$\tilde{H}(\omega) = \left[\frac{\Gamma_s(\omega)}{\Gamma_s(\omega) + \Gamma_d(\omega)}\right]. \tag{1.2}$$

where $\Gamma_s(\omega)$ and $\Gamma_d(\omega)$ are the power density spectra (PDS) for the processes $\underline{s}$ and $\underline{d}$. Because the signal $\underline{s}$ is only short-term stationary, a short-term Power Density Spectrum (stPDS)[1] for speech frame j, say $\hat{\Gamma}_s(\omega; j)$ is used. $\hat{\Gamma}_d(\omega; j)$ is a PDS estimated from adjacent periods of silence near frame j. Replacing the PSD's of Equation 1.2 with their short-term counterparts results in the a time-varying optimum frequency domain transfer function of,

$$\tilde{H}(\omega; j) = \left[\frac{\hat{\Gamma}_s(\omega; j)}{\hat{\Gamma}_s(\omega; j) + \hat{\Gamma}_d(\omega; j)}\right] \tag{1.3}$$

for frame j. The resulting frequency domain estimate for $\hat{s}(n)$ is:

$$\hat{S}_s(\omega; j) = \tilde{H}(\omega; j) S_y(\omega; j). \tag{1.4}$$

[1] This speech processing specific notation is taken from [34].

Note that $\tilde{H}(\omega; j)$ is a *zero-phase* transfer function, as both $\hat{\Gamma}_d(\omega; j)$ and $\hat{\Gamma}_s(\omega; j)$ are *real* PDS's. Weiner filtering, like spectral subtraction before it, uses the phase of the noisy spectrum for the estimate $\hat{s}(n)$. It processes only in the spectral magnitude domain. Techniques using Weiner filtering usually add additional constraints in order to effectively deal with the non-stationarity of speech. These constraints are usually based upon known properties of speech production; the constraints are, in essence, constraints on an assumed speech model. Because of this, further discussion of these Weiner filtering techniques themselves is deferred until after the next section, which describes the basis of many of these constraints.

1.5.2.2 Enhancement via All-Pole Modeling

An all-pole model is often used for the modeling of speech. It is well-known that certain phoneme classes, most notably the vowels, involve vocal-tract configurations that are acoustically resonant [41, 42, 48]. On the other hand, other phoneme classes, such as nasals and fricatives, require zeros in the model for precise modeling. Given the assumption of short-term stationarity, most researchers have found the use of an all-pole model a matter of analytic necessity. The reason for this is that a very powerful and simple computational technique of linear predictive analysis exists for deriving an all-pole model from a speech frame.

The name Linear Predictive Coding, or LPC, is collectively given to the computational technique for estimating these parameters – even though the application at hand might not be speech coding. A spectral *magnitude-only* response can be modeled with LPC poles[2]. For this reason, the spectral shape can be parsimoniously modeled by the poles of an all-pole model.

Using an all-pole model for speech production, the classical speech production of Figure 1.2 is simplified and becomes Figure 1.6.

This all-pole model has the following form for frame j:

$$\Theta(z; j) = \frac{\Theta_0(j)}{1 - \sum_{i=1}^{M} \hat{a}(i; j) z^{-i}} \tag{1.5}$$

where the $\hat{a}(i; j)$'s are the short-term LPC coefficients, M is the model order, and $\Theta_0(j)$ is a gain term for the j^{th} frame. For the case of clean speech (i.e., noise component is negligible), there are many methods to estimate these variables and most, if not all, of the methods are equivalent [151] to either the covariance method [6], the autocorrelation method [104, 96, 103], or the lattice method [97, 20].

[2]This is due to the mathematical requirement that the poles are either real or occur in complex conjugate pairs.

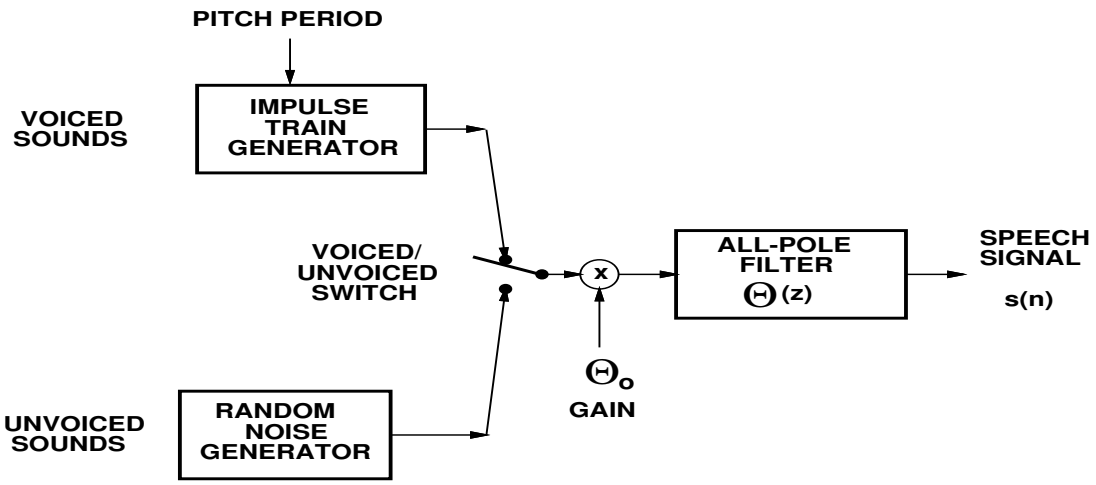

Figure 1.6: LPC Based Speech Model

Robust Linear Prediction (LP) estimates for the case of noisy speech (which is the case for speech enhancement) is a present day focus of research [169, 95]. As the estimation problem is nonlinear, iterative techniques are used which minimize various norms of the error. Recent research at Rutgers University on obtaining LP parameters which minimize various norms of the error [153] is being performed to help determine which norm spaces are most appropriate. Two successfully demonstrated methods are briefly described here to illustrate the nature of the techniques.

1.5.2.2.1 Maximum a posteriori estimation of LP coefficients

The first method by Lim and Oppenheim [90, 91] is based on *maximum a posteriori* (MAP) estimation of the LP coefficients and gain of the noise-free speech. Let $f_s(n;j)$ be the n^{th} sample in the j^{th} frame for the clean speech and $f_y(n;j)$ be similarly defined for the observed noisy speech. Consider a specific speech frame J. Then define:

$\mathbf{s} \triangleq [f_s(1;J)\ f_s(2;J)\ ...\ f_s(N;J)]$, the wanted, but unknown, speech vector.

$\mathbf{s}_I \triangleq$ given or estimated initial conditions for $\mathbf{s}_k$ vector.

$\mathbf{s}_k \triangleq$ k^{th} estimate of the vector $\mathbf{s}$.

$\mathbf{y} \triangleq [f_y(1;J)\ f_y(2;J)\ ...\ f_y(N;J)]$, the observable vector of noisy speech.

$\mathbf{a}_k \triangleq k^{\text{th}}$ estimate of the LP vector $\hat{\mathbf{a}}$.

$g_k \triangleq$ k^{th} estimate of the model gain, $\hat{\Theta}_0$.

The method generally follows the following two steps:

$$1.\ \text{Find}\ \ \mathbf{a}_k = \underset{\mathbf{a}}{\text{argmax}}\ \ p(\mathbf{a} \mid \mathbf{s}_{k-1}, \mathbf{y}, g_{k-1}, \mathbf{s}_I).$$

$$2.\ \text{Find}\ \ \mathbf{s}_k = \underset{\mathbf{s}}{\text{argmax}}\ \ p(\mathbf{s} \mid \mathbf{a}_k, \mathbf{y}, g_k, \mathbf{s}_I).$$

where $p(\ \mathbf{x} \mid \mathbf{z}\ )$ is the conditional probability of $\mathbf{x}$ given $\mathbf{z}$. For each iteration, k, $\mathbf{a_k}$, is found by LP parameter estimation using $\mathbf{s}_{k-1}$ and one of the aforementioned techniques. The g_k can be obtained using Parseval's theorem. The second step, obtaining $\mathbf{s}_k$, is performed by Wiener filtering $\mathbf{s}_{k-1}$. To do this, the optimum Wiener filter must be formed by using the stPSD's of both the $\mathbf{s}_{k-1}$ and an estimate of the noise PSD. In the case of white noise (of variance $\hat{\sigma}_d^2$), the transfer function of this filter would be:

$$\tilde{H}_k(\omega; j) = \left[\frac{\hat{\Gamma}_{s_1}(\omega; j)}{\hat{\Gamma}_{s_1}(\omega; j) + \hat{\sigma}_d^2} \right]. \tag{1.6}$$

Hansen [62] extended Lim and Oppenheim's procedure for colored noise. The optimum Wiener filter then becomes:

$$\tilde{H}_k(\omega; j) = \left[\frac{\hat{\Gamma}_{s_1}(\omega; j)}{\hat{\Gamma}_{s_1}(\omega; j) + \hat{\Gamma}_d(\omega; j)} \right]. \tag{1.7}$$

where $\hat{\Gamma}_d(\omega; j)$ is the PSD estimate of the colored noise for frame j.

1.5.2.2.2 LP estimation using estimate-maximize theory

The *estimate-maximize* (EM) algorithm was first introduced by Dempster, et al. [35] as a technique for obtaining maximum likelihood estimation from incomplete data. In the EM algorithm, the observations were considered "incomplete" with respect to some original set of "complete" data. The algorithm iterates between estimating the sufficient statistics of the complete data given the observations and a current set of parameters (E step), and maximizing the log likelihood of the complete data, using the estimated sufficient statistics (M step).

Using the same vector notation of the previous section, the observed data vector $\mathbf{y}$ is viewed as being incomplete with respect to some complete data set $\mathbf{s}$. Then, $\mathbf{y}$ is related to $\mathbf{s}$ by

$$\mathcal{H}(\mathbf{s}) \;=\; \mathbf{y}, \tag{1.8}$$

where $\mathcal{H}(\,.\,)$ is a noninvertible (many-to-one) transformation and is defined by the speech model parameters $\mathbf{a}$ considered (i.e., these $\mathbf{a}$'s need not necessarily be the same parameters of the previous section).

The EM algorithm, at each iteration k, is directed at finding the vector $\mathbf{a}_k$ such that

$$\mathbf{a}_k = \underset{\mathbf{a}}{\text{argmax}} \;\; \mathbf{E}\,[\,\log p_{\underline{y},\underline{a}}\,(\,\mathbf{s},\;\mathbf{a})\;|\;\mathbf{y},\;\mathbf{a}_{k-1}\,], \tag{1.9}$$

where $p_{\underline{\mathbf{y}},\underline{\mathbf{a}}}(\,.\,)$ is the pdf for *random vectors* $\underline{\mathbf{y}}$ and $\underline{\mathbf{a}}$. The key to this technique is the procedure to find $\mathbf{a_k}$ of Equation 1.9.

The EM approach is similar to the two-step MAP estimation previously, the main difference is the error criteria here is the expected log-likelihood function given the observed or estimated speech data.

1.5.2.3 Constrained Iterative Enhancement

Although the traditional adaptive Wiener filtering of Section 1.5.2.1 is mathematically straightforward, its practical application for speech enhancement is difficult. The non-stationary nature of speech makes the estimation of the speech signal's stPDS difficult. This estimation is further compounded if the noise is also non-stationary.

Hansen and Clements [63, 64, 65] considered an iterative Wiener filtering technique that added speech-specific constraints to the estimates of each iteration. Hansen and Clements [62] noted that although the MAP estimation of LPC parameters is shown to increase the joint likelihood of the speech waveform and the all-pole parameters of the model, a heuristic convergence criteria must be employed.

The enhancement procedure imposes spectral constraints both interframe (across time) and intraframe (in the current frame). The interframe constraints impose that pole movement (in frequency) constraints that limit the frame-to-frame pole jitter. The intraframe constraints ensure the all-pole speech model is stable and that the pole locations are not too close to the unit circle. As pole bandwidth is related to its distance from the unit circle, poles too close imply a pole bandwidth that is known to be too small for the (clean) speech signal to have produced.

In order to increase numerical accuracy and eliminate inconsistencies in pole ordering[3], the Line Spectral Pair (LSP) transformation was used to implement the interframe requirements. LSP is an alternative set of parameters that are theoretically equivalent to and can be derived from the LPC parameters [75].

Further refinements to this basic technique can be found for the case of pole-zero modeling of speech by Musicus and Lim [121], and for a method based on Kalman filtering by Paliwal and Basu [131].

1.5.3 Adaptive Noise Canceling Enhancement

Adaptive Noise Canceling (ANC) techniques refer to a class of adaptive enhancement algorithms based on the availability of a primary input source and a secondary reference source. The primary input source is assumed to contain speech plus additive noise,

$$y(n) = s(n) + d_1(n), \tag{1.10}$$

where these sequences are realizations of stochastic processes $\underline{y}$, $\underline{s}$, and $\underline{d}_1$. The secondary channel receives an input $d_2(n)$, the realization of a stochastic process $\underline{d}_2$ that is assumed to be correlated with $\underline{d}_1$ but not with $\underline{s}$. All random processes

[3] The pole ordering is critical for the interframe constraints, as non-uniformity in order and number of poles would defeat the imposition of the criteria.

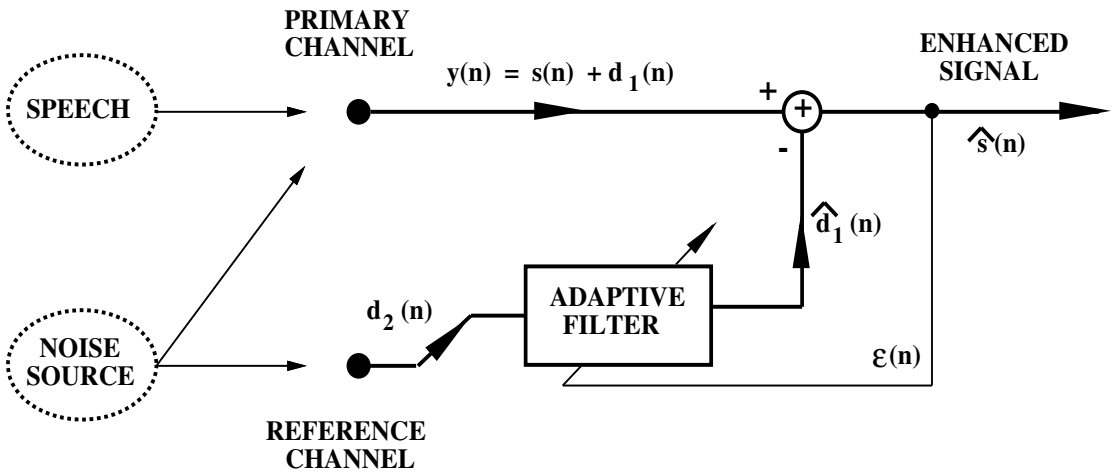

Figure 1.7: Adaptive Noise Canceling Block Diagram

are assumed Wide Sense Stationary (WSS) and appropriately ergodic. This adaptive filtering technique can be found in virtually any adaptive filtering textbook [205, 193, 73, 68] and a block diagram illustrating the signal arrangement is shown in Figure 1.7.

The overall output from the canceler $\hat{s}(n)$ is used as an "error" or control input to adjust the coefficients of the adaptive filter. The basic premise is that it should be possible to adjust the adaptive filter such that it produces an output $\hat{d}_1(n)$ that is the part of $d_2(n)$ that is correlated to $d_1(n)$. The other part (which is $d_1(n) - \hat{d}_1(n)$) is uncorrelated to $d_2(n)$, as is the signal $s(n)$ (which is assumed to be uncorrelated to $d_2(n)$). Since there is nothing in signal $d_2(n)$ that is correlated to either $s(n)$ or $(d_1(n) - \hat{d}_1(n))$, it is not possible for the filter to cancel these signals. Therefore, the filter simply adjusts its coefficients so as to minimize the energy of ϵ.

This classical noise canceling approach is based on the least mean-squared error (LMS) criterion that was first formulated by Widrow and Hoff [205]. The major advantage of the approach is that it does not require any *a priori* knowledge of the signal. The drawback is that it requires the use of a secondary channel. However, Sambur [165] has shown that a secondary channel can be *simulated* from voiced

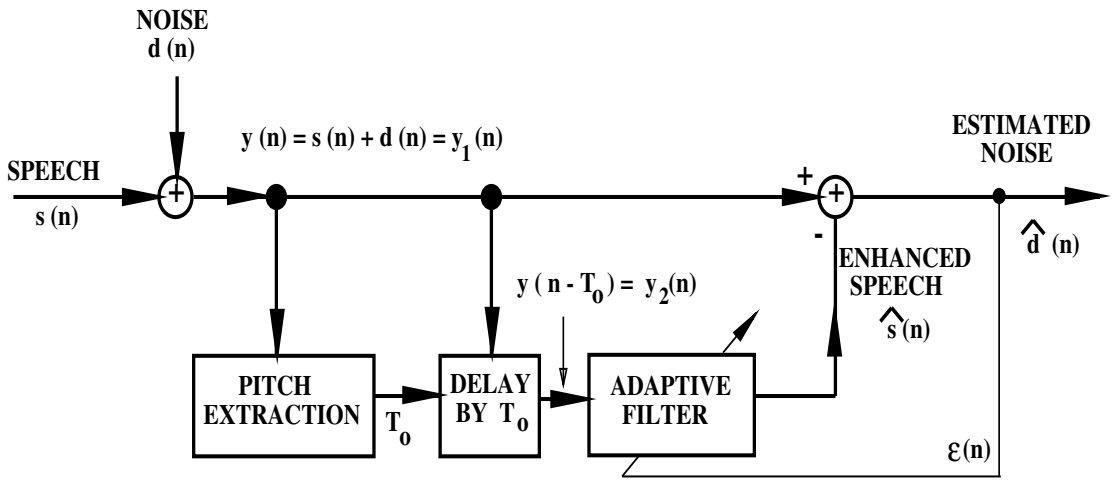

Figure 1.8: Single Channel ANC Block Diagram

speech. Because this technique requires the use of pitch tracking techniques, it is discussed in the next section.

1.5.4 Fundamental Frequency Tracking Based Enhancement

The fourth, and final, class of enhancement techniques are those based on the periodicity of voiced speech. The three major enhancement techniques that exploit periodicity are the single-channel ANC, adaptive comb filtering, and harmonic selection.

1.5.4.1 Single-Channel ANC

The dual-channel ANC discussed in Section 1.5.3 was extended for single channel inputs by simulating the reference channel. The basic idea is to obtain a reference *speech* channel during voiced sections of speech. Such a reference signal can be formed by delaying the primary (and only) speech channel by one or two pitch periods and using the LMS algorithm to generate an estimate of the $\hat{s}(n)$. A block diagram of this processing is shown in Firgure 1.8.

To see how this is possible, consider the noisy speech signal $y_1(n)$,

$$y_1(n) = s(n) + d(n). \tag{1.11}$$

Now let the reference signal, $y_2(n)$, be a delayed version of the *voiced* speech signal $y_2(n) = y_1(n - T_0)$, where T_0 reresents the pitch period. Then, under ideal periodic speech conditions, we have:

$$y_2(n) = s(n - T_0) + d(n - T_0) = s(n) + d(n - T_0). \tag{1.12}$$

The delayed speech signal $s(n - T_0)$ will be highly correlated with the original speech $s(n)$, whereas the delayed noise signals will have low correlation with the speech signals and with themselves. The adaptive filter, adapted by the LMS algorithm or a derivative of it, generates the enhanced speech signal $\hat{s}(n)$. When silence or unvoiced speech is processed, one could either not process this portion of speech or let the signal pass through the system while keeping the adaptive filter response constant (that is, not updating the filter coefficients).

Sambur [165] observed that this approach improves the quality for additive white noise in the range of 0-10 dB SNR, with higher levels of improvement as the amount of the noise increases. The main limitation of this technique is the requirement of accurate pitch estimation at low SNR's.

1.5.4.2 Adaptive Comb Filtering

Adaptive comb filtering is similar in its basic assumptions to the single channel LMS-based ANC. Since voiced speech is quasi-periodic, its magnitude spectrum contains harmonic structure. If the noise is non-periodic, its energy will be distributed throughout the spectrum. The essence of comb filtering is to build a filter that passes the harmonics of speech while rejecting noise frequency components between the harmonics. Ideally, the frequency spacing of each "tooth" in the comb filter should correspond to the fundamental frequency F_0 in Hz and should remain constant throughout the voiced section of speech. Unfortunately,

speakers normally vary their pitch and therefore require the comb filter to be adaptive. The comb filter is usually implemented in the time domain as:

$$\hat{s}(n) = \sum_{i=-L}^{L} c(i)\, y(n - iT_0) \tag{1.13}$$

where $c(i), -L \leq i \leq L$, are the $2L+1$ comb filter coefficients, $y(n)$ is the noisy speech, T_0 the fundamental period in samples, L a small constant (typically 1-6). L represents the number of pitch periods both backward and forward used in the filtering process. Choosing the delay to correspond to the fundamental period in $y(n)$ results in an averaging process that strengthens all components with that period, and attenuates or cancels others that have no period or a period different from the original fundamental.

When processing unvoiced or silence signal portion, the signal is unprocessed and scaled by 0.3 to 0.6. This scaling has been found to be necessary since applying the ACF to voiced sounds reduces the noise energy present. Failure to do this results in an unnatural emphasis of unvoiced sounds with respect to voiced sounds.

Clearly, the success of this technique is inherently dependent on the accuracy of the estimate of F_0 for the desired signal. As we shall see when we discuss the PMMM in Chapter 2, F_0 estimation is a difficult task in high noise environments as speech signals often vary from one pitch period to the next. The problem of F_0 changing within an analysis window is handled by the comb filtering technique[4] by the adding "alignment factors", ζ_i, to Equation 1.13, resulting in:

$$\hat{s}(n) = \sum_{i=-L}^{L} c(i)\, y(n - iT_0 + \zeta_i). \tag{1.14}$$

Here, individual pitch period alignments are required over the entire data analysis window and these measurements are then used as timing adjustments

[4]In the PMMM discussion we will note why this is not the optimum way to do this.

ζ_i to align adjacent pitch periods. As the number of pitch periods increase, so does the need to include the alignment factors, ζ_i. Simple ACFs set $\zeta_i = 0$ and keep the number of pitch periods L used in the filtering to a minimum (e.g. $L = 1$ results in the averaging of three pitch periods). More sophisticated techniques incorporate larger number of pitch periods and require more alignment terms. It is desirable to use as many pitch periods as possible, as L is inversely proportional to the bandwidth of each "tooth" of the comb filter.

However, due to the short-term stationary property of speech, the number of pitch periods *must not* represent more than 30 to 40 milliseconds of speech. This limits the practical range of L. We will later see that the PMMM will overcome this limitation.

Most reported results using this technique [90, 140] use *a priori* pitch information obtained from the *clean* speech to drive the adaptive comb filter. In practice, however, it is difficult to obtain estimates of F_0 in noisy environments. The PMMM processing to be described later does not require *a priori* pitch information for the enhancement.

1.5.4.3 Harmonic Selection

If the degrading noise source is from a competing speaker, then an enhancement technique similar to comb filtering can be formulated in which the spectral harmonics of the competing speaker are separated based on pitch estimates of one or both of the speakers. Parsons proposed such a method in which a high-resolution, short-term spectrum is used to separate competing speech from competing speakers [137].

A time domain alternative to Parsons technique is Time Domain Harmonic Scaling (TDHS) developed by Malah and Cox [98]. The technique involves pitch-synchronous block decimation and interpolation. The essential difference between TDHS noise reduction and ACF is that TDHS moves the noise under the pitch harmonics (i.e., masking the background noise), while comb filtering seeks to filter

out the noise in between the harmonics [34].

Both the Harmonic Selection and the TDHS techniques work well for the case of constant pitch utterances. They both have, as the underlying limitation, the property that their performance is bounded by the amount of averaging used (due to the short-term stationary nature of speech) and the need for high-resolution pitch estimation.

1.6 Need for Long-Term Model for Speech

There are many speech applications where the use of longer frame sizes would be beneficial. For example, speech enhancement accomplished via comb filtering improves as the impulse response of the comb filter (in time) is lengthened, provided that the speech is stationary for that time interval (see [130, 92] and discussion of comb filtering techniques previously in Section 1.5.4.2). Another example is the co-channel voiced speech separation problem [23], a harmonic separation problem, where separability will improve with long frame lengths, provided that the stationarity assumption holds.

A well appreciated fact in communications theory is that it is possible to minimize the effects of zero-mean statistically independent noise by the design of longer (in time) signaling sets [207]. The signaling sets for communications are designed to be stationary. Communications receivers integrate the output of a filter "matched" to the transmitted signaling constellation[5]. The longer the amount of time the noise is integrated, the smaller the noise contribution on the result of the integration will be. Impulse noise is typically mitigated by judicious selection and time extent of signaling sets [159, 194, 88].

If speech was or could be made more stationary, the above techniques and others could be used to further improve speech enhancement techniques.

---

[5] A "matched" filter is a filter whose response is a delayed, time-reversed version of the transmitted signal. For a constellation comprised of orthogonal basis elements $\varphi_i(t)$ with time extent T, the optimum matched filter would have an impulse response of $\varphi_i(T-t)$

The Pitch Mode Modulation Model is such a long term speech model that *non-linearly* time warps the speech signal such that the quasi-periodic harmonic components of speech become periodic within some time interval. Once that the signal has been made periodic, the above *linear* techniques that critically depend upon (frequency domain) stationarity can be applied. Later in this thesis, it is demonstrated that the PMMM-based time warping is performed *without* a-priori pitch information and the PMMM parameters, including a robust pitch estimate, are estimated accurately in the presence of arbitrary noise.

1.7 Shortcomings of Present Speech Enhancement and Speech Modeling

All speech enhancement techniques discussed critically depend upon some form of stationarity, either for the speech signal, the noise signal, or both, for desired enhancement. Those techniques in the first three categories, Short-Term Spectral methods, Speech Modeling and Wiener Filtering methods, and Adaptive Noise Canceling methods all depend on short-term stationarity, both in amplitude *and* frequency, for enhancement. Those in the fourth category, Fundamental Frequency methods, depend upon frequency (i.e., pitch) stationarity.

The short-term stationarity assumed in speech models and in the enhancement techniques impose processing on short segments of speech where the stationarity assumption is not seriously violated. In some speech models to be described, the amplitude or frequency non-stationarity of the signal is parametrically modeled, typically as a power series expansion. As will be seen, these constraints limit the length of speech signal to be processed, and therefore limit the degree of enhancement possible.

The Pitch Mode Modulation Model is a long-term model for speech which accommodates non-stationarity present in normal speech and does not require the parameterization of the non-stationarity to have a specific form. The PMMM processing will be seen to *non-linearly* time warp the speech signal such that

the quasi-periodic harmonic components of speech become periodic within some time interval. Once that the signal has been made periodic, the above *linear* techniques that critically depend upon (frequency domain) stationarity can be applied. Because the PMMM does not require the frequency non-stationarity to be parameterized by a specific form, the PMMM can process much longer speech frames. The PMMM-based time warping is performed without *a priori* pitch information and the PMMM parameters, including a robust pitch estimate, are estimated accurately in the presence of arbitrary noise.

1.8 Thesis Outline

In Chapter 2 the Pitch Mode Modulation Model for speech is described. In this chapter it is demonstrated that warping the speech signal by a specific time warping function results in all speech quasi-harmonic components becoming periodic. In this framework the (frequency) non-stationarity of a typical voiced utterance is seen as a type of *coherent* phase modulation of the periodic components which are the components of the PMMM decomposition of speech. The PMMM decomposition can be viewed as a generalization of the Fourier Series representation for quasi-periodic signals. Communications theory principles are used both to derive the time warping function, and to describe the effects this time warping has on the short-term speech spectrum. As an example of the ability of the PMMM to model the non-stationarity of speech in long speech frames, a 256 ms speech frame is analyzed.

In Chapter 3, the PMMM is compared and contrasted to classical speech models, sinusoidal/harmonic speech models and other speech models that parameterize speech non-stationarity. In Chapter 4, recent research on auditory fusion and cochlea modeling are expounded in order to show that the PMMM parameterization of speech is in direct agreement with these new research findings. The PMMM can, therefore, also be construed to have an auditory basis.

In Chapter 5 the primary applications of speech enhancement for speaker identification and PMMM based time-scale and pitch modification of speech are discussed. In Chapter 6 other potential applications for the PMMM, including low bit rate speech coding, are presented.

Chapter 7 presents a summary of the innovations of this research and proposed areas of continuing research.

Chapter 2

The Pitch Mode Modulation Model

2.1 Modulation Model Description

The speech modeling and enhancement techniques described here have been developed by utilizing a modulation model paradigm in which the speech signal $s(t)$ is decomposed into amplitude and phase modulated "components". Speech waveforms are, by their nature, complicated waveforms and can be decomposed into signal components in an infinite number of ways.

The following two sections describe two possible types of modulation model-based decompositions, namely the Formant Mode Modulation Model and the Pitch Mode Modulation Model. Both types of decompositions are useful for different applications. The Pitch Mode Modulation Model is used here for speech modeling and enhancement. The Formant Mode Modulation Model is presented below in order to highlight its similarity to the Pitch Mode Modulation model and because it has been successfully applied to related speech processing tasks such as speaker identification [5] and word spotting [179, 180].

2.2 Formant Mode Modulation Model

The Formant Mode Modulation Model (FMMM) [5] represents a given signal $s(t)$ as the superposition of a small number of amplitude modulated (AM) and phase modulated (PM) components $s_i(t)$. Thus the FMMM of a sampled speech signal $s(t)$ is given by

$$s(t) = \sum_{i=1}^{N} s_i(t) = \sum_{i=1}^{N} a_i(t)\cos(\phi_i(t)) = \sum_{i=1}^{N} \mathrm{Re}[a_i(t)e^{j[\phi_i(t)]}], \tag{2.1}$$

where $a_i(t)$ and $\phi_i(t)$ are the amplitude and phase modulation functions respectively of the i^{th} component of $s(t)$. For telephone quality speech the number of components, N, is usually four or five. The FMMM parameters, $a_i(t)$ and $\phi_i(t)$, are generally designed to be narrowband waveforms. Each modulation model component will usually represent a formant and, as such, their component center frequencies will track formant frequencies in time [3].

The primary benefits of FMMM processing are that $s(t)$ can be parsimoniously represented in the representation, and that $a_i(t)$ and $\phi_i(t)$ can be exploited for purposes of speech and speaker recognition. The FMMM parameters can be found as the solution to the constrained optimization problem (from [3]):

Minimize :

$$\mathcal{E} = \sum_{k=0}^{K-1} (\tilde{s}(k) - s(k))^2 \tag{2.2}$$

Subject to:

$$\tilde{s}(k) = \sum_{i=1}^{N} a_i(k)\,\cos(\phi_i(k)) \tag{2.3}$$

where K is length of the discrete data sequence. These equations, without additional constraints or *a priori* knowledge, form an underdetermined optimization problem which, in general, has an infinite number of solutions. See [3] for a description of FMMM constraints studied and techniques for solving Equation 2.2 with these constraints.

2.3 Pitch Mode Modulation Model: Fundamentals

The Pitch Mode Modulation Model (PMMM) [155], in contrast to the FMMM, decomposes $s(t)$ into AM and PM sinusoidal components whose center frequencies are integer multiples of a 'fundamental' frequency (i.e., the pitch frequency for

voiced speech). This type of decomposition is also described by Equation 2.1 except that the number of components N is usually higher (e.g., 15 to 25 for telephone quality speech).

The PMMM is a generalization of the Fourier Series representation for quasi-periodic signals with slowly time-varying amplitudes. A time warping will be defined ($\tau = g(t)$) such that its application to the quasi-periodic signal, $s(t)$, will result in a warping that makes the signal periodic in the new time base τ. Therefore, $s(t)$ in the new time base τ can be written as:

$$s(\tau) = \sum_{i=1}^{N} s_i(\tau) = \sum_{i=1}^{N} \mathrm{Re}\,[a_i(\tau) e^{j\,\phi_i(\tau)}], \tag{2.4}$$

$$= \sum_{i=1}^{N} \mathrm{Re}\,[a_i(\tau) e^{j\,(i\bar{\omega}_1\tau + K_i)}] \tag{2.5}$$

$$= \sum_{i=1}^{N} \mathrm{Re}\,[a_i(\tau) e^{j\,(i\frac{2\pi}{T_{AVG}}\tau + K_i)}] \tag{2.6}$$

where T_{AVG} is the average period of the quasi-periodic signal and K_i is a harmonic dependent phase constant. The $a_i(\tau)$, unlike a strict Fourier Series decomposition, are allowed to be slowly time varying to account for the amplitude variation of each component in real speech signals.

The $a_i(t)$ and $\phi_i(t)$, in the original time base t, are the amplitude and phase modulation functions, respectively, of the i^{th} PMMM component, or quasi-harmonic, of $s(t)$. The PMMM, therefore, models each quasi-harmonic, $s_i(t)$, as a phase modulated sinusoid with carrier frequency $i\bar{\omega}_1$ ($\bar{\omega}_1$ being the average fundamental or pitch frequency during a specific time interval) and a phase deviation about that carrier.

For a given sound, the N function pairs $\{a_i(t), \phi_i(t)\}$ possess redundant information. This redundancy is due to the synchronous variation of all the components. Using this model, the quasi-harmonics of clean, voiced speech have been found to be *time-locked* to a suitably defined time warping function, $\beta(t)$, such that:

$$\phi_i(t) \approx i\ \bar{\omega}_1\ [t + \beta(t) + \Theta_i] \quad \forall\ i,\ t \in (t_a, t_b], \tag{2.7}$$

where $\bar{\omega}_1$ is the average fundamental (pitch) frequency during a time interval $(t_a, t_b]$ and Θ_i is a harmonic dependent (phase) constant. Stated differently, a suitably defined time warping function, $\beta(t)$, can be found such that Equation 2.7 will hold approximately for all i. The time warping used in Equation 2.5 is thus $\tau = g(t) = [t + \beta(t)]$, for a suitably defined $\beta(t)$.

To find the time warping function $\beta(t)$ that satisfies Equation 2.7, several estimates of it can be made – one from each quasi-harmonic component. Assuming, for the moment, that $\phi_i(t)$ can be found for each quasi-harmonic i, an estimate of the time warping function, $\beta_i(t)$, can be obtained by solving the following for $\beta_i(t)$:

$$\phi_i(t) = i\ \bar{\omega}_1[t + \beta_i(t) + \Theta_i] = i\ \bar{\omega}_1\ t + i\ \bar{\omega}_1\ \beta_i(t) + i\ \bar{\omega}_1\ \Theta_i, \quad t \in (t_a, t_b]. \tag{2.8}$$

The term $(i\ \bar{\omega}_1\ \beta_i(t))$ is seen as $\phi_i(t)$ minus the linear term $(i\ \bar{\omega}_1\ t)$ and the constant term $(i\ \bar{\omega}_1\ \Theta_i)$. Finding independent time warping estimates from each quasi-harmonic component is tantamount to demodulating each peak in the Fourier spectrum as if it were a phase modulated sinusoid with a "carrier frequency" equal to its average frequency, $\bar{\omega}_i = i\bar{\omega}_1$. The mechanics of solving for the time warping functions $\beta_i(t)$ for all quasi-harmonics is non-trivial and is the cornerstone tool in PMMM analysis. However, as our purpose in this section is to glean analytical understanding of the PMMM, the numerical specifics concerning how to do this are deferred to Section 2.5.1.

It has been found that all harmonically related time warping estimates, $\beta_i(t)$, from a clean, voiced utterance are essentially identical, and all approximate a single time warping function, $\beta(t)$. This is the basis for the PMMM form for the phase function of the i^{th} quasi-harmonic component given in Equation 2.7. The

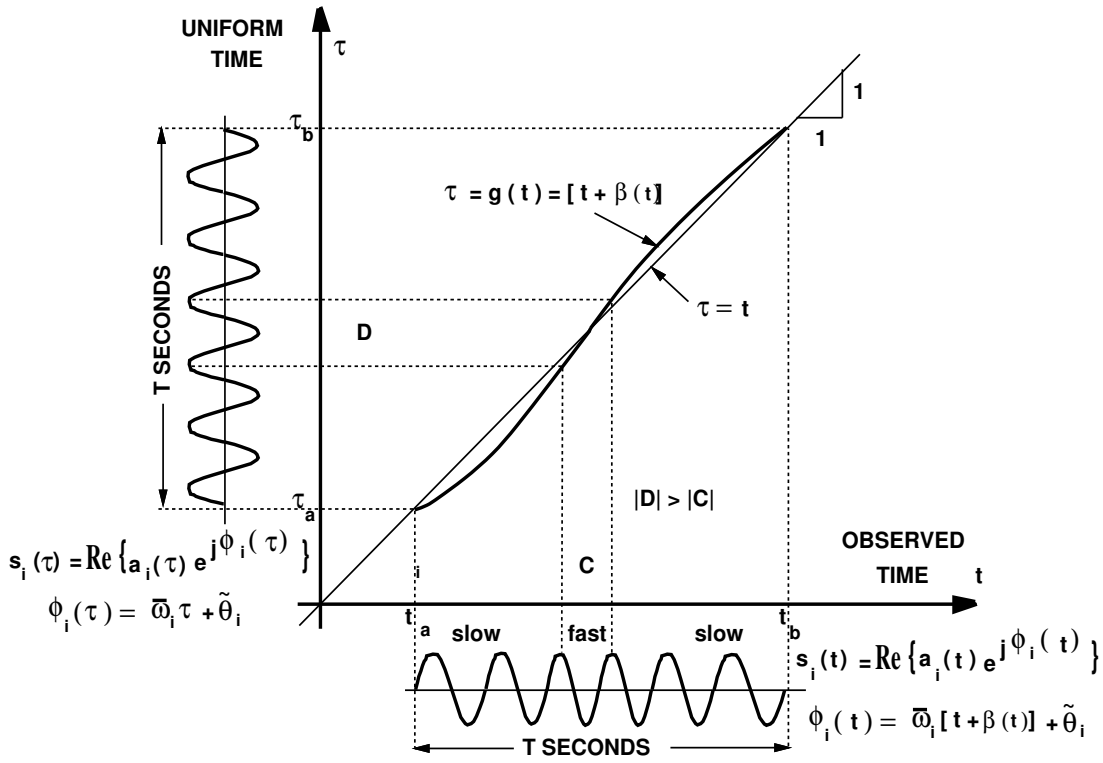

Figure 2.1: The Mapping from $\tau = g(t) = [t + \beta(t)]$

effect of a common $\beta(t)$ is a *coherent* phase modulation of the components. A detailed analysis of this coherent phase modulation follows in Section 2.4.1.

Equation 2.7, rewritten as Equation 2.9 below, can also be viewed as a mapping, $\tau = g(t)$, from the original time base t to a time base τ in which the quasi-periodic pitch components are mapped into *perfectly* periodic components with a period equal to the average period. That is:

$$\phi_i(t) \approx i\,\bar{\omega}_1[t + \beta(t) + \Theta_i] = i\,\bar{\omega}_1\tau + (i\bar{\omega}_1\Theta_i) \;\; \forall\, i, \qquad (2.9)$$
$$t \in (t_a, t_b],\; \tau \in (\tau_a, \tau_b],\; \beta(t_a) \triangleq 0,$$

where $\tau = g(t) = [t + \beta(t)]$, $\bar{\omega}_1$ is the average fundamental frequency during a time interval $(t_a, t_b]$ and the term $(i\,\bar{\omega}_1\,\Theta_i)$ is, as previously, a harmonic dependent constant.

The implication of Equation 2.9 is made clear in Figure 2.1. Figure 2.1 shows the mapping function, $\tau = g(t) = [t + \beta(t)]$, needed to map a quasi-periodic sinusoid in time base t to a perfectly periodic sinusoid in time base τ. It is seen that this mapping compresses the time scale for portions of the quasi-harmonic oscillating slower than the average frequency and expands portions oscillating faster than the average. Since $(\tau_b - \tau_a) = T$ seconds $= (t_b - t_a)$, the average number of cycles per second for each quasi-harmonic sinusoid, $\bar{\omega}_i$, is the same for both time bases.

Note that the time warping function $\beta(t)$ is *defined* as the *deviation* of the mapping $\tau = g(t)$ from the line $\tau = t$ during time interval $(t_a, t_b]$, and that $\beta(t_a) \triangleq \beta(t_b) \triangleq 0$. Also note that the mapping $\tau = g(t) = [t + \beta(t)]$ is, in general, a *nonlinear* mapping from t to τ.

The variation of $\beta(t)$ is exceedingly small for voiced speech; typical values for $\max |\beta(t)| \approx 0.001 * (t_b - t_a) \sim 0.005 * (t_b - t_a)$ for voiced frame sizes $(t_b - t_a)$ of $35 \sim 200$ milliseconds. In other words, the plots of $\tau = t$ and $\tau = [t + \beta(t)]$ for voiced speech would be virtually indistinguishable if plotted together on the same scale. This claim is substantiated in the next section.

2.3.1 Beta Function Extraction Example

As a simple example of the statements already made, Figure 2.2 presents the magnitude of a 4096 point FFT of an vowel /æ/, as in the word "bat", sampled at 16 kHz. This particular utterance is unusually long (256 milliseconds) and spoken with an effort to maintain a steady pitch. Although not typical in this regard, it is shown here to illustrate the redundancy of voiced speech.

If this speech were truly periodic, we would expect to see an approximate line spectra in Figure 2.2[1]. As we shall soon demonstrate, the broadening of the higher order harmonics is mostly due to pitch non-stationarity.

[1] For a perfect line spectra, the component amplitudes would be required to be constants and the 4096 samples would need to be an exact multiple of the period.

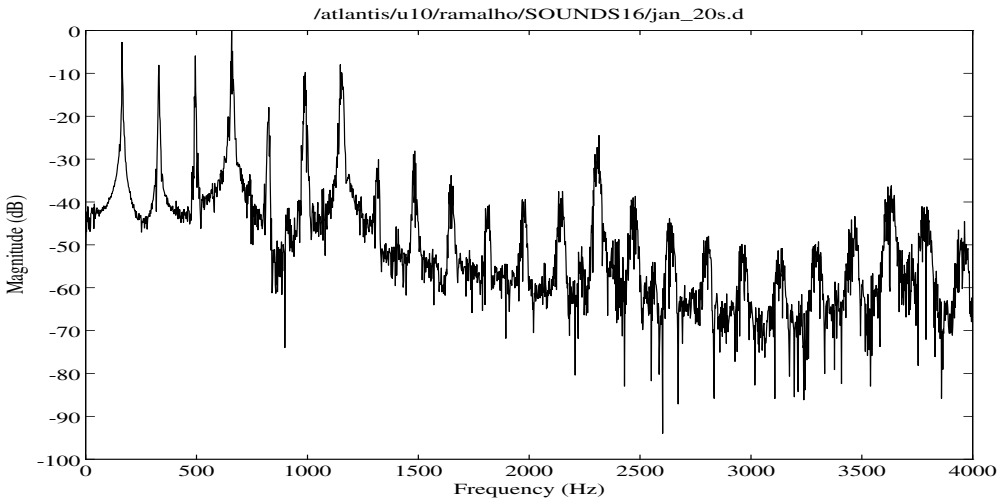

Figure 2.2: STFT Magnitude of vowel /æ/ in "bat" (4096 samples @ 16kHz sampling)

Figure 2.3 presents the $\beta_i(t)$ estimates for the first ten components of the same utterance as the Short Term Fourier Transform (STFT) of Figure 2.2. The time warping functions, $\beta_i(t)$ $i = 1...10$, estimated from the components of this utterance are consistent, that is, they tend to cluster about an average $\beta(t)$. A representative time warping function, $\beta(t)$, can be obtained from this cluster of $\beta_i(t)$'s such that Equation 2.9 holds approximately for all quasi-harmonic components. Note that the *deviation* from uniform time necessary to make the quasi-harmonics periodic is very small, about 0.2% (about eight samples out of the 4096 sample duration).

The harmonic components of Figure 2.2 were clearly identified and successfully demodulated because the spectra of the quasi-harmonics did not overlap. This was because the utterance was spoken with an effort to maintain constant pitch. As most voiced utterances are not spoken in this manner, the harmonic structure of the higher harmonics are not always evident from the STFT Magnitude. Iterative techniques to successfully obtain the phase estimates, $\phi_i(t)$, in this case are outlined in Section 2.5.4. Before this, however, it is instructive to observe the effect that a *coherent*, $\beta(t)$ induced, phase modulation has on the spectrum of a

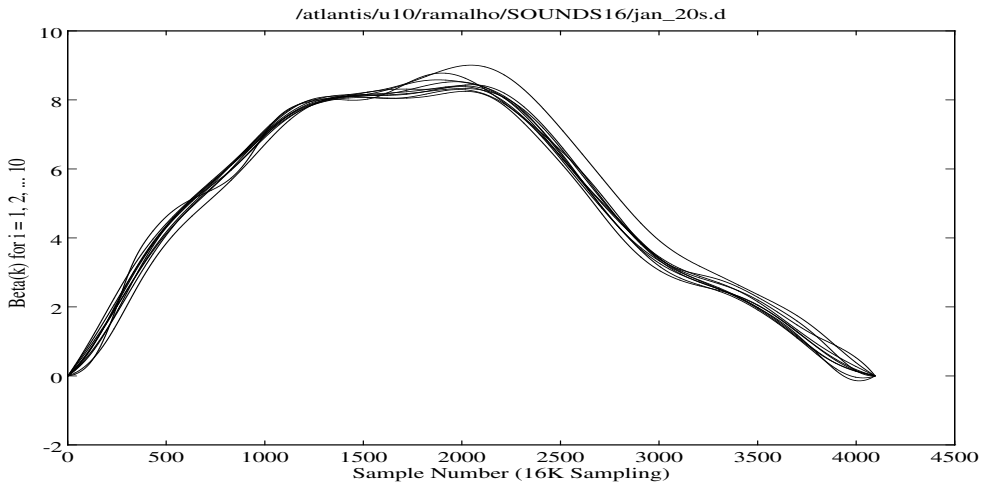

Figure 2.3: $\beta_i(t)$ Estimates for $i = 1, 2, ...10$

voiced speech signal. This is done in the following section.

2.4 Pitch Mode Modulation Model: Detailed Analysis

2.4.1 Spectral Implications of $\beta(t)$ Modulation

This section addresses the effects that a *coherent*, $\beta(t)$ induced, phase modulation has on the spectrum of a voiced speech signal. To do this, we first assume that the PMMM form for the phase holds exactly for all quasi-harmonics, that is:

$$\phi_i(t) = i\,\bar{\omega}_1[t + \beta(t) + \Theta_i] \quad \forall\, i,\ t \in (t_a, t_b]. \tag{2.10}$$

This conjecture is then examined in the frequency domain to determine the spectral effect of the $\beta(t)$ induced modulation.

Figure 2.1 illustrates yet another way to interpret the mapping $\tau = g(t) = [t + \beta(t)]$. The phase modulation term of Equation 2.10, $(i\,\bar{\omega}_1\beta(t))$, can be viewed as the phase modulation function necessary to modulate each harmonic component of a purely periodic harmonic series of sinusoids into the set of quasi-harmonic sinusoids that are the PMMM decomposition of speech. To see this,

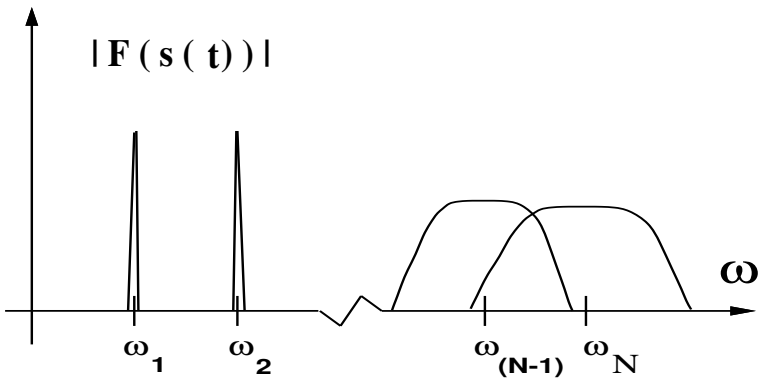

Figure 2.4: Illustrative Fourier Magnitude for $s(t)$ with $\beta(t)$ Modulation

the instantaneous frequency of the i^{th} quasi-harmonic component (i.e., the time derivative of the phase function $\phi_i(t)$ [2]) is given by:

$$\omega_i(t) = \frac{d\phi_i(t)}{dt} = i\,\bar{\omega}_1 + i\,(\bar{\omega}_1 \frac{d\beta(t)}{dt}). \tag{2.11}$$

Borrowing from communications theory lexicon, the first term can be interpreted as a carrier frequency ($i\,\bar{\omega}_1$) and the second term as the frequency modulation about that carrier[3]. Since the ($\bar{\omega}_1 \frac{d\beta(t)}{dt}$) part of the frequency modulation is common to all quasi-harmonics, the pitch harmonics with higher center frequency carriers (higher harmonic number i) have a correspondingly higher instantaneous frequency deviation.

To illustrate the phase modulation effect of $\beta(t)$, Figure 2.4 depicts the case where the $\beta(t)$ of Equation 2.9 is non-zero. For diagrammatic simplicity, Figure 2.4 assumes the $a_i(t)$ of Equation 2.1 are set to one, that is:

$$s_{\text{FIG 2.4}}(t) = \sum_{i=1}^{N} \text{Re}[e^{j[\phi_i(t)]}] = \sum_{i=1}^{N} \text{Re}[e^{j\,i\bar{\omega}_1[t+\beta(t)+\Theta_i]}]. \tag{2.12}$$

[2] Strictly speaking, there are many definitions for the instantaneous frequency for both amplitude and frequency modulated signals. Most definitions approach, in some sense, the time derivative of the phase function for small amplitude modulation in the time interval considered. For more information concerning instantaneous frequency definitions see [77, 17]. For details on *discrete-time* instantaneous frequency definition and computation see [190].

[3] It is recognized in communications literature that one cannot tell from a given waveform whether it has been phase or frequency modulated. The essential difference is the nature of the dependency on the modulating signal. The following analysis follows from a frequency modulation viewpoint.

In this figure, it is seen that the spectral width of the higher harmonics are more broad than those of the lower harmonics. In general, the energy contained in the higher pitch harmonics will overlap in frequency, making the demodulation of the higher harmonics difficult. This effect will become more clear shortly.

Phase Modulation, unlike amplitude modulation, is a *nonlinear* form of modulation. It is not possible to evaluate the spectra of any PMMM component for an arbitrary $\beta(t)$ because the superposition principle does not apply [189, 191, 86]. In AM modulation, if one knows the spectra for the modulating signal, then the spectra of the modulated signal is a frequency translated version of this spectra (due to the modulation property of the Fourier Transform). If the AM modulating signal can be decomposed into a *linear sum* of elementary functions (e.g., a harmonic or power series), the spectrum of the modulated signal would be a frequency shifted version of the *linear sum* of their respective spectra. No such decomposition for the purposes of estimating PM spectra is possible. Theoretically, one would have to perform an analysis for every $\beta(t)$ one would expect to encounter.

In order to gain analytical and mathematical insight to this type of modulation, we are forced to consider a specific form for the $\beta(t)$ modulation in which a closed form is known to exist – the case of sinusoidal modulation. Thereafter, the results for this case will be heuristically generalized in a completely analogous manner to the case of Frequency or Phase Modulation in communications theory.

2.4.2 Analysis for Sinusoidal $\beta(t)$

Consider the specific case where the modulation function $\beta(t)$ is sinusoidal. From communications theory, we know that if $\beta(t)$ is sinusoidal with frequency ω_{mod}, $\beta(t) = \alpha \sin(\omega_{\text{mod}} t)$, then the resultant spectra would have 'sidebands' spaced symmetrically on either side of the pitch 'carrier' $(i\,\bar{\omega}_i)$ at frequency separations ω_{mod}, $2\omega_{\text{mod}}$, $3\omega_{\text{mod}}$, etc., [191, 189, 173]. To see this for our case of coherent $\beta(t)$ modulation, consider a quasi-harmonic component with $\beta(t) = \alpha \sin(\omega_{\text{mod}} t)$. The

resulting quasi-harmonic component is given by:

$$s_i(t) = a_i(t)\mathrm{Re}[e^{j\,i\bar{\omega}_1(t+\Theta_i)}e^{j\,i\bar{\omega}_1\alpha\sin(\omega_{\mathrm{mod}}t)}]. \tag{2.13}$$

We will find it useful to define $\psi = i\,\bar{\omega}_1\alpha$ as the modulation index; the reason for this terminology will be made clear shortly. The rightmost complex exponential of the previous equation can be expanded into its Fourier series representation as:

$$e^{j\,i\bar{\omega}_1\alpha\,\sin(\omega_{\mathrm{mod}}t)} = e^{j\psi\,\sin(\omega_{\mathrm{mod}}t)} = \sum_{n=-\infty}^{\infty} J_n(\psi)e^{jn\omega_{\mathrm{mod}}t}, \tag{2.14}$$

where,

$$J_n(\psi) = \frac{1}{T_{\mathrm{mod}}}\int_{-T_{\mathrm{mod}}/2}^{T_{\mathrm{mod}}/2} e^{j\psi\sin(\omega_{\mathrm{mod}}t)}e^{-jn\omega_{\mathrm{mod}}t}dt. \tag{2.15}$$

Making a change of variables $\xi = \omega_{\mathrm{mod}}\,t = (2\pi/T_{\mathrm{mod}})\,t$, we get:

$$J_n(\psi) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{j(\psi\sin\xi - n\xi)}d\xi\ . \tag{2.16}$$

This integral equation occurs in many physical problems and has been tabulated extensively. It is a function of n and ψ, denoted by $J_n(\psi)$, and is called the Bessel function of the first kind (signified by the "J") of order n and argument ψ.

Rewriting Equation 2.13 by using the expansion of Equation 2.14 results in:

$$s_i(t) = a_i(t)\mathrm{Re}\left[e^{j\,i\bar{\omega}_1(t+\Theta_i)}\sum_{n=-\infty}^{\infty} J_n(\psi)e^{jn\omega_{\mathrm{mod}}t}\right]\ , \tag{2.17}$$

$$= a_i(t)\sum_{n=-\infty}^{\infty} J_n(\psi)\cos([i\bar{\omega}_1 + n\omega_{\mathrm{mod}}]t + i\bar{\omega}_1\Theta_i)\ , \tag{2.18}$$

the last step being allowed because the Bessel functions, $J_n(\psi)$, are real. Furthermore, due to the fact that $J_n(\psi) = J_{-n}(\psi)$ for n even, and $J_n(\psi) = -J_{-n}(\psi)$ for n odd, we can further expand Equation 2.18 as:

$$
\begin{aligned}
s_i(t) &= a_i(t)\,\{\ J_0(\psi)\ [\cos(i\bar{\omega}_1 t + i\bar{\omega}_1\Theta_i)] \qquad (2.19)\\
&+\ J_1(\psi)\ [\cos([i\bar{\omega}_1 + \omega_{\mathrm{mod}}]t + i\bar{\omega}_1\Theta_i) - \cos([i\bar{\omega}_1 - \omega_{\mathrm{mod}}]t + i\bar{\omega}_1\Theta_i)]\\
&+\ J_2(\psi)\ [\cos([i\bar{\omega}_1 + 2\omega_{\mathrm{mod}}]t + i\bar{\omega}_1\Theta_i) + \cos([i\bar{\omega}_1 - 2\omega_{\mathrm{mod}}]t + i\bar{\omega}_1\Theta_i)]\\
&+\ J_3(\psi)\ [\cos([i\bar{\omega}_1 + 3\omega_{\mathrm{mod}}]t + i\bar{\omega}_1\Theta_i) - \cos([i\bar{\omega}_1 - 3\omega_{\mathrm{mod}}]t + i\bar{\omega}_1\Theta_i)]\\
&+\ \ \ldots\ \ \}.
\end{aligned}
$$

This is what we set out to prove. Namely, that if $\beta(t)$ is of the form of a sinusoid of frequency ω_{mod}, the resultant spectra would have 'sidebands' symmetrically on either side of the pitch 'carrier' ($i\bar{\omega}_i$) at frequency separations ω_{mod}, $2\omega_{\mathrm{mod}}$, $3\omega_{\mathrm{mod}}$, etc. We now see that the number of these sidebands is infinite, and the magnitude of a given sideband is given by the solution of the Bessel function of the first kind for the specific modulation index, $\psi = i\,\bar{\omega}_1\alpha$, and order, n. This is in sharp contrast to Narrowband AM modulation. If the amplitude of a constant frequency sinusoid of frequency $\omega_{carrier}$ was modulated at a frequency ω_{mod} ($\omega_{\mathrm{mod}} \ll \omega_{carrier}$), the resultant spectra would only have *one* sideband no matter how large the modulation. AM is called a *linear* form of modulation, while angle modulation (FM or PM) is inherently *nonlinear*.

Although the PM introduced by $\beta(t)$ theoretically introduces an infinite number of sidebands, the magnitudes of the higher-order sidebands become negligible and, for all practical purposes, the power is contained within a finite bandwidth. To show this, Figure 2.5 shows a plot of the coefficients $J_n(\psi)$ plotted as a function of the modulation index ψ.

For small frequency modulation indices, $\psi = i\,\bar{\omega}_1\alpha \leq 0.2$, the modulation is approximated well as a *linear* form of modulation. This is because the only significant $J_n(\psi)$ are $J_0(\psi)$ and $J_1(\psi)$. This case is referred to as Narrowband Frequency Modulation in communications literature.

However, for large modulation indices the modulation becomes increasingly nonlinear, and the number of sidebands of significant energy grows. The case of

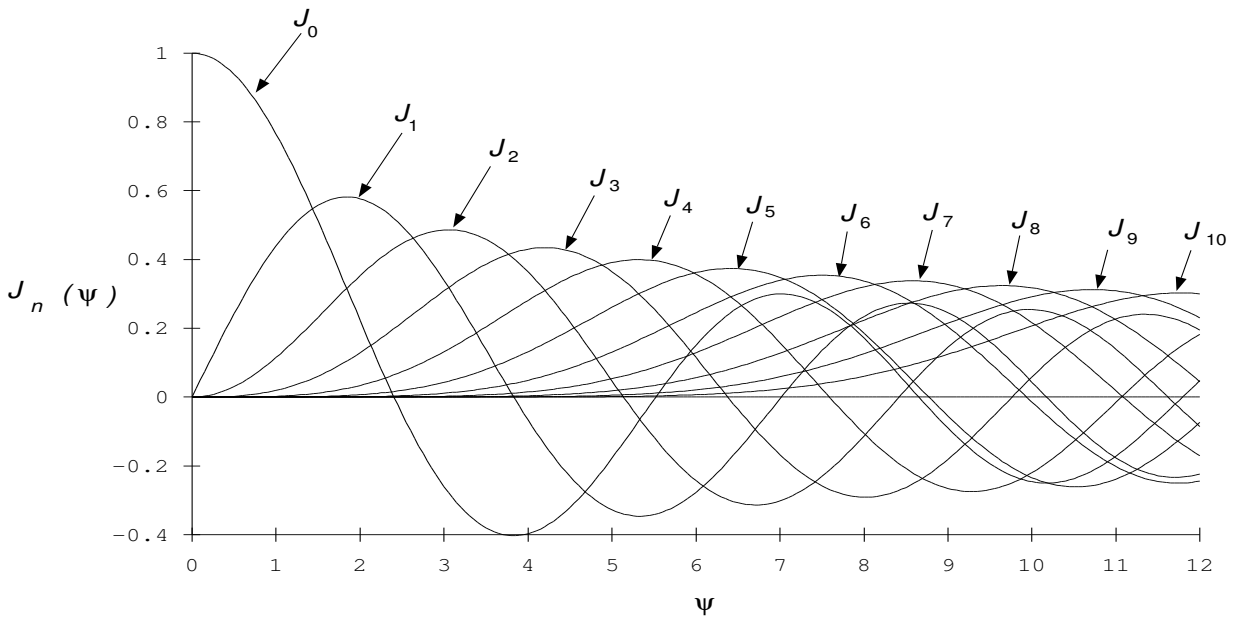

Figure 2.5: $J_n(\psi)$ as a Function of Modulation Index ψ

$\psi = i\bar{\omega}_1\alpha > 0.2$ is referred to as Wideband Frequency Modulation in communications literature. Depending on the modulation index $\psi = i\bar{\omega}_1\alpha$, the low harmonic index pitch harmonics may be approximated well by narrowband PM, while for increasingly higher pitch harmonics the modulation is increasingly wideband. This fact makes the demodulation of the higher pitch harmonics increasingly difficult.

Two more points should be made in regard to the Bessel functions, $J_n(\psi)$, and angle modulation in particular. The Bessel functions are normalized such that $J_0^2(\psi) + 2J_1^2(\psi) + 2J_2^2(\psi) + 2J_3^2(\psi)... = 1$ for any ψ. This equality states that the sum of the power in the sidebands (whose power is proportional to their $J_n^2(\psi), \quad n \neq 0$) plus the power in the carrier ($J_0^2(\psi)$) equals the power of an unmodulated carrier ($J_0^2(0) = 1$ and $J_n^2(0) = 0, \; \forall \; n \neq 0$), or

$$J_0(0) \; = \; 1 \; = \left[J_0^2(\psi) + 2 \sum_{n=1}^{\infty} J_n^2(\psi) \right]. \tag{2.20}$$

Thus the power which was concentrated at a particular frequency of an unmodulated sinusoid is "spread out" over the sidebands where $J_n^2(\psi)$ is significant[4]. Figure 2.4 graphically makes this point by illustrating that the Fourier magnitude decreases with the increasing bandwidth of the higher quasi-harmonics.

The second point deals with the term bandwidth. A rule commonly adopted by FM communications theorists is that a sideband is considered significant if its magnitude exceeds 1% of the unmodulated carrier, i.e., if

$$|J_n(\psi)| \geq 0.01. \tag{2.21}$$

Utilizing this rule of thumb, a simple formulation for determining the bandwidth W of a FM signal was formulated by J. R. Carson. It is called Carson's rule [189, 191, 86] and is given by:

$$W \approx 2\omega_{\text{mod}}(1 + \psi), \tag{2.22}$$

and it approaches the correct limit for both very large and very small modulation indices ψ. For our $\beta(t) = \alpha \sin(\omega_{\text{mod}} t)$ analysis above, it says

$$W \approx 2\omega_{\text{mod}}(1 + i\bar{\omega}_1\alpha). \tag{2.23}$$

In other words, when the second term in parenthesis dominates, the bandwidth increases linearly with increasing harmonic number. Thus, the bandwidth increases so as to maintain a constant Q. This type of nonlinear relationship is reminiscent of critical band theory in psychoacoustics [170, 1, 129, 119] and Mel-Warping of Spectra in speech processing (e.g., Mel-Cepstrum [34]). This constant Q relationship is revisited in Section 4.2, where a PMMM tie to cochlear modeling is discussed.

[4] See [191] for detailed analysis and [143] for proof of the Bessel function identity in Equation 2.20.

It is helpful for our later discussion to make some additional comments in regard to the detailed spectral shapes of the quasi-harmonic components in the composite spectrum of $s(t)$. The STFT Magnitude of Figure 2.2 and some plots of spectra to follow show the trend of increasing harmonic bandwidth with increasing harmonic number. One may expect that the spectral shapes should be somehow similar since the identical form of the modulation (up to multiplication by the harmonic number) is identical for all harmonics. Dissimilar harmonic spectra have lead many speech coding researchers to (incorrectly) surmise that the excitation for these frequency bands of dissimilar spectra were created by a different excitation than the other bands[5]. Because the $\beta(t)$ modulation is a *nonlinear* form of modulation, we *should not* assume similar spectral shapes - superposition arguments do not hold[6]. We should instead *expect* significant visual differences in the individual harmonic spectra. To further confuse things, it is well recognized that similar spectra can be created by different forms of modulation, and that significantly different spectra can be formed by seemingly similar modulation forms [134, 203].

Another interesting special case to note is when the modulation index ψ is such that $J_0(\psi) = 0$, for example when $\psi \approx 2.4$ or 5.75. When this occurs, the spectra of the quasi-harmonic component is missing its "carrier" frequency - despite the fact that the signal is a frequency modulated sinusoid at that carrier frequency!

The main point of this discussion on spectra is to note that it is unreasonable to expect similar spectral shapes for the case of quasi-harmonics that have been modulated in a wideband manner. The above analysis considered the case of sinusoidal $\beta(t)$ and made some observations about what we should expect for the spectra of $s(t)$. Fortunately, frequency modulation practices in communications show that these results, called tone modulation in communications, and observations can be heuristically extended to the case of an arbitrary $\beta(t)$ modulation

[5] This topic will be discussed in greater detail later in the Harmonic Coding Section 3.2.

[6] See discussion at the end of Section 2.4.1

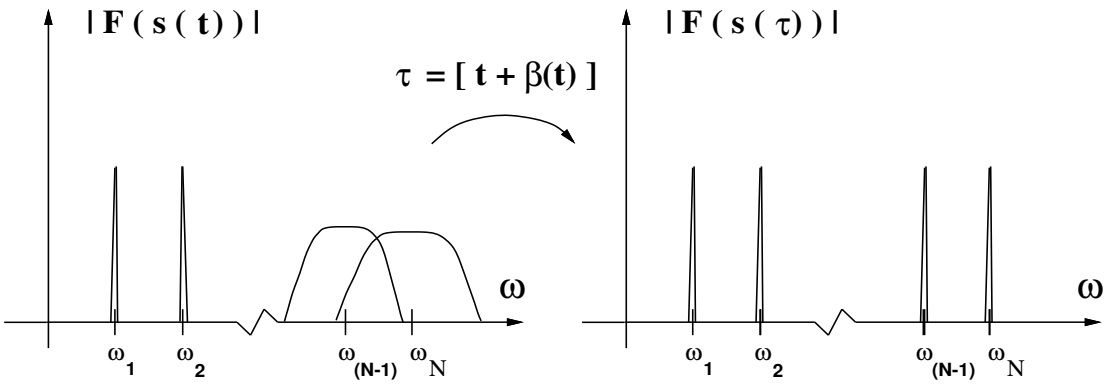

Figure 2.6: Illustrative Fourier Magnitude for $s(t)$ and $s(\tau)$

(see [189, 191, 86]).

2.4.3 The Goal of PMMM Preprocessing

In summary, the modulation caused by the mapping $\tau = g(t) = [t + \beta(t)]$ results in a *nonlinear* form of modulation. Finding $\beta(t)$ in the PMMM paradigm is tantamount to demodulating each quasi-harmonic. Higher harmonic number pitch harmonics become increasingly difficult to demodulate, as their bandwidths increase with harmonic number. In fact, the bandwidths for the higher harmonic number quasi-harmonics may overlap in frequency – making demodulation in one step impossible.

The goal of PMMM processing is to obtain a $\beta(t)$ such that the relation $\tau = [t+\beta(t)]$ can be applied to time-warp the quasi-periodic $s(t)$ into a periodic $s(\tau)$. If successful, this time-warping procedure will result in $s(\tau)$ having an approximate line spectra. This goal is illustrated in Figure 2.6. This time-warping procedure, which we refer to as "de-warping", will undo the effect of the $\beta(t)$ induced phase modulation – the harmonic peaks will sharpen and their spectral magnitudes will increase.

Novel techniques for estimating the time warping functions, $\beta_i(t)$, from each component have been developed as a part of this research. As the alert reader may have surmised, iterative techniques for obtaining the $\beta(t)$ estimates are needed

since the higher order quasi-harmonics cannot be demodulated if their spectra significantly overlap (in energy)[7]. In this case, a $\beta(t)$ is estimated from the components that can be demodulated and the relation $\tau = [t + \beta(t)]$ is applied to de-warp $s(t)$ to become more periodic. After this de-warping, the pitch harmonics should have become more sharpened (their deviation from their harmonic carrier frequencies should be lessened) and additional pitch peaks may now be able to be demodulated. A second $\delta\beta(t)$ is found and applied to make the signal even more periodic. This process is repeated until all pitch harmonics are of uniform bandwidth. This iterative procedure for robust $\beta(t)$ estimation is discussed thoroughly in Section 2.5.4.

2.5 PMMM Parameter Estimation Algorithm

The PMMM exploits the fact that a subset of independent time warping estimates can be intelligently combined to provide a robust estimate of $\beta(t)$. This robust $\beta(t)$ can be used to help reduce the unwanted noise for the application at hand.

This ability to obtain a representative $\beta(t)$ for the example of Section 2.3.1 was demonstrated previously for the case of mild form of $\beta(t)$ modulation. The methodology reported below has been "robustized" to handle the case of arbitrary noise and does not assume any *a priori* knowledge about the speech signal other than the fact that it is voiced. In this section, we will outline the specifics of techniques developed to obtain a robust $\beta(t)$ necessary for the de-warping of $s(t)$.

2.5.1 Beta Function Extraction Methodology

Independent time warping estimates are found for each observable quasi-harmonic component of the Fourier spectrum by solving: $\phi_i(t) = \bar{\omega}_i[t + \beta_i(t) + \Theta_i]$ for $\beta_i(t)$. In this section, we describe how $\phi_i(t)$ is estimated from each quasi-harmonic.

[7] For commercial FM transmission, a significant overlap in adjacent channel spectra is prohibited by Federal Communications Commission regulations. If the overlap was significant, the receivers would not be able to demodulate the signal correctly.

We assume, for the moment, that the spectral smearing is such that the quasi-harmonics of the STFT of the speech considered do not significantly overlap (as in the speech spectrum of Figure 2.2). We later discuss when this assumption is not true. In the discussion which follows below, the continuous time representation is used to avoid cumbersome indexing and for continuity with the previous analysis. When particular algorithmic details require a reference in discrete-time, the sample index k is used.

The first task is to sift through the STFT for the block of speech samples considered looking for a local peak which *may* be a pitch harmonic. To test this hypothesis, we consider this peak as having the form of a frequency modulated signal. Then, to obtain a $\phi(t)$ estimate, an analytic signal is created by isolating this peak via frequency domain filtering with a specified bandwidth. The time warping, $\beta(t)$, estimate for this peak is generated from the $\phi(t)$ estimate.

$$s(t) = \sum_{i=1}^{N} s_i(t) = \sum_{i=1}^{N} a_i(t)\cos(\phi_i(t)) = \sum_{i=1}^{N} \mathrm{Re}\left[a_i(t)e^{j[\phi_i(t)]}\right], \tag{2.24}$$

We express each $s_i(t)$ of Equation 2.1, which is rewritten above for convenience as Equation 2.24, by means of an analytic signal, denoted $S_i(t)$ (with an upper case S). We obtain $S_i(t)$ by taking the Hilbert Transform of each $s_i(t)$ [163, 125]. The analytic signal representation is $S_i(t) = [s_i(t) + j\hat{s}_i(t)]$, $\hat{s}_i(t)$ being the phase quadrature signal to $s_i(t)$.

Expressed in analytic signal representation, the Equation 2.24 becomes:

$$S(t) \;=\; \sum_{i=1}^{N} S_i(t) \;=\; \sum_{i=1}^{N} [s_i(t) + j\hat{s}_i(t)] \;=\; \sum_{i=1}^{N} a_i(t)\; e^{j\phi_i(t)} \tag{2.25}$$

An analytic signal is formed for the i^{th} component by single-side bandpass filtering about the Fourier domain spectral peak. Specifically, a frame of speech is first windowed using a Hamming window and a Fast Fourier Transform (FFT) is computed. Then the FFT is single-side bandpass filtered about the positive frequency component peak and an Inverse FFT (IFFT) is performed to obtain

$S_i(k) = a_i(k)e^{j\phi_i(k)}$. The phase estimates $\phi_i(k)$ from only the center half of the Hamming windowed frame are used[8] and the index k_{begin} is used to denote the index of the first phase estimate used in the frame. Next, the average frequency, $\bar{\omega}_i$, of the component is calculated as:

$$\bar{\omega}_i = \frac{1}{|t_b - t_a|} \int_{t_a}^{t_b} \dot{\phi}_i(t)dt \tag{2.26}$$

where $\dot{\phi}_i(t)$ is the derivative of $\phi_i(t)$ with respect to time. Numerically, the derivative is approximated by a first difference, $\dot{\phi}_i(k) = \phi_i(k) - \phi_i(k-1)$, then clipped to within the median $\dot{\phi}_i(k)$ plus or minus 125 Hz within the frame interval considered. The clipping function is *not* needed if the peak considered is a harmonic peak and the signal to noise ratio is sufficient. The clipping was experimentally found to improve the $\beta(t)$ clustering process, and the reason for its inclusion will be explained shortly. The integration of Equation 2.26 is calculated numerically using the trapezoidal rule.

A new $\phi_i(t)$, denoted $\hat{\phi}_i(k)$, is calculated from the clipped $\dot{\phi}_i(k)$ such that the derivative of $\hat{\phi}_i(k)$ yields the clipped $\dot{\phi}_i(k)$. This is done by cumulative summing the elements of $\dot{\phi}_i(k)$, then adding $\phi_i(k_{\text{begin}})$ to all the frame elements. If no elements were clipped in the generation of $\dot{\phi}_i(k)$, $\hat{\phi}_i(k)$ will be equal to the original $\phi_i(k)$.

Next a $\beta_i(k)$ is found by solving the equation $\hat{\phi}_i(k) = \bar{\omega}_i[k + \beta_i(k) + \Theta_i]$ for $\beta_k(k)$ subject to the constraint $\beta_i(k_{\text{begin}}) = 0$. The $\beta_i(k)$ so constructed from the $\hat{\phi}_i(k)$ will also satisfy the constraint $\beta_i(k_{\text{end}}) = 0$[9].

It has been found via experimentation that the time warping estimates can be improved if more than one estimate is made for the Fourier domain peak considered. To do this, we form three different analytic signals from the same

[8] Using synthetic signals it has been found that this method of obtaining the *phase* of a known frequency modulated signal was superior to most other methods and is relatively straightforward to implement. Later, when we are interested in the *amplitude* of a harmonic, a different windowing strategy is used.

[9] If we did not compute a the new $\hat{\phi}_i(k)$ from the clipped $\dot{\phi}_i(k)$, then $\beta_i(k_{\text{end}})$ would not equal zero when clipping took place. This is because $\bar{\omega}_i$ was calculated using the clipped $\dot{\phi}_i(k)$. The constraint $\beta_i(k_{\text{end}}) = 0$ is important for the beta clustering to be discussed in Section 2.5.2.

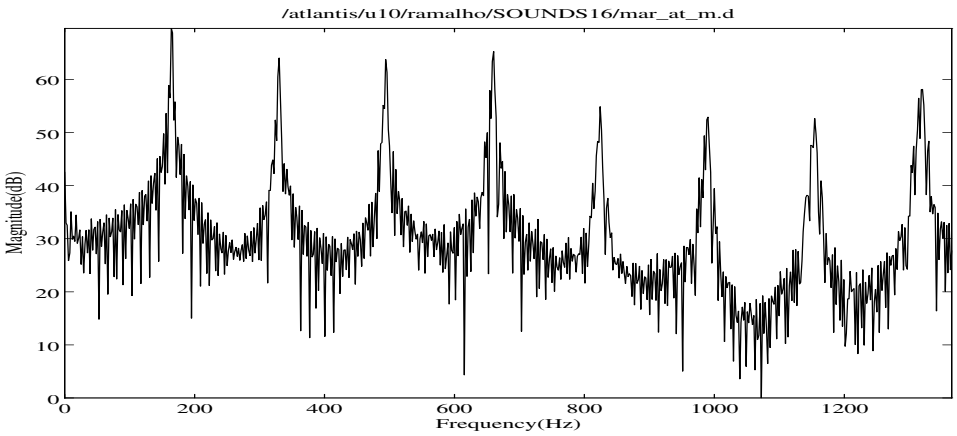

Figure 2.7: First eight pitch harmonics of vowel /æ/ in "bat"

Fourier domain peak, each one using a slightly different bandwidth. These time warping estimates so obtained should be substantially similar to each other if the peak was a pitch harmonic – the only substantial difference is that this new estimate used slightly different "sideband" information in forming the estimate. It has been found that if this peak was a spurious noise peak, the time warping estimates will not be similar to each other. This is because noise is poorly modeled as a frequency modulated signal.

To make the above more concrete, consider the following example. Figure 2.7 shows the first eight harmonics of a voiced utterance. Figure 2.8 shows three $\beta(t)$ estimates derived from the first pitch harmonic of Figure 2.7. These three estimates were derived using three different bandwidths in the generation of the analytic signal from which the $\beta_1(t)$ was estimated. Likewise, Figure 2.9 shows three $\beta_2(t)$ estimates derived from the second pitch harmonic. Note that Figures 2.8 and 2.9 are visually similar with $\beta(t)$ extremes of approximately -1 to +4.5 samples.

In contrast, Figure 2.10 presents three beta function estimates from an analytic signal generated from a portion of the spectrum in between the first and second pitch harmonics. As the discussion above suggested, the three beta estimates

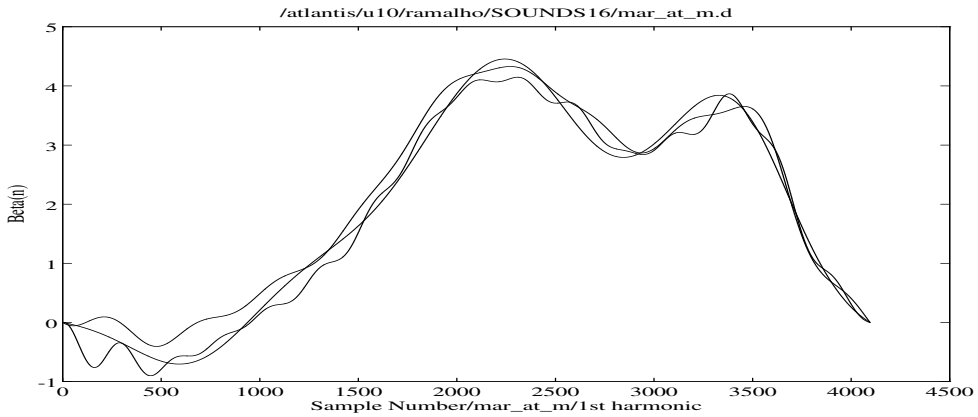

Figure 2.8: Three beta function estimates for first pitch harmonic

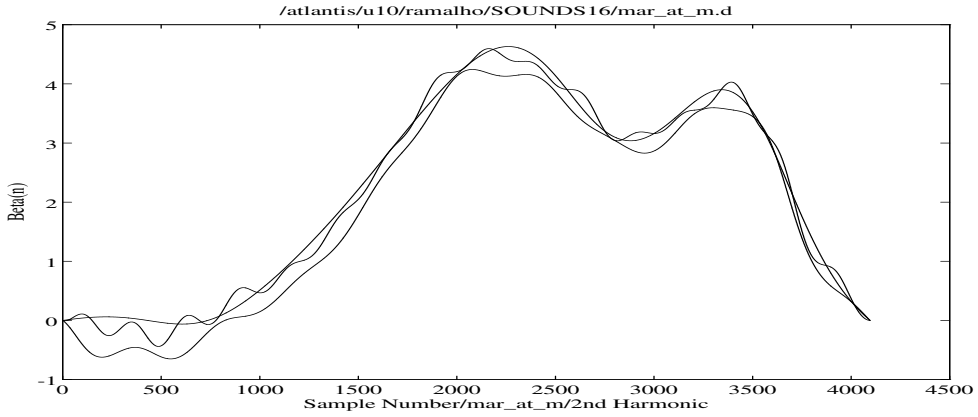

Figure 2.9: Three beta function estimates for the second pitch harmonic

bear no resemblance to each other. They also have extremely large $\beta(t)$ extremes of -110 to +20 samples. This is due to the fact that this portion of the spectrum cannot be demodulated – it does not correspond to a frequency modulated signal.

The $\dot{\phi}(t)$ clipping was *not* used to create the beta estimates of Figure 2.10. If the $\dot{\phi}(t)$ clipping was used, the three estimates would still be dissimilar, but their extremes would be much less (approximately $\pm 10 - 12$ samples).

For robustness to noise, the bandwidths used for the $\phi_i(t)$ estimates need to be adaptive. The peaks should, if possible, be demodulated using the largest possible bandwidth consistent with the noise floor about the peak. The final time

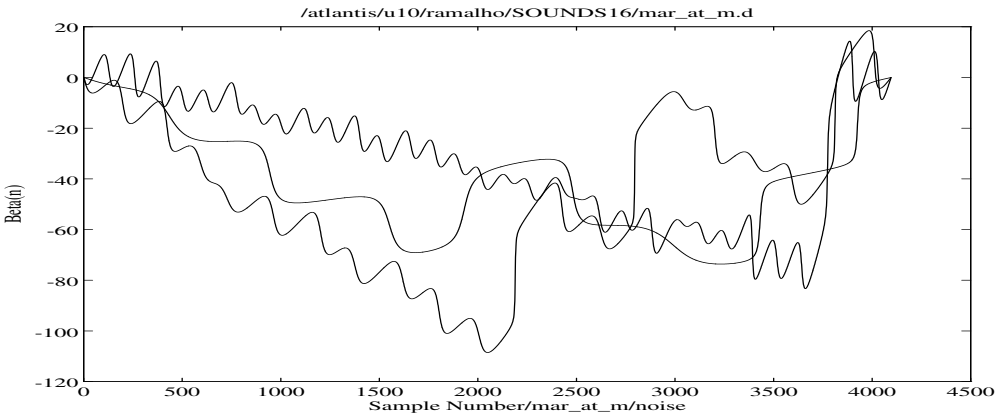

Figure 2.10: Three beta function estimates where there is not a pitch harmonic

warping estimate for a given quasi-harmonic i, $\beta_i(t)$, is obtained by averaging three $\beta_i(t)$ estimates using different bandwidths for the bandpass filters. The largest bandwidth used, in number of FFT frequency bins, is taken to be $2L$, where L is the first local minimum of:

$$BW(L,k) = \left[\left(\frac{1}{psd(k)} \sum_{l=-L}^{L} psd(k-l) \right) - 1 \right] / 2L \qquad (2.27)$$

as L is increased from one, k is the index of the FFT bin peak being considered, and $psd(\cdot)$ is the magnitude of a given FFT bin. The two other bandwidths used were the nearest integer approximations of $4L/3$ and $2L/3$. The intent of this algorithm is to find the largest bandwidth that is reasonable to demodulate the signal $s_i(t)$. The metric of Equation 2.27 will continue to decrease for as long as the FFT magnitude continues to monotonically decrease from the peak value. For well separated peaks with high peak to noise ratio, this algorithm uses a large maximum bandwidth, and the converse is true for small peaks and high noise.

2.5.2 Time-Warping Function $\beta_i(t)$ Clustering

The PMMM exploits the fact that a subset of independent $\beta_i(t)$ estimates from each quasi-harmonic can be intelligently combined to provide a robust estimate of

time warping function. The procedure used, as discussed above, does not assume any *a priori* knowledge about whether a Fourier domain peak corresponds to a pitch harmonic or noise. The underlying premise is that, if we obtain a time warping estimate from every possible peak, that the voiced harmonic components from a given utterance will tend to cluster about an "average' or representative $\beta(t)$. In this section, we discuss the clustering method used to obtain the representative $\beta(t)$.

A K-Means clustering algorithm [52] was designed and slightly modified to obtain a representative $\beta(t)$ from the $\beta_i(t)$'s. After convergence of the K-Means algorithm, the centroid of the cluster with the most elements is chosen to represent $\beta(t)$. To further improve the estimate, a subset of M of the most consistent time warping functions is selected from this cluster. The centroid of these M time warping functions is then calculated and is used as the representative time warping function, $\beta(t)$. The precise manner in which this is done is specified in the $\beta(k)$ pseudocode of Figure 2.11.

The $S_i(t)$'s that correspond to the subset of M time warping functions are, with high confidence, a proper subset of the desired speakers N components. The average frequencies of those M quasi-harmonics, $\bar{\omega}_i$, are found to be harmonically related to a very high accuracy. An algorithm has been designed that computes the common integral factor of these M component frequencies, and it provides an extremely accurate pitch estimate. This algorithm is described in detail in the following section.

At this point some observations concerning this clustering algorithm and its design should be noted. If all the $\beta_i(n)$ estimates that came from FFT spectra in which the quasi-harmonics were well separated and only quasi-harmonic peaks were demodulated, we would have a case like Figure 2.3. A simple "1-class" clustering would suffice. However, as discussed above, the PMMM decomposition has, as a goal, *not* to assume any a priori information concerning the signal. For example, if one knew the pitch, then perhaps only the peaks corresponding to

Perform the classical K-Means algorithm with $k = 3$ on all time warping estimates, $\beta_i(k)$. Initialize two of the three classes with the $\beta_i(t)$ from the most positive and from the most negative extremes. Initialize the remaining class with a vector of all zeros. Iterate until convergence is obtained.

The class with the most elements in it is chosen to be the representative $\beta(k)$ class.

Compute a class mean, $\hat{\beta}(k)$, from the members in this class.

Compute a vector $\hat{\mathbf{d}}$ that has the $\mathcal{L}_1$ distances of each $\beta_i(k)$ in the representative $\beta(k)$ class to the class mean (i.e., the elements of $\hat{\mathbf{d}}$, $\hat{d}(j)$, are $\hat{d}(j) = \sum_{k_{\text{begin}}}^{k_{\text{end}}} |\beta_j(k) - \hat{\beta}(k)|$).

Form a vector $\mathbf{d}$ by sorting, in ascending distance, the elements in $\hat{\mathbf{d}}$. Keep track of the indices i that the distances in $\mathbf{d}$ are associated with. Let M equal the number of elements in $\mathbf{d}$.

Compute the quantity $\frac{2}{(M-1)} \sum_{l=1}^{(M-1)} d(l)$.

If $\frac{2}{(M-1)} \sum_{l=1}^{(M-1)} d(l)$ is less than $d(M)$, let $M = M - 1$ and repeat the previous step. If not continue.

$\beta(k)$ is taken as the average of the $\beta_i(k)$ that are associated with the distances $d(l)$, $l = 1, ...M$.

Figure 2.11: Pseudocode for $\beta(k)$ Clustering Algorithm

multiples of this pitch frequency would be demodulated. Instead, the algorithm has been designed simply to sift the Fourier domain magnitude for peaks, demodulate these peaks, and obtain a $\beta_i(k)$ from each peak. The rationale for the decision to demodulate all *potential* quasi-harmonic peaks will be made clear later in the discussion to follow concerning robustness to noise. Using this methodology, some $\beta_i(k)$ estimates are allowed to be "bad" in the sense that the FFT peaks that they correspond to are unwanted signals or spurious noise peaks.

The clustering technique above was designed to accommodate these unwanted $\beta_i(k)$'s, while obtaining the wanted representative $\beta_i(k)$. We now note that Figure 2.10 did *not* use the $\dot{\phi}_i(t)$ clipping strategy outlined above. As a result the $\beta(k)$ extremes were -110 to $+20$ samples – well beyond the actual $\beta(k)$ extremes of -1 to $+5$ for the utterance considered. The $\dot{\phi}_i(t)$ clipping strategy was designed to limit the $\beta_i(t)$ extremes for these noise peaks to something bounded by two or three times the extremes of the most dynamic $\beta(k)$'s found in practice. The $\beta_i(k)$ so formed would: 1) meet the constraint $\beta_i(k_{\text{begin}}) = 0 = \beta_i(k_{\text{end}})$ necessary for proper clustering, and 2) produce an "unwanted" $\beta_i(k)$ whose positive and negative going extremes were bounded as described above. Knowing this, the operation of the K-Means algorithm is now made clear. When a peak other than a quasi-harmonic peak is demodulated, the unwanted $\beta_i(k)$ is unrelated to the wanted $\beta(k)$, but is nonetheless bounded within a reasonable distance from the wanted representative $\beta(t)$. The representative $\beta(k)$ class (the cluster with the "wanted" time warping function estimates in it) is usually the "middle" $\beta(k)$ cluster.

Figure 2.12 presents an example of this beta clustering strategy. This figure presents the beta function estimates for all the pitch peaks of a 2048 sample segment of the utterance "How are you?", sampled at 16 kHz. Previous to generating these estimates, the signal had been iteratively de-warped ten times to make the utterance as periodic as possible. This allows for *all 29* of the pitch peaks from 0 to 3500 Hz to be demodulated. Some of the pitch peaks demodulated have signal

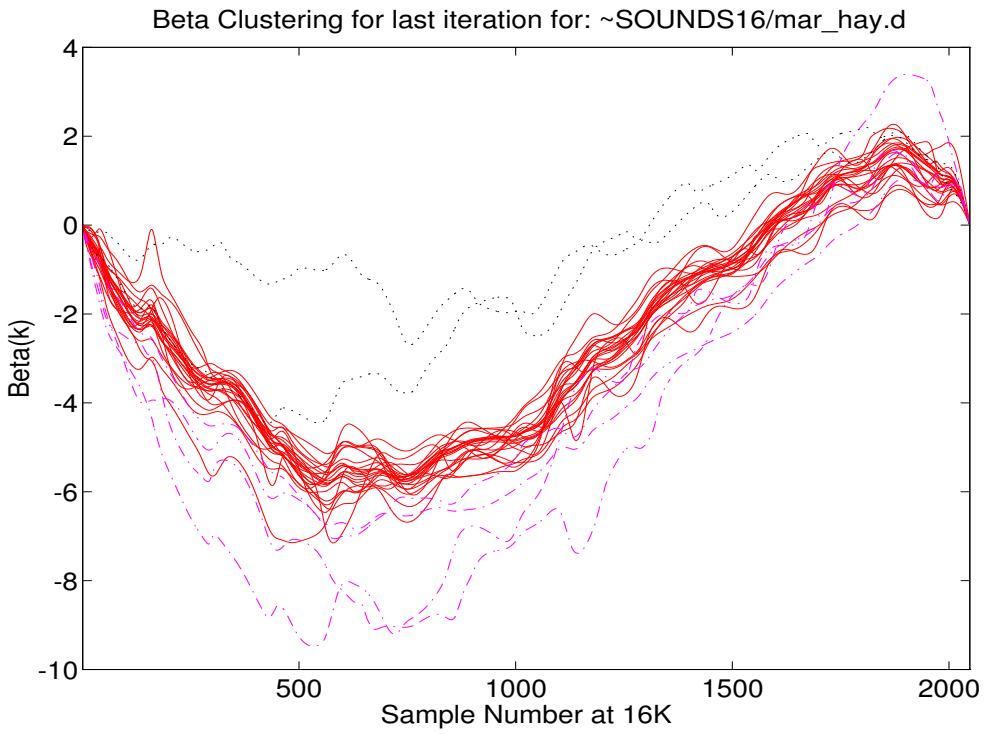

Figure 2.12: Beta clustering example for 2048 sample segment of "How are you?"

power 40 dB down from the peak signal. As a result, some beta function estimates are impaired by the noise, resulting in a noisy beta function estimate. The beta functions have been classified into three classes, the beta functions depicted in these classes are plotted with different line characteristics. For example, the class with the largest positive beta extremes are plotted as dotted lines. Clearly seen is the "middle" cluster of consistent beta estimates. The final beta function estimate used for this utterance is the average of the M beta functions from the algorithm above. For this particular case, M was equal to 9 and the $\beta_i(k)$'s corresponded to harmonic numbers 7, 8, 9, 15, 16, 24, 25, 27, and 28 of the 29 possible harmonics in the 0 to 3500 Hz range.

2.5.3 PMMM-Based Fundamental Frequency Estimation

This section describes the algorithm used to determine the Fundamental, or pitch, frequency for the frame j, $\bar{\omega}_{1_j}$, from the M time warping functions used in the

$\beta(t)$ clustering algorithm of the previous section. This estimate is an average fundamental frequency estimate for the entire *frame* (i.e., for $t \in (t_a, t_b]$) - it is not an instantaneous pitch estimate. A PMMM-based *instantaneous* pitch estimate can be formalized by taking the derivative of the phase function of the PMMM model of the first harmonic component, $\phi_{1_j}(t) = \bar{\omega}_{1_j}[t + \beta(t) + \Theta_1]$, resulting in $\omega_{1_j}(t) = \bar{\omega}_{1_j}[1 + \frac{d\beta_j(t)}{dt}]$. Discussion of this instantaneous pitch estimate is deferred until Section 6.3.

Experimentation with the M average frequency estimates, $\bar{\omega}_i$, has shown that if $\bar{\omega}_A$ is the average radian frequency of the A^{th} harmonic component and $\bar{\omega}_B$ is that of the B^{th} harmonic, then:

$$\frac{\bar{\omega}_A}{A} = \frac{\bar{\omega}_B}{B} \pm X\% \tag{2.28}$$

where X is dependent on the frame size. X is approximately 1% for 32 millisecond frame sizes and decreases (improves) for larger frames. For a wholly voiced 256 millisecond utterance to be analyzed in Section 1.6, X was less than 0.1%.

The fundamental frequency estimation algorithm associates a harmonic number, H_i, with each of the M average frequency estimates. Then M fundamental frequency estimates are obtained by dividing the average frequencies by their associated harmonic numbers, that is $\bar{\omega}_i/H_i$. Since the frequencies are harmonically related, these M estimates should be identical. In practice, most, if not all, of the M estimates will be close to the true fundamental frequency. To protect biasing the final fundamental frequency estimate from one or more inaccurate ones, pitch estimates more than one standard deviation from the median of the estimates are eliminated. The final average fundamental frequency is chosen as the average of the remaining estimates.

The most challenging part of the above is to associate a harmonic number, H_i, with each on the M average frequencies, $\bar{\omega}_i$. This is done via a voting procedure whereby each of the M estimates are individually compared to the remaining

$M-1$ estimates to generate an intelligent guess for the harmonic number H_i.

Figure 2.13 presents pseudocode for the algorithm. The variables $FREQ_{MAX}$ and $FREQ_{MIN}$ are the maximum and minimum expected pitch frequency, respectively, for the signal. In the absence of any *a priori* pitch information, these are initialized to 270 and 70 Hz, respectively.

The algorithm of Figure 2.13 uses a rational approximation function call to make the harmonic number association. The algorithm adjusts an internal variable TOL if the harmonic number association would result in a fundamental frequency outside of the range ($FREQ_{MIN}$, $FREQ_{MAX}$). These frequency ranges can be set relatively large because, in practice, an incorrect harmonic assumption rarely occurs. This is because the M frequency estimates $\bar{\omega}_i$ are, with high certainty, a subset of the true harmonic frequencies. The exception cases are in the algorithm mainly to guard against a grossly incorrect harmonic number association with $\bar{\omega}_i$. It is important to note that with 70 and 270 Hz as the allowable pitch ranges this algorithm will correctly estimate the pitch of all but the lowest pitch males and highest pitch females. With additional *a priori* pitch information for the population considered, the maximum and minimum pitch range can be modified.

A voting procedure is used in the harmonic association algorithm because it is sometimes impossible to tell what the fundamental is from the ratio of two of the $\bar{\omega}_i$'s. For example, suppose we are attempting to make a harmonic association of the second harmonic by comparing it to the first and fourth harmonic. Performing the ratio test in the pseudocode for the second harmonic and comparing it to the first results in $\frac{\approx 1\bar{\omega}_{1_j}}{\approx 2\bar{\omega}_{1_j}} \approx \frac{1}{2} = \frac{A}{H}$, yielding the correct harmonic association of $H = 2$. Performing the ratio test in the pseudocode for the second harmonic and comparing it to the fourth results in $\frac{\approx 4\bar{\omega}_{1_j}}{\approx 2\bar{\omega}_{1_j}} \approx \frac{2}{1} = \frac{A}{H}$, yielding an *incorrect* association of $H = 1$. When the $\bar{\omega}_i$'s happen to correspond to harmonic numbers which are prime numbers, the harmonic association will be correct. Recognizing that this method is correct most of the time (since not all $(M-1)$ harmonic

FOR $i = 1, 2, ...M$

Set $H_{MIN} = ceil\,[\omega_i/(2\pi FREQ_{MAX})]$, the smallest harmonic number to consider for $\bar{\omega}_i$.

Set $H_{MAX} = floor\,[\omega_i/(2\pi FREQ_{MIN})]$ to be the largest harmonic number to consider for $\bar{\omega}_i$.

FOR $j = 1, 2, ...M$ and $j \neq i$

Set TOL, a tolerance for the ratio test below, based on frame size and $\bar{\omega}_i$.

Perform RATIO TEST, finding smallest integers A and H satisfying: $\frac{A}{H} = \frac{\bar{\omega}_j}{\bar{\omega}_i} + \epsilon$, where ϵ may take any value in $[-TOL, +TOL]$.

IF $H < H_{MIN}$

Adjust TOL downward ($TOL = TOL/1.5$). Go back to Ratio Test.

ELSEIF $H > H_{MAX}$

Adjust TOL upward ($TOL = 1.5\,TOL$). Go back to Ratio Test.

ELSE

Vote for H as the harmonic number for $\bar{\omega}_i$.

END (If $H < ...$)

END (For $j = ...$)

Set H_i, the final harmonic estimate for $\bar{\omega}_i$, equal to the harmonic number with the largest number of votes. If two or more harmonic numbers possess the same number of votes, choose the smallest harmonic number.

END (For $i = ...$)

Figure 2.13: Harmonic Number Association Pseudocode for $\bar{\omega}_i$

number estimates for a given $\bar{\omega}_i$ will be related by a confusing ratio), the voting procedure outlined in the pseudocode is acceptable.

Note that the frame fundamental frequency estimate $\bar{\omega}_1$ is *not* used for calculating $\beta(t)$, but rather it is obtained *after* clustering of the estimated time warping functions. This marks a significant difference to other speech parameterization techniques that use harmonic models which require accurate robust pitch estimation *prior* to processing the speech. Because of its significance, this point is discussed in detail in Chapter 3, when these other speech models are compared to the PMMM.

2.5.4 Iterative Parameter Estimation

The iterative PMMM parameter estimation methodology described here allows for an iterative improvement of the $\beta(t)$ estimate. As previously mentioned, an iterative technique for obtaining $\beta_i(t)$ estimates is required for the higher order quasi-harmonics if their spectra overlap significantly in frequency.

The strategy used is that outlined in Section 2.4.3. That is, a $\beta(t)$ is estimated from the components that can be demodulated and relation $\tau = [t + \beta(t)]$ is applied to de-warp $s(t)$ to become more periodic. After this de-warping, the pitch harmonics should have become more sharpened (their deviation from their harmonic carrier frequencies should be lessened) and additional pitch peaks may now be able to be demodulated. A $\delta\beta(t)$ is found and applied to make the signal even more periodic. This process is repeated until all pitch harmonics are of uniform bandwidth. A block diagram of this processing is shown as the solid blocks and lines of Figure 2.14. The solid lines and blocks of Figure 2.14 depict the iterative procedure to obtain the necessary PMMM parameters for a given frame: the time-warping function, $\beta(k)$, and the average fundamental frequency for a particular frame j, $\bar{\omega}_{1_j}$.

As previously mentioned, the PMMM algorithm has been "robustized" to handle arbitrary noise scenarios. In the following subsection, a detailed analysis of

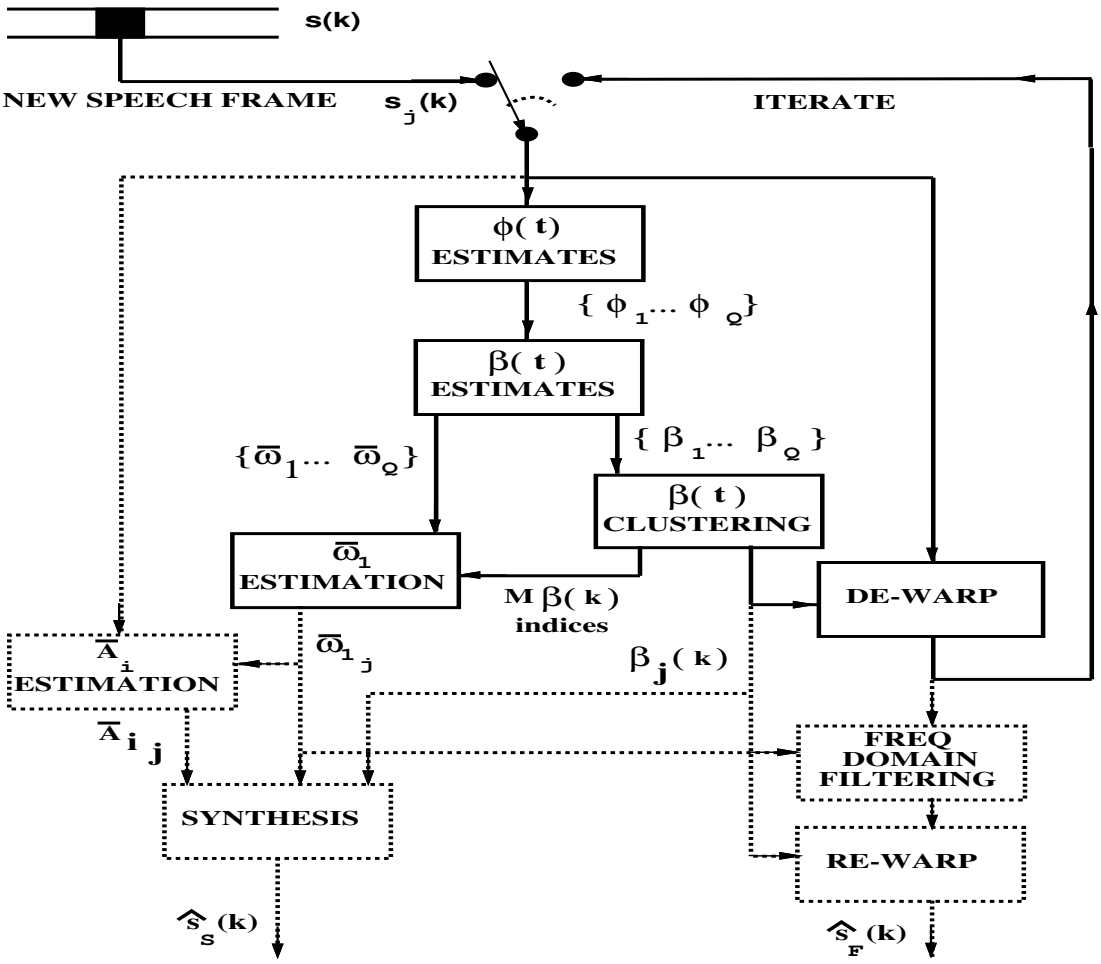

Figure 2.14: PMMM Iterative Parameter Estimation Block Diagram

the iterative parameter estimation algorithm is presented and rationale for the signal processing specifics is given.

2.5.5 Iterative Parameter Estimation Algorithm: Detailed Analysis

Specifically, the speech is first separated into frames of X samples and involves the processing steps outlined in the pseudocode of Figure 2.15 for a particular frame j.

Since the $\beta(t)$ estimation is iterative, it is not required to converge to the

[10] X is chosen to be a power of two. Recall from Section 2.5.1 that the $\phi_i(k)$ estimates from which the $\beta_i(k)$ are obtained are the center X samples of a $2X$-sample Hamming-windowed FFT.

[11] The uniform time base sample index is denoted by τ_k.

L = NUMBER OF ITERATIONS.

FOR $l = 1, 2, ...L$

$f_{MAX} = f_{BEGIN} + (l-1) f_{INCREMENT}$.

f_{MAX} = MIN $[f_{MAX}, f_{CUTOFF}]$.

Calculate $\beta_i(k)$ for each observable Fourier domain peak in frequency range 0 to f_{MAX} Hz using algorithm of Section 2.5.1[10]

IF $l = 1$ (the first iteration)

Initialize the 3 clusters as per Section 2.5.2.

ELSE (not the first iteration)

Initialize the 3 classes with the representative classes of previous iteration.

END (IF $l = 1$)

Cluster the $\beta_i(k)$'s to obtain the representative $\beta(k)$ using algorithm of Section 2.5.2.

Obtain a candidate pitch frequency for iteration l, $\tilde{\omega}_{1_l}$, by finding the fundamental frequency of the $\bar{\omega}_i$'s corresponding to the subset of the M estimates of $\beta(k)$'s in the clustering process using algorithm of Section 2.5.3.

De-warp $s(k)$ with the relation $\tau_k = [k + \beta(k)]$[11]

END (FOR $l = ...$)

Calculate a final fundamental frequency estimate, $\bar{\omega}_{1_j}$, for frame j based on the $\tilde{\omega}_{1_l}, l = 1, ...L$.

Figure 2.15: Iterative Parameter Estimation Pseudocode

optimal $\beta(t)$ in one step. With clean speech, it is often possible for the iterative process to converge in no greater than three iterations. However, as the processing is intended to perform well in noise, the procedure has been designed to purposely slow the convergence process in order to help assure the correct $\beta(t)$ result. In experimentation with very noisy signals (SNRs approaching 0 dB), it has been found that it is extremely important for the first time warping estimate to warp the signal in "the correct direction". By this we mean a time warping such that the signal has been made, at least marginally, more periodic. If this doesn't happen, the signal will become less periodic and subsequent iterations *may* not be able to recover from the additional warping.

This has been done algorithmically by limiting the frequency range that the peak picking algorithm demodulates $\phi_i(t)$'s from. In our implementation, this range is from 0 to f_{BEGIN}, where f_{BEGIN} is set to 1800 Hz. Said another way, the first time warping estimate only uses information from 1 to 1800 Hz to obtain the first $\beta(t)$[12].

After the important first iteration, the frequency domain region that the $\beta(t)$ are estimated from is increased by $f_{INCREMENT}$ until a maximum search frequency of f_{CUTOFF} is reached. These frequencies are set to 350 and 3700 Hz, respectively, for the 8 kHz sampled signals. Therefore, the iterative procedure is completed in six or seven iterations, instead of the two or three needed for the no noise case.

After all iterations are complete, a final fundamental frequency estimate, $\bar{\omega}_{1_j}$ is obtained from the individual iteration estimates $\tilde{\omega}_{1_l}$ in a manner similar to the algorithm of Section 2.5.3. To protect biasing the final estimate from one or more inaccurate ones, estimates more than one standard deviation from the median of all of the iteration estimates are eliminated. The final average fundamental frequency is chosen as the median of the remaining estimates.

[12]Recent auditory research suggests that the low frequency region below 1500 Hz is used to establish time synchrony in humans. More on this in Section 4.2.

The de-wapring procedure, in discrete time, is an interpolation processes. The de-warping procedure interpolates new signal samples at the new (discrete time) uniform time sample points τ_k. Computer simulations have shown that an interpolation "granularity" to the nearest $1/50^{\text{th}}$ of a sample yields an interpolation error (after warping) which is flat and no less than 60 dB below the peak signal magnitude. Many such iterative warpings, as inferred by Figure 2.14, would lead to an unacceptable error because the warping errors are cumulative. In practice, the de-warping is accomplished by warping the *original* speech frame by the combination (i.e., sum) of all the warpings to the present iteration. This procedure, with $1/50^{\text{th}}$ interpolation granularity, insures that all interpolation errors are insignificant.

If the application requires the harmonic amplitudes, $a_i(k)$, of the PMMM to be calculated, an analytic signal needs to be formed from the de-warped signal, but in a slightly different manner. Recall that when obtaining the phase estimates, $\phi_i(k)$, the speech signal was windowed with a Hamming window. This window was experimentally found to yield superior phase estimates than other windows tried. For good amplitude estimates for the center X samples of the analytic signal, a window with a value of unity for these X samples is needed [157]. Experimentation has shown good results with a Tukey window. Thus, for $a_i(k)$ estimation, analytic signals about the harmonics of $\bar{\omega}_{1_j}$ are formed from the speech signal after it has been de-warped and windowed by an appropriate Tukey window. The PMMM-based speech synthesis to be described later uses this $a_i(k)$ estimation technique.

2.6 PMMM: Ability to Model Long-Term Speech Non-Stationarity

The Pitch Mode Modulation Model successfully models long-term speech non-stationarity in a manner heretofore unreported in the literature. A detailed comparison of the PMMM to these other models is deferred to Sections 3.3.1 and 1.6. The two significant differences to other models that afford it significant advantage

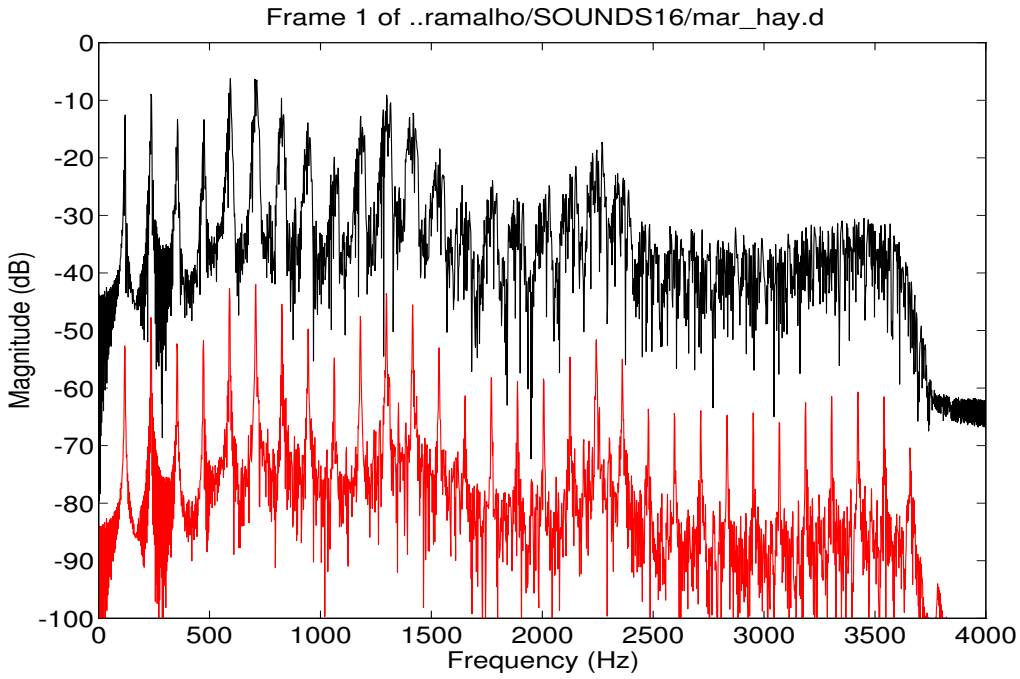

Figure 2.16: "How are you": Spectra of Original and De-Warped Spectra (De-Warped plotted 40 dB down)

are that it: 1) does not require pitch information to perform the necessary time de-warping and, most importantly, 2) it does not assume a specific parametric form for the time warping function $\beta(t)$ – just that it be narrowband.

To demonstrate the capabilities of the PMMM and its estimation scheme, a 4096 sample portion of the utterance "How *are* you?" was input to the PMMM processing described in Section 2.5.4. The portion of the utterance in *italics* was contained in the 4096 samples, and it was sampled at 16 kHz, resulting in slightly over 1/4 second of voiced speech. The utterance was spoken naturally, with no effort to maintain pitch or speaking timbre.

Figure 2.16 presents two spectra; the top spectra is from the original 4096 speech sample, and the bottom is from the same 4096 sample of speech after the PMMM de-warping algorithm of Section 2.5.4 was performed. The top spectra is gain normalized such that its maximum value is 0 dB and the bottom spectra is plotted 40 dB below the top spectra for comparison purposes. Clearly seen is

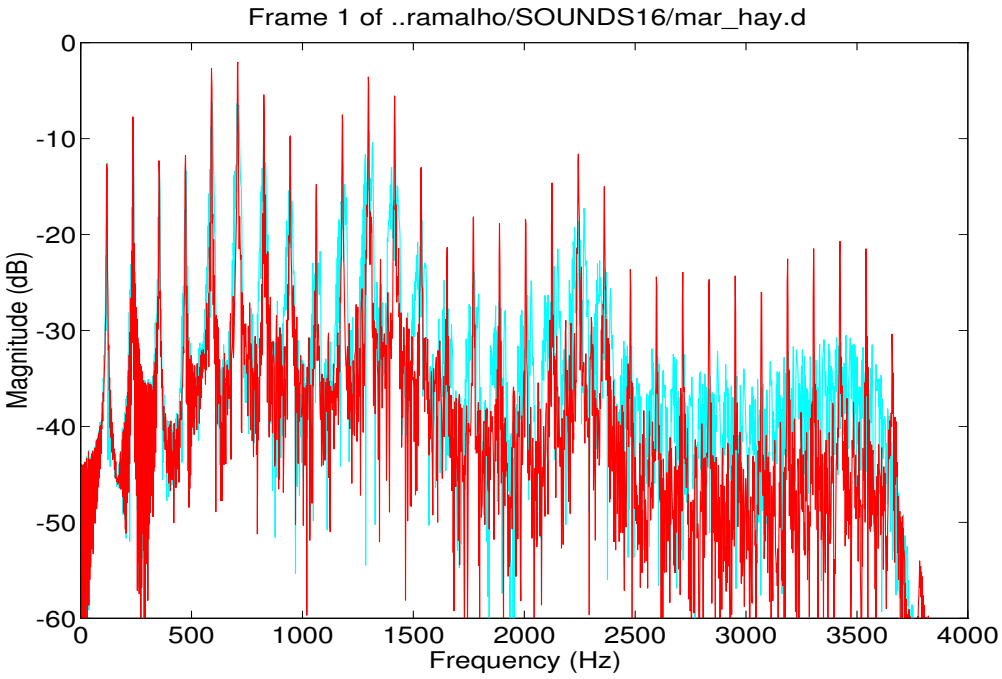

Figure 2.17: "How are you?": Original (grey) and De-Warped (black) Spectra

the sharpening of the pitch harmonics - the energy in both spectra are identical. Consistent with the explanation in Section 2.4.3, the pitch harmonic magnitude has increased. To see this more clearly, Figure 2.17 plots both original and warped spectra on one plot.

Figure 2.17 demonstrates that the quasi-harmonic magnitude has increased by approximately 10 dB for the higher order quasi-harmonics. It should be noted that if broadband noise was warped by the $\beta(t)$ used to de-warp the speech, that the noise spectra would appear essentially similar to the unwarped noise spectra. Thus, the amplitudes of the quasi-harmonics are "pulled" out of the noise and can be much more accurately estimated. Using the PMMM-based demodulation technique, unobservable quasi-harmonics that are "buried" or obscured by the noise can now be seen and estimated.

These plots demonstrate that the broadening of the higher order pitch harmonics is primarily due to pitch non-stationarity (i.e., not the amplitude non-stationarity). This justifies a similar statement made in Section 2.3.1, whose

proof was deferred until here. In fact, the higher order pitch harmonics are not discernible from the un-warped spectra. The iterative procedure of Section 2.5.4 was required to demodulate this signal. The fact that the phase demodulation of the quasi-harmonics is successful for voiced utterances is proof of the PMMM theory presented earlier.

Also note that two additional pitch peaks are created in the range *above* the lowpass cutoff frequency of approximately 3600 Hz. These peaks are created from harmonic energies that had been modulated *below* 3600 Hz, but were rewarped to their present location! There is no other reason for these peaks to occur. This bring out the fact that pitch harmonics can be modulated *past* the Nyquist frequency. This energy cannot be recovered. In addition energy from pitch harmonics above the Nyquist frequency may be modulated into the lowpass filtered bandwidth. Both of these effects make it unwise to attempt to demodulate pitch peaks close to the Nyquest rate[13].

It is instructive to note here that the vocal tract is not stationary over the 256 milliseconds of this utterance. The amplitude $a_i(t)$ of each of the components is "unconstrained" by the time-warping procedure applied to the signal. The author is unaware of any long-term speech model that has successfully accomplished the de-warping indicated in Figure 2.17.

The PMMM de-warping is accomplished independent of the quasi-harmonic component magnitude and does *not* require a specific parametric form for the $\beta(t)$ de-warping. These features mark a significant difference to other speech models that attempt to model long-term speech non-stationarity. A detailed comparison of the PMMM parameterization to these other models that model speech non-stationarity is deferred to Sections 3.3.1 and 1.6.

---

[13] In addition, the presence of noise is known to cause difficulty for *discrete-time* instantaneous frequency computations via analytic signals. Reference [190] thoroughly discusses this topic and suggests oversampling by a factor of 2 for non-stationary signals.

2.7 PMMM: Exploiting Redundancy in Voiced Speech

The speech modeling and enhancement techniques developed here exploit the fact that independent estimates of the time-warping function $\beta_i(t)$ can be estimated from every pitch quasi-harmonic. This important finding implies that there is about a 25-fold[14] redundancy contained within telephone quality voiced speech (4 kHz bandwidth) which heretofore has not been exploited in speech signal processing.

The relation $\phi_i(t) \approx i\,\bar{\omega}_1[t + \beta(t) + \Theta_i]$ applies to *clean* speech and essentially states that $\beta_i(t) \approx \beta(t)$ for *all* i. The techniques described in Section 2.5.2 to calculate a representative $\beta(t)$ from all $\beta_i(t)$ that allow for the fact that *some* $\beta_i(t)$ generated could be inaccurate and should not be used in finding a $\beta(t)$ representative for the time segment considered. A specific $\beta_i(t)$ could be inaccurate if, for instance, the speech signal in the frequency range that the i^{th} quasi-harmonic was extracted from was somehow contaminated. In this case, the $\beta_i(t)$ estimate should not be used in obtaining a representative $\beta(t)$ (i.e., that $\beta_i(t) \not\approx \beta(t)$ for that particular i).

The time warping function estimation technique, described in Section 2.5.1, takes advantage of the redundancy in voiced speech by using intelligent clustering techniques to obtain a robust, representative $\beta(t)$. This time warping function estimation strategy is very important in real-world speech applications in which the noise and interference environment is unknown. For example, if narrowband noise or other frequency selective impairment (e.g., narrowband fading on radio channels) contaminates the speech signal, a representative $\beta(t)$ can still be obtained using the above methodology. This redundancy is exploited in this thesis to help restore noisy or otherwise contaminated speech [155]. The author knows of no other research that takes advantage of the $\beta(t)$ "time-synchrony" redundancy contained in voiced speech in this manner.

---

[14] There are approximately 25 pitch harmonics in 0 to 4 kHz bandwidth speech.

The ability of this technique to obtain $\beta(t)$ in the presence of realistic noise marks a significant difference to other, independently derived, speech research which use time-warping principles to make quasi-periodic pitch harmonics periodic [56, 23, 188]. If, for example, narrowband noise or other frequency selective impairment contaminate the speech signal, an accurate estimate of $\beta(t)$ and the pitch are still obtained. As a result, the pitch estimate obtained has been found to be robust to *arbitrary* noise impairment. This robustness of the pitch estimate is significant, as the performance of many speech coding [108] and enhancement techniques [130] are critically dependent on accurate fundamental frequency tracking. Also note that the $\beta(t)$ processing above allows for the identification of pure tones since their $\beta_i(t) = 0$. Identifying pure tones (for their removal) is also important for real world applications [74] as many signal processing techniques operate on the principle of minimizing, in some sense, the energy of an error signal. The presence of unrelated signal energy (e.g., a tone) is known to degrade the performance of these algorithms [2, 105, 188, 23].

This clustering or "grouping" of quasi-harmonics with the same "time-synchrony" may, in fact, be the same fundamental process that the auditory system uses to fuse portions of harmonically related sounds together. This conjecture is supported by both the auditory fusion principles and the cochlea modeling information to be presented in Section 4.1.

This research has experimentally shown that the PMMM decomposition also models so called mixed-voiced portions of speech well [15]. This is due to the successful demodulation of the higher pitch quasi-harmonics via the nonlinear mapping $\tau = g(t) = [t + \beta(t)]$. In performing this demodulation successfully, a processing gain is obtained which is similar to that specifically designed into Spread Spectrum communications systems. This processing gain and the analogy to Spread Spectrum communications systems is explored later in Section 2.8.

[15] Mixed-voiced portions of speech have Short-Time Fourier Spectrums with obvious harmonic structure for the low harmonics and a wideband noise-like structure for higher frequencies [58, 59].

Additionally, PMMM decomposition also allows a speech harmonic coding/synthesis technique in which separate models for unvoiced and mixed-voice sounds are not needed [58, 59]. This application, along with trivial time-scale and pitch-scale modification of speech, is also discussed in Chapter 6.

2.8 PMMM Processing Gain: A Spread Spectrum Analogy

The de-warping example presented in the last section demonstrated that the harmonic magnitudes increase as the quasi-harmonic energy becomes concentrated in a narrow frequency range. The small remaining bandwidths about the peaks are primarily due to the amplitude variation of the components[16].

The analysis of Section 2.4.1 implies that if something other than the correct $\beta(t)$ is used to time-warp the speech signal, one *should* expect an additional modulation of the quasi-harmonics. The effect would be to further broaden and lower the amplitude of the higher order quasi-harmonics. This effect is diagrammatically illustrated for two speakers in Figure 2.18.

According to PMMM theory, the speech of speaker 1, s_1, can be modeled as a time-warped version of speech which has an approximate line spectra. This effect is illustrated by the spectra of $\tilde{s}_1$ in the figure. Likewise, the speech of speaker 2, s_2, can also be modeled as a time-warped version of speech modeled as an approximate line spectra. The important point to note is that, in general, the time warping of the first speaker, $\beta_1(t)$, is different than that for the second speaker, $\beta_2(t)$. *If* one could find the time-warping function $\beta_1(t)$ needed to de-warp s_1, this time-warping function will, in general, further modulate the second speakers quasi-harmonics. Unwarping the combined speech, $s_1 + s_2$, with $\beta_1(t)$ "spreads" the harmonic energy of s_2 and compresses the harmonic energy of speaker 1, thus providing a method for more precise estimation of the PMMM parameters of speaker 1.

[16] The other dominant effects are the window used and the fact that the FFT length was not constrained to encompass an integer number of pitch periods.

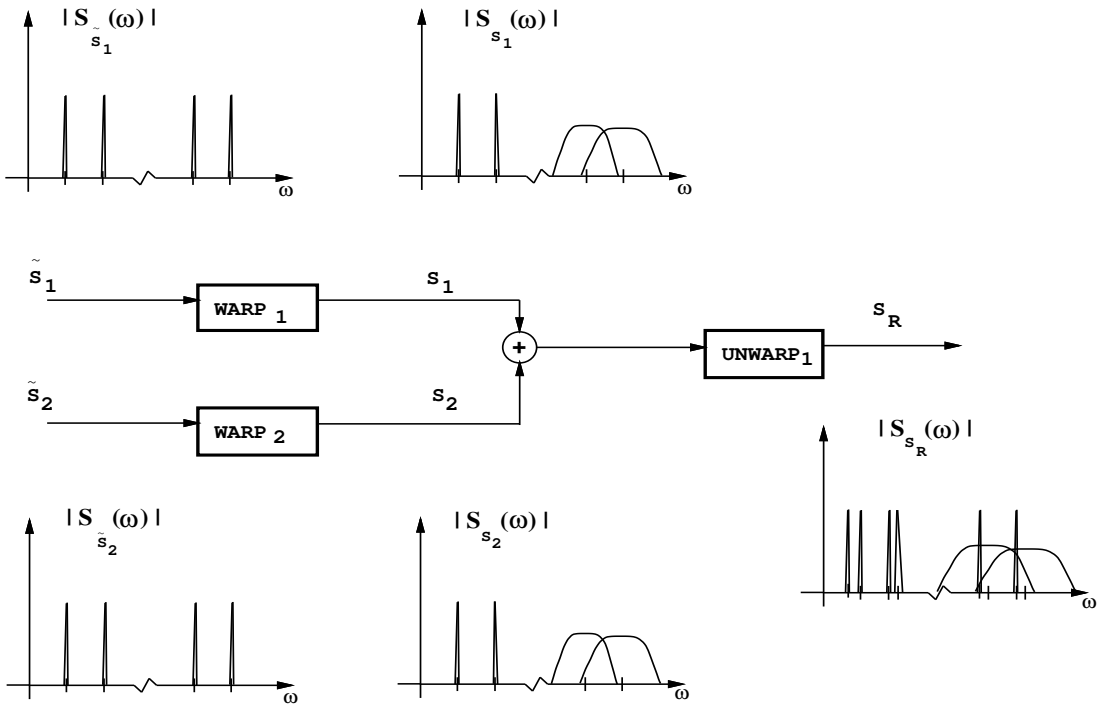

Figure 2.18: PMMM Spread Spectrum Analogy

The same basic principle has been used in Spread Spectrum communications systems [31, 142]. In these systems, the transmitted signal is spread over such a wide bandwidth that its power in any particular frequency band is negligible. The spread spectrum signal appears to any other communications system using a small portion of the bandwidth simply as noise. When the demodulator, knowing the manner in which the signal was spread, demodulates the signal, the signal is "compressed" into its original form.

Spread spectrum techniques have been developed for two primary reasons. First, the systems were developed as a form of private (i.e., secret) communications. Without knowing the exact form of the modulation, demodulating the signal is extremely difficult or impossible. Because the effect of the spread spectrum signal on other users of a portion of this bandwidth is that of a low-level noise source, the presence of the spread spectrum signal may not be apparent to the other users. The second reason for spread spectrum development is that

the modulated signal has a high immunity to interference. The redundancy introduced by the spread spectrum processing allows for the reception of a high quality demodulated signal despite impairments in the many frequency bands the modulated signal had occupied. For military applications, these properties are extremely important as an adversary will have a difficult time interrupting (e.g., jamming) the signal unless it knows or can figure out the "key" to the modulation [142].

The coherent, $\beta(t)$ induced, phase modulation that the PMMM exposes possesses these spread spectrum attributes for the higher quasi-harmonics. This type of coherent phase modulation type has not, to the authors knowledge, been reported in Spread Spectrum literature.

The term "processing gain" has been coined in Spread Spectrum communications. For PMMM processing, the term describes how much the unwanted signal amplitude in the Fourier domain has been "spread" in comparison to the signal amplitudes of the wanted components. For PMMM processing of long frames, such as the 256 millisecond "How are you?" utterance of the previous section, we can expect that the higher order pitch harmonics to "whiten" and become broadband. As was shown previously, the processing gain for the high order harmonic components is on the order of 10 dB. For the processing of shorter frames, we can expect this gain to decrease. The reduction in the processing gain is expected for two reasons. First, the unwanted signal is, with respect to the frame length, more stationary, and the result is that the signal will not be as broadband. Secondly, for the wanted signal, the Fourier domain effects of the windowing of these shorter frames result in broader harmonics and a higher uncertainty in the Hilbert transform derived phase. Processing gains to be reported in the results section are on the order of 6 - 8 dB for 32 to 64 millisecond frame sizes.

The point of this section is that although this spread spectrum-like signal processing is possible, it is uncertain whether this mechanism is, in fact, used by the mammalian auditory system. We revisit this issue in Section 4.3 which

discusses PMMM ties to recent research in auditory processing.

2.9 Summary PMMM Attributes

The PMMM is a long-term speech model that is able to model the non-stationarity found in normal voiced and mixed voice speech. The PMMM components are found to vary in time synchrony with a specific time warping function, $\beta(t)$, which causes a coherent phase modulation of the components. Its major attributes as a long term model are:

- No specific parameterization of the time warping function (e.g., a time series) was used as a model constraint in order to estimate $\beta(t)$.
- No specific parameterization of the amplitude of the PMMM components (e.g. constants) was used as a model constraint.
- No sensitivity of the PMMM technique to *a priori* pitch estimates, as none are used.
- The PMMM is the first long-term model able to associate, on the harmonic-level, quasi-harmonic speech components and disassociate them from non-speech components.
- The PMMM is the first mathematical formulation that describes the spectral effect of a coherent time warping of PMMM components using phase modulation principles.
- A robust pitch estimate is obtained *after* PMMM processing, using signal components *known* to be correlated (in time modulation characteristic) to other speech components.
- The PMMM is the first technique to obtain and qualitatively describe a spread spectrum-like processing gain for amplitude estimation of PMMM components.

In the following chapter, many of these attributes will be used to describe the differences between the PMMM and other speech models, including some that attempt to model speech non-stationarity.

Chapter 3

PMMM Comparison to Other Speech Models

The primary advantage of introducing a speech model is to transform the highly redundant speech waveform to a form in which the parameters have a lower bandwidth. The PMMM captures the redundancy in the quasi-periodic nature of the speech signal in the beta function, $\beta(t)$, the average frequency, $\bar{\omega}_1$, and the set of PMMM component amplitudes, the $a_i(t)$'s. We show in Section 6.2 that the $\beta(t)$ can be approximated very parsimoniously by a few low precision coefficients, and that the set of $a_i(t)$'s are indeed low frequency waveforms. Because the PMMM is a long-term speech model, speech synthesis from the PMMM captures the timbre, or voice quality, of voiced and mixed voice speech. Section 6.2 demonstrates that the bit rate of a PMMM-based parameterization is on the order of 3.1 Kbps *before* serious attempts to minimize the bit rate.

Traditional vocoders using classical speech models and operating at these bit rates have had much difficulty in generating natural sounding speech. As these traditional models operate on the principle of the short-term stationarity of speech, modifications of these vocoders were necessary to improve the naturalness of the speech synthesis.

In this section, we review and compare the PMMM speech decomposition to other speech models to contrast the features and benefits of the PMMM as a long-term speech model.

3.1 Vocoders Based on Classical Speech Models

In the classical speech model discussion of Section 1.2, it was noted that a parametric-type speech model usually models both the excitation and the vocal tract separately. The underlying motivation was that the excitation and the system parameters tend to vary slowly in time due to the physical constraints of the vocal tract and its excitation sources.

The introduction of such a speech model allows for the identification of the vocal tract and excitation parameters independently. For synthesis applications, it also allows for the modification of the parameters independently. For example, for the enhancement of speech spoken in a helium-oxygen mixture, a nonlinear frequency warping of the spectral envelope is desired without modifying the excitation parameters [162].

It has long been recognized that although vocoders based on this class of speech models have been quite successful in the synthesis of *intelligible* speech, they have not been useful in synthesizing high quality speech. The poor quality of the synthesized speech is partly due to fundamental limitations of the model and partly due to inaccurate estimation of the model parameters. In addition, classical vocoder performance tends to be talker-sensitive, and fragile in that extraction of source and vocal tract parameters are adversely affected by additive acoustic noise and signal distortion [139]. As a result, these vocoders have not been widely used in applications such as the pitch or time-scale modification of speech [148], speech enhancement [34], or high quality bandwidth compression applications [58].

Improved quality LPC-based vocoders, such as the Residual Excited LPC, Multipulse LPC, and the Code Excited LPC coders extend the traditional models toward a higher quality synthesis. The essential operating properties of these coders is described shortly. They, in part, are seen to artificially model (in the frequency domain) or mechanically introduce a non-stationarity in speech that

the PMMM captures.

There are, of course, other vocoders that are of significance to speech research. Some are: the Channel Vocoder [34] (the first and oldest, implemented by Dudley), the Phase Vocoder, the Cepstral (or Homorphic) Vocoder, and the Formant Vocoder. However, as they do not possess a significant tie to the Pitch Mode Modulation Model, they are not described here.

3.1.1 Classical Spectral Envelope Analysis Methods

Classical approaches to spectral envelope estimation include Linear Prediction [104] (i.e., All-Pole Modeling of Section 1.5.2.2), windowing the cepstrum [126] (i.e., smoothing the log magnitude spectrum), windowing the autocorrelation function [11] (smoothing the magnitude squared spectrum) and linear interpolation of the log magnitude spectrum at multiples of the pitch frequency [139]. In all the approaches, the pitch period often interferes with the spectral envelope estimation procedure. For example, for speech frames with high F_0, the widely spaced pitch harmonic locations in the spectrum tend to cause pole locations and bandwidths to be poorly estimated by linear predictive methods. Methods that window the cepstrum or autocorrelation function obtain a poor envelope estimate for high F_0 frames due to interference of the peak at the fundamental with the spectral envelope information on the low time portion of the cepstrum or autocorrelation function. The linear interpolation method suffers from a fewer number of frequency points in the interpolation.

These methods are limited by *both* the assumption of the short-term stationarity of the vocal tract and by the excitation of the tract at multiples of F_0. These methods result in a system identification problem where the vocal tract is modeled by a few resonances in the frequency domain of interest (typically four in 4 kHz bandlimited speech), stationary for only a short-time, and excited at a relatively few frequencies.

The PMMM model, on the other hand, makes no assumption concerning amplitude stationarity. The amplitude modulation of the individual PMMM components, the $a_i(t)$'s, are *not* constrained by the processing.

3.1.2 Classical Fundamental Frequency Estimation

Classical approaches to pitch period estimation [72] include the Gold-Rabiner parallel processing method [55], choosing the minimum of the average magnitude difference function (AMDF) [164], choosing the peak of the cepstrum [102], and choosing the peak of the autocorrelation function [151].

In these fundamental approaches, the short-term spectral envelope often interferes with the pitch period estimation procedure. For example, methods that choose the peak of the cepstrum or autocorrelation function often obtain a poor pitch period estimate for short pitch periods due to interference of the spectral envelope information present in the low-time portion of these signals with the pitch period peak. Ross et al. [164] remark in their description of the AMDF pitch estimator that the limiting factor on the accuracy is the inability to completely separate the fine structure from the effects of the spectral envelope. Deller notes in [34] that there are two problems with the autocorrelation function method: 1) that if the first formant is particularly strong, that it can create a "competing" periodicity in the autocorrelation and, (2) that since the speech is only quasi-periodic, that the peaks are less prominent – making peak picking (for the pitch estimate) more difficult.

The Simple Inverse Filter Tracking (SIFT) algorithm attempts to compensate for the spectral envelope prior to pitch estimation by the use of a whitening, or inverse, filter. In the SIFT algorithm, the speech is lowpass filtered (typically to 1 kHz.) and decimated in order to "suppress superfluous high-frequency content". To create the inverse filter, a low-order LP analysis ($M \approx 4$) is used, since we would expect no more than two formants in the nominal 1 kHz bandwidth. An LPC residual is generated from this, and it is expected that it should appear as a

phase-scattered version of the excitation and reasonably periodic. Then the autocorrelation function is used to detect the periodicity. However, this technique boosts the low energy regions of the spectrum which tend to be dominated by noise. The noise reduces the "periodic signal"-to-noise ratio. Consequently, although performance is improved by reducing the effects of the spectral envelope (via whitening), the performance in noise results in a degradation of performance when it is needed most. This noise degradation of the estimate is the reason for the 1 kHz lowpass filtering. As a result, this and most other classical methods of pitch estimation only use signal energy from the low-frequency, highest-energy portions of the signal. The SIFT method also relies upon short-term stationary of speech for the whitening process.

The PMMM, on the other hand, makes no assumption concerning the stationarity of speech. Using the redundancy of speech, it is able to use information from the "best" harmonic estimates to generate an accurate pitch estimate. In the beta clustering example of Section 2.5.2, the harmonics chosen for the fundamental frequency estimation were 7, 8, 9, 15, 16, 24, 25, 27, and 28 of the 29 possible harmonics in the 0 to 3500 Hz range. The average frequency of each PMMM component corresponding to harmonic numbers 24, 25, 27 and 28 are all above 2700 Hz. This frequency range would not even be used by the SIFT algorithm for fundamental frequency estimation.

The PMMM is also able to generate accurate pitch information in the presence of arbitrary noise. This is particularly important for the case where the interference source is narrowband, but possesses high power. Most classical estimation schemes for this case will degrade, as these techniques work on the short-term energy (or sometimes magnitude) of the noisy speech frame. In contrast, the PMMM will, in most cases, disassociate these narrowband components from the speech – their energy will have little or no effect on the accuracy of the resulting pitch estimate.

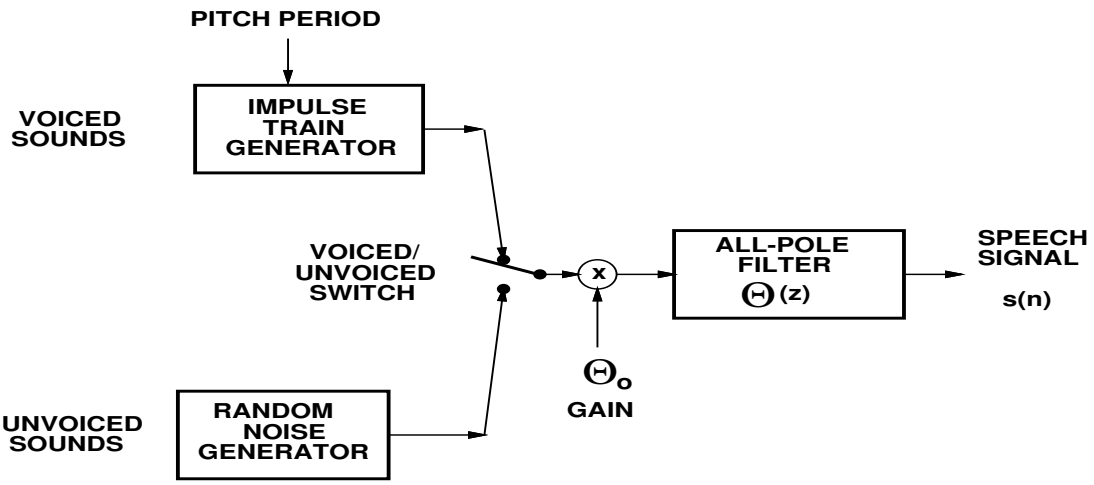

Figure 3.1: LPC Based Speech Model

3.1.3 Classical LPC Based Vocoders

The most widely used and well-known speech model is the LPC vocoder. The LPC-based, all-pole, speech model was introduced in Section 1.5.2.2 and it, in turn, was based on the classical speech model discussed in Section 1.2. The figure for this model is reproduced as Figure 3.1 for convenience. Because robust, short-time spectral envelope estimation is required for many speech enhancement techniques, we have already made mention of the techniques used for estimating the LPC coefficients, $\hat{a}(i,j)$ of Equation 1.5.

As implied by Figure 3.1, the speech frame is classified as being either voiced and unvoiced. If voiced, an estimate of the pitch period is needed. Then the gain Θ_0 necessary for the required input is calculated.

The vocoder models in the following sections extend the basic LPC coder methodology toward the attainment of a higher quality synthesis.

3.1.3.1 Residual Excited LPC Vocoder

Speech quality in LPC-based coders can be improved by computing a residual signal and transmitting this residual to the synthesizer. The residual is simply the difference of the synthesized (LPC model based) speech and the input speech.

This is tantamount to "putting back" what the original LPC vocoder didn't capture. The Residual Excited LPC Coder, therefore, does not attempt to improve upon the underlying all-pole model. In practice, the residual is whitened using an inverse filter based on the LPC coefficients and specialized techniques are used to reduce the bit rate necessary to transmit the residual. The goal is to send a perceptually similar, but bit rate reduced, residual for use in place of the original residual in the synthesis.

3.1.3.2 Multipulse LPC Coders

One of the shortcomings of the Residual Excited LPC Vocoder is the requirement to send the coded residual for the speech synthesis. The residual regeneration scheme for the low bit rate residual results in a crude approximation of the higher frequencies in the original residual. The Multipulse LPC is an analysis-by-synthesis method, due to Atal and Remde [7], which has the basic configuration Figure 3.2. The multipulse excitation consists of a short sequence of pulses whose amplitudes and locations are chosen to minimize the residual between the LPC generated speech and the input speech. Typically, the amplitude and locations of these pulses are optimized one at a time.

The LPC coefficients are generated in the usual manner. Then the present speech frame is subtracted from the LPC generated speech to form a residual error $\zeta_j(n)$. This error is weighted by a perceptual weighting filter $W(z)$, and the result is presented to an error minimization algorithm whose task is to select the best location and amplitude of the next pulse. This procedure is continued until the error is reduced below some acceptable level or the number of pulses reaches its maximum. This is the basic scheme; other pulse optimization strategies can be used (e.g., see [89]).

This coder had been effective for synthesizing good-quality speech at 9600 bps. We note that it has attained this quality by focusing, albeit in an ad-hoc manner, on the specific details of the excitation. We note that such a procedure is

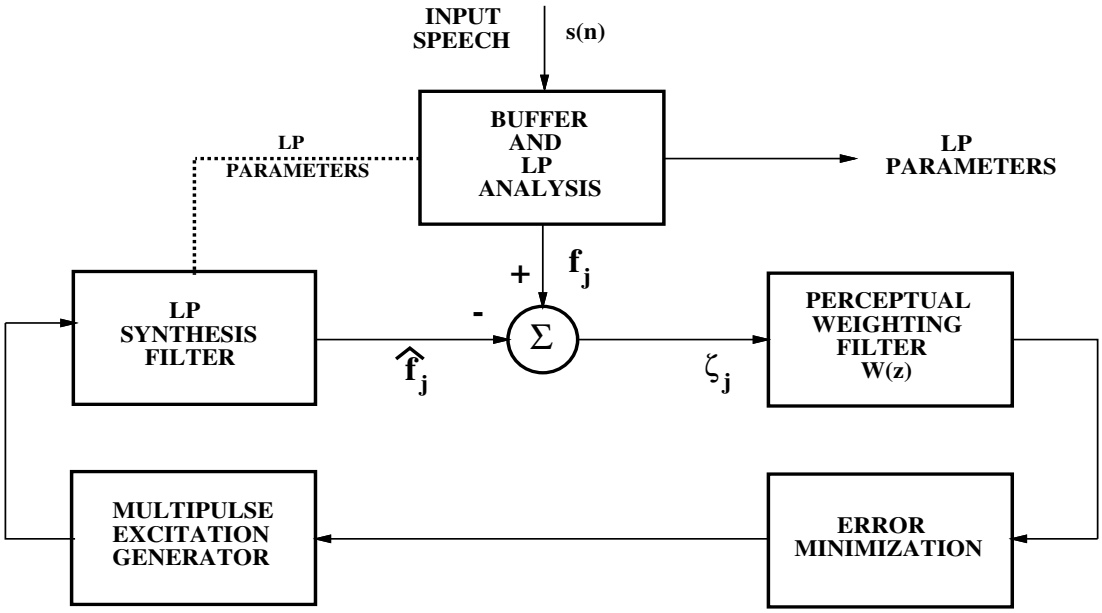

Figure 3.2: Multipulse LPC Excitation Methodology

quite capable of modeling a quasi-periodic source that the PMMM has uncovered. Indeed, Singhal and Atal [182] have noted that, for voiced speech, significant correlation exists in the multipulse excitation from one pitch period to the next.

3.1.3.3 Code-Excited Linear Prediction

Multipulse LPC vocoders produce good quality speech signals, but at a cost of increasing the amount of information associated with the excitation. In an effort to further reduce amount of information associated with the excitation, the Code-Excited Linear Prediction (CELP) Coder was invented [171].

The methodology behind the CELP coder is similar to the Multipulse LPC coder in that it uses, as an error metric, a perceptually weighted residual. The residual is, like the multipulse LPC case, the difference of the speech synthesized with the present parameters and the speech frame input. The difference is that it selects, as its excitation source, a specific stored sequence from a Gaussian excitation codebook. The Gaussian codebook is designed prior to its use in the system. The technique amounts to performing an exhaustive search through the

entire codebook and transmitting the index of the codebook entry that had the minimum (perceptual) error for the speech frame considered.

The primary practical advantages of CELP coders over Multipulse LPC coders are that the search can be made more quickly than the multipulse iterative procedure, and that the required bit rate for a given quality is reduced. Again, we note that the quality over the base LPC coder was obtained by improving the excitation.

3.2 Spectrum Based Harmonic Coding

In this section we compare the PMMM to two other speech models that model much, if not all, of the speech signal as a sum of harmonically related sinusoids.

3.2.1 Sinusoidal/Harmonic Models of McAulay and Quatieri

The sinusoidal model of McAulay and Quatieri [108] models the speech waveform by a sum of sine waves with time-varying amplitudes, frequencies, and phases. Their model for an analysis frame (typically 20 to 30 ms) of speech is:

$$x(n) = \sum_{k=1}^{M} a_k \cos[\omega_k n + \phi_k] \tag{3.1}$$

where ω_k is the average frequency for the frame and the a_k and ϕ_k are the sets of amplitude and phase constants for each harmonic component. In the analysis stage, these parameters are obtained from a STFT that is computed every 5 to 10 ms over a 20 to 30 ms analysis window. The frequencies, ω_k, are estimated by STFT Magnitude (STFTM) peak picking and the phases, ϕ_k, are computed as the phase of the STFM at those frequencies.

In a subsequent paper by Quatieri and McAulay [148], they improve upon the above model toward the end of time-scale and pitch modification of speech. They note that their earlier model did not attempt to maintain the original phase

relationship among the sine waves, that such a model tended to cause an objectionable "reverberant" quality to the reconstruction, and that this property is essential for the purposes of time-scale and pitch modification of speech. Virtually all of [148] is involved in describing how to estimate these phases and ensure that the reconstruction under time-scale and pitch modification maintains these phase relationships.

In yet another paper by Quatieri using this model for the purposes of speech coding [111], another modification is introduced to simulate the noise-like portions of the spectra of mixed-voice and unvoiced sounds. The modification is essentially to randomize the phases of the harmonic components in these regions; that process results in the perception of a noise-like sound (since the phases are updated every 5 to 10 ms.).

The PMMM, as a long-term speech model, maintains a coherent time-locked relationship between the phases of the PMMM components. The PMMM synthesis, to be described in Section 5.2, does not attempt to duplicate the phases of the original speech[1], although this can easily be done[2]. However, the time warping function, $\beta(t)$, maintains the important *relative* phases of the harmonic components over both time-scale and pitch modification. Since the PMMM successfully models mixed voice speech well, there is no need for ad-hoc techniques to make the sinusoidal speech appear noise-like. Lastly, due to the wideband phase modulation property of the large $\beta(t)$ modulation, Section 5.2 demonstrates that it is also possible to generate an arbitrary noise-like spectrum with the PMMM.

3.2.2 Multi-Band Excitation Vocoder of Griffin and Lim

The Multi-Band Excitation (MBE) vocoder of Griffin and Lim [58, 59] is similar to the coder of McAulay and Quatieri in that it is a harmonic coder. It is different in that the amplitudes are chosen by estimation of the short-term spectral envelope

[1] McAulay and Quatieri also note that this is not required for synthesis if time-scale or pitch change is not the application, provided the frame rate is approximately 1000 Hz or greater.

[2] The PMMM phases, $i\bar{\omega}_1\Theta_i$, of Equation 2.8 can be easily calculated from previously determined quantities.

and a binary Voiced/Unvoiced (V/UV) decision is made for each harmonic. The MBE parameters are, therefore, the spectral envelope, the fundamental frequency, a V/UV decision for each harmonic, and the phase of each harmonic declared voiced. It also relies on short-term stationarity of speech for the estimation of its parameters.

The method of estimating the parameters involves a two step procedure. In the first step, the pitch period and spectral envelope are estimated to minimize the error between the original spectrum and the synthetic spectrum. Then, the V/UV decisions are made based on the closeness of fit between the original and synthetic spectrum at each harmonic.

Griffin argues that the individual V/UV decisions are required in order to match the spectrum in mixed voicing regions of typical mixed voice phonemes such as a /z/. When an unvoiced region in the spectrum occurs, the coder simply replaces that portion of the spectrum with a (correctly shaped) noise. In previous works, such as the newest Quatieri coder [111], the V/UV decisions controlled large regions of the spectra. Griffin argues that these approaches are too restrictive to model many speech segments, especially voiced speech in noise.

However, the PMMM is not only able to effectively deal with mixed-voicing, but is able to model its cause. The large time warp *caused* the wideband modulation of the pitch harmonics, which, in turn, *caused* the noise-like appearance in the spectra.

3.3 Non-Stationary Speech Models

In the introduction of this thesis, a case was made for speech modeling that used longer frame sizes and adequately captured the non-stationarity in speech. It appears that long term modeling or models that address the (pitch) non-stationarity of speech were investigated only by those applications that absolutely required it.

Below are the models which describe the present art in time warping research, of which only the model of Stettiner et al., can truly be considered a "long-term"

model.

3.3.1 Non-Stationary Model of Almeida and Tribolet

Almeida and Tribolet describe a non-stationary spectral model for voiced speech in [2] in which they consider voiced speech to be the output of a linear time-varying filter $h(t,\tau)$ that has been excited by a quasi-periodic pulse train $p(t)$. The objective of their modeling is to define an "occurrence" function $\phi(t)$, that would time warp the quasi-periodic pulse train into a periodic pulse train[3]. They note that this function is "quite arbitrary" in that its values are only defined at the pulse instants. They formulate what they call "generalized harmonics" and then a "generalized spectral line" using this model. Their goal is to find a mapping such that the spectrum of their generalized spectral lines formulation with the time warping $\phi(t)$ matches the observed speech spectrum.

In their formulation, the time warping function, $\phi(t)$, is approximated by a power series,

$$\phi(t+\tau) = \phi(t) + \Omega(t)\tau + \sum_{l=2}^{\infty} \frac{1}{l!} \frac{d^{l-1}\Omega}{dt^{l-1}} \tau^l \tag{3.2}$$

where $\Omega(t)$ is seen as the fundamental frequency of the speech in a frame of speech about time t. They immediately note that this sum must be truncated in practice, and that a modeling error will be introduced. They further note that their method is *extremely* sensitive to the fundamental frequency estimate that the power series is "expanded" around and formulate many methods to minimize this error. Owing to the complicated nature of their formulation, they also constrained the *amplitudes* to be constants. Due to the short-term stationarity of speech, this limited the length of the speech frames considered to between 32 and 64 ms. Their best results were obtained with a 32 ms window.

[3]This is an unfortunate choice, as the PMMM reserves $\phi(t)$ for individual phase functions of the PMMM components.

This model is then used to design a harmonic coder in this paper and in a subsequent paper by Marques and Almeida [105]. In the subsequent paper the sinusoidal basis components for a frame, $o_k(t)$ are modeled as

$$o_k(t) = e^{-\mu_k t^2} e^{\lambda_k t} e^{j(\Delta_k t^2 + \omega_{0k} t)} \tag{3.3}$$

where it is seen that only a linear ramp (of $2\Delta_k$) has been parametrically added to the harmonic model. The amplitudes are parameterized as Gaussian envelope parameters μ_k and λ_k. Like the previous work, they do not attempt to actually de-warp the speech signal. Instead, they iteratively derive the model parameters from an initial estimate of the fundamental frequency. They note, and emphasize, that if the initial estimate is bad, then the iterative estimation process "will give meaningless results" due to the nonlinear nature of the estimation process.

Note that this model parameterizes both the frequency *and* the amplitude parameters and possesses the parametric sensitivities mentioned.

3.3.2 Long-Term Model of Stettiner, et al.

In an independently derived paper of Stettiner, Malah, and Chazan at ICASSP '93 [188], an iterative time warping procedure similar in spirit to the PMMM de-warping was reported. Their aim was to develop a long-term model for speech for signal processing applications that require a high frequency resolution for signal separation.

The model used in this work is:

$$s(t) = g(t) \sum_{k=-\infty}^{\infty} d_k(t)\, c_k\, e^{jk\phi(t)}. \tag{3.4}$$

They note that after both the time warp *and* the amplitude variations have been inverse filtered out, that the model will reduce to:

$$s_2(t) \triangleq \sum_{k=-\infty}^{\infty} C_k \, e^{jk\phi(t)} \tag{3.5}$$

Although their model does not constrain their time warping function, $\phi(t)$[4], to have a parametric form, it imposes that the amplitude estimates be constants, C_k, after amplitude normalization.

Their procedure uses the EM algorithm to iteratively estimate improved time-warping and amplitude estimates. Specifically, the input is pre-whitened by an inverse filter calculated from an All Pole speech model and gain normalized. Then it is passed through another whitening filter to diagonalize the covariance matrix of the Weighted Least Squares formulation used to obtain the model parameters. Then they solve for the time warping function $\phi(t)$ by an iterative application of a simple comb filter and time warping. Their technique is the *only* other one known to this author that does not constrain the time warping function, $\phi(t)$, to have a particular parametric form.

Note, however, that the model is "limited" by the *short term* stationarity assumption of the all-pole model that the inverse filter was designed from. They note that the effect of this (amplitude non-stationarity caused) error is most pronounced for the speech samples near the ends of their analysis frames. It is this authors speculation that this also limits the length of the speech frame considered.

They report a spectrum of a 100 ms time-warped segment of speech in which they show a much reduced smearing of the harmonics in the frequency domain. Although improved, the compressing of the pitch harmonic energy is in no manner as dramatic as that demonstrated by the PMMM model in Figures 2.16 and 2.17 of Section 2.6.

Although one of their stated purposes is the speaker separation task, the algorithm was only demonstrated for the case of a single speaker, presumably with

[4]This is an unfortunate choice, as the PMMM reserves $\phi(t)$ for individual phase functions of the PMMM components.

little appreciable noise. No discussion of the effects of additive noise on the algorithm was reported.

Note that the C_k of Equation 3.5 are constrained to be constants after the spectral envelope normalization. Although this is possible for the no-noise case, it is unlikely to be the case for a competing speaker or for arbitrary noise. The inverse filter has no way of determining which portions of the spectrum are dominated by the speech and which sections are dominated by the noise. Therefore, it is an educated speculation by this author that this algorithm will degrade with additive noise, since the noise will be amplified and considered of equal importance to the speech energy in the EM optimization algorithm. This inverse filtering problem with noise was also seen as a pitfall of some of the fundamental frequency estimation techniques in Section 3.1.2.

The PMMM not only does *not* constrain the amplitudes of the PMMM components, it is also able to *associate* harmonic peaks of the Fourier spectrum with the wanted speech. As a result, accurate de-warping and pitch estimation can be obtained by the PMMM in high additive noise conditions.

3.3.3 DTW-Based Comb Filtering of Graf and Hubing

The method of Graf and Hubing [56] is a pitch synchronous averaging technique where the warping of the pitch period immediately prior to and the pitch period immediately following the current pitch period are non-linearly warped into the current pitch using a Dynamic Time Warping (DTW) technique. The DTW was performed on a sample-by-sample basis with 10 kHz sampling. The resultant signal is comb filtered about the pitch frequency of the current frame.

Results for this processing are compared to a standard Adaptive Comb Filter and were shown to yield improved SNR results for SNRs greater than 5 dB. For SNRs more negative than 5 db, the DTW technique was inferior to the ACF, as the proper alignment of adjacent pitch periods with high noise peaks is increasingly difficult.

The comb filtering results for *noisy* speech were generated using pitch estimates from the *clean* speech. This is consistent with the reported performance of many speech enhancement and speaker separation techniques, although this author believes this practice is unfair when compared to the PMMM processing which does not require *a priori* pitch estimates for its processing.

This work is reported here because it is a first-order time-warping procedure. In general, using PMMM parlance, the difference between time base t and the uniform time base $\tau = [t + \beta(t)]$, has been shown to be very small (typically .1 to .5 %) in Section 2.3.1. Stettiner, in a private communication with this author, has verified that a sample by sample warping is much too course a warping for this purpose and that such a warping introduces its own (highly nonlinear) significant errors.

3.4 Summary of PMMM Processing and Modeling Advantages

In summary, the PMMM is a new long-term speech model that is significantly different than the prior art in several ways. The significant differences are enumerated below:

1. In contrast to classical (short-term) speech models, there is no need to characterize the vocal tract and excitation parameters separately. No spectral envelope estimation is necessary, as the PMMM component amplitudes are "unconstrained" by PMMM processing. No independent fundamental frequency estimates need be made. The PMMM processing can obtain robust F_0 estimates, even with arbitrary noise (this will be shown later in Section 5.1).

2. In contrast to improved quality LPC coders, the PMMM captures the naturalness of unvoiced and mixed-voice speech. No improvements to the "excitation" of the model is necessary, since the quasi-periodicity of the speech is contained in the time warping functions, $\beta(t)$.

3. In contrast to other harmonic or sinusoidal models, there are three significant things to note. First, the need for harmonic amplitude estimation via either all-pole modeling (Griffin) or STFTM harmonic amplitude estimation (Quatieri) is not necessary. In fact, due to the harmonic sharpening of the time warping functions, the PMMM is capable of more precise amplitude estimation in noise. Secondly, there is no need to explicitly create a noise-like spectral region. If it were present in the original speech, the $\beta(t)$ function captured what was necessary to regenerate it. Thirdly, there is no explicit need to calculate the relative phases of the harmonics, as the coherent $\beta(t)$ modulation maintains the relative relationships under time-scale and pitch change.

4. In contrast to the non-stationary model of Almeida, the PMMM does not require the rigid parametric form for the warping or the harmonic amplitudes. The rigidity and parameter sensitivity of these parameterizations limit successful application.

5. In contrast to the long-term model of Stettiner, no inverse filtering is required to satisfy the amplitude stationarity assumptions that a fixed C_k requires. The inverse filtering is probably detrimental to its performance in noise, and it places a practical upper bound on the length of the speech frame considered.

In addition, *none* of the speech models have observed the coherent (quasi-periodicity produced) phase modulation analogy to communications theory that was outlined in Section 2.4.2.

Of all of the prior art for long-term speech models, the independently derived model of Stettiner et al., is the closest to the spirit of PMMM-based time unwarping, as it does not constrain its time warping function to a parametric form.

Chapter 4

PMMM Ties to Recent Speech Perception Research

4.1 Auditory Fusion Principles

The task of separating sounds that have arisen from different sources has been studied by auditory theorists for quite a long time. It is acknowledged by these theorists that this process is difficult and that segregation of sounds into streams of "belongingness" is not based purely on cochlear information. Integration of sounds is also a function of experience and higher level processing of the acoustic information by the brain and the central auditory paths between the cochlea and the brain.

Focusing on the fundamental acoustic information portion of this problem has led to many studies as to what types of acoustic stimuli bind or fuse individual sounds so as to create the illusion that the combined sounds had originated from one source. Many studies have found experimentally that if you modulate tone complexes (harmonic or non-harmonic) in amplitude and/or frequency in a coherent way, the sounds tend to fuse. This property is called "Micromodulation" or "Auditory Fusion" and is illustrated in Figure 4.1.

In this example, all the tones are coherently modulated in frequency with a constant difference on a *log frequency* scale, except the tone at frequency "f4". If the tones were played at reasonable loudness, this complex would be perceived as two sounds: a pure tone at frequency "f4" and another "sound" with a certain character (or timbre). This fusion phenomenon has been demonstrated with both amplitude and/or frequency modulation of tones. Brian Moore in his book *Introduction to the Psychology of Hearing* [118], states:

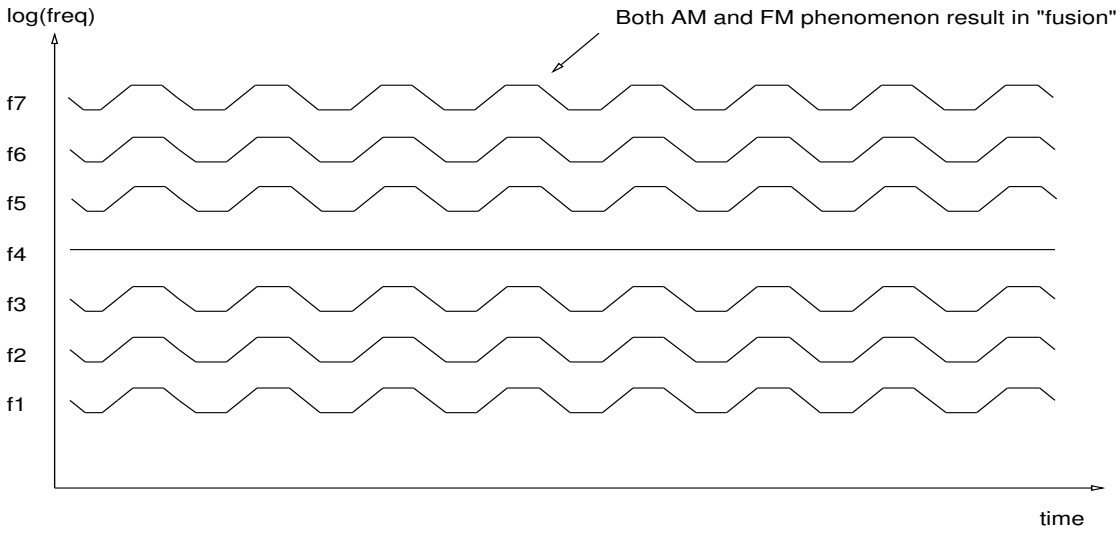

Figure 4.1: Auditory Fusion Principle

> For example, if we present a steady complex sound containing many components with randomly distributed frequencies, a single noise-like sound will be heard, with a certain timbre. A subgroup of components in the complex can now be made to stand out perceptually from the rest by making them vary in a coherent way in either frequency, amplitude, or both. This group will be perceived as a prominent "figure" against a steady background, and both the figure and the background will have a timbre different from that of the original unvarying sound.

Many of these findings are relatively new [18]. In particular, the term "micromodulation", which refers to modulation in the range of tens to hundreds of hertz, is a relatively new term [106].

The PMMM exploits the property that the PMMM components that are coherently phase modulated belong to the same sound source. The PMMM components have average frequencies that are harmonically related. The average frequencies of these components, when plotted on a *log frequency* scale, would be appropriately spaced. However, the *deviation* of the instantaneous frequency of each PMMM component from their respective center frequencies, when plotted

on a *log frequency scale*, would have a *constant difference*[1] – in exact agreement with the condition auditory theorists have found that fuse tones!

The PMMM is also in direct agreement with these findings with respect to both the frequency modulation *and* the amplitude modulation. The case for the perception of coherent amplitude modulation based on the PMMM is made shortly.

4.2 Cochlear Modeling and Time Synchrony Analysis – Implications for Harmonic Association

Cochlea modeling and central auditory processing research have motivated researchers toward the efficient representation of acoustic signals and provided new insight toward novel representations for robust speaker/sound recognition [208]. Although segregation of sounds into streams of belongingness is not based solely on cochlear information, recent research suggests that the ear performs a preliminary type of component extraction based on these auditory principles.

The most fundamental processing the cochlea performs is the segregation of incoming frequencies in a tonotopically ordered manner along its length. Functionally, the cochlea may be viewed as a parallel bank of overlapping bandpass filters. Above about 800 Hz, the impulse response of these filters are related to each other by a dilation – they maintain a constant Q-factor. Below 800 Hz, the frequency scale is not purely logarithmic, but rather becomes progressively more linear, especially below 500 Hz [71]. Therefore, most cochlea modeling assumes that the auditory system performs an approximate affine wavelet transform above 800 Hz [94, 79, 174, 211]. Recall from Section 2.4.2 that the bandwidth for a sinusoidal $\beta(t)$ (i.e., $\beta(t) = \alpha \sin(\omega_{\text{mod}} t)$) was given by:

[1] Note that $\phi_i(t) = i\,\bar{\omega}_1[\,t + \beta(t)\,]$ and $\omega_i(t) = i\,\bar{\omega}_1\,(1 + \frac{d\beta(t)}{dt})$. Thus, $\log[\omega_i(t)] = \log[i] + \{\log[\bar{\omega}_1] + \log[1 + \frac{d\beta(t)}{dt}\,]\}$. Since the term in the brackets is common to all pitch harmonics, the modulation is a constant difference, when plotted on a log scale.

$$W \approx 2\omega_{\mathrm{mod}}(1 + i\bar{\omega}_1\alpha) = 2\omega_{\mathrm{mod}}(1 + \bar{\omega}_i\alpha), \tag{4.1}$$

which is also constant Q for $\bar{\omega}_i\alpha \gg 1$. For a typical α from Section 2.3.1 of 0.002 and setting $\bar{\omega}_i\alpha$ equal to something solidly in the constant Q region, say $\bar{\omega}_i\alpha = 10$, yields:

$$f_i = \frac{10}{2\pi 0.002} = 795\mathrm{Hz}. \tag{4.2}$$

This is in agreement with the auditory findings above. Equation 4.2 says that the *bandwidth* of frequency components above about 800 Hz are predicted by the PMMM to be constant Q[2].

Shihab Shamma in his paper [177] discusses the role of Lateral Inhibition Networks (LINs) in central auditory processing. A specific LIN output effectively emphasizes correlations in adjacent bandpass filter outputs of the cochlea. There are many LIN's (thousands) along the length of the cochlea. In this paper, Shamma states:

> Summing these LIN outputs results in a waveform that is amplitude modulated at the difference of the harmonics (i.e., the pitch) ...the *envelope* of this waveform exhibits properties and dependencies on the complex sound ...

Shamma's observation above suggests that a similar auditory fusion may occur prior to processing by the brain.

Previous speaker separation research at Rutgers University [154] found that the *amplitude* envelopes of the Formant Mode Modulation Model (FMMM) decomposition for a specific speaker, $a_{i_{FMMM}}(t)$, had a coherent amplitude variation at their pitch frequency. That is, the FMMM components of each speaker could

[2]We note that the bandwidth of the higher frequency components is dominated by the quasi-periodicity induced phase modulation.

be associated with the correct speaker by noting correlations in the FMMM amplitude envelope variations[3]. The reason for this is clear for the higher frequency cochlea bandpass filters. Whenever 2 or more pitch harmonics fall within one of these passbands, its *amplitude envelope* maximums are in approximate time-synchrony with the pitch period. A simple example of a signal $s_{BPF}(t)$ representing three perfectly periodic pitch harmonics (i.e., $\beta(t) = 0$) with amplitudes a_{n-1}, a_n, a_{n+1} and fundamental frequency, $\bar{\omega}_P$, written in analytic signal form is:

$$\begin{aligned} S_{\mathrm{BPF}}(t) &= a_{n-1}e^{j(n-1)\bar{\omega}_P t} + a_n e^{jn\bar{\omega}_P t} + a_{n+1}e^{j(n+1)\bar{\omega}_P t} \\ &= a_n e^{jn\bar{\omega}_P t}\ [1 + \frac{a_{n-1}}{a_n}e^{-j\bar{\omega}_P t} + \frac{a_{n+1}}{a_n}e^{j\bar{\omega}_P t}] \\ &= [1 + \frac{2a_{n-1}}{a_n}\cos(\bar{\omega}_P t) + \frac{(a_{n+1} - a_{n-1})}{a_n}e^{j\bar{\omega}_P t}]\ \ a_n\ e^{jn\bar{\omega}_P t} \\ &= A_{AM}(t)\quad e^{jB_{FM}(t)} \end{aligned} \tag{4.3}$$

where it has been assumed that $|a_n| \geq |a_{n-1}| \geq |a_{n+1}|$ (which is the usual case due to the roll-offs for the bandpass filters) and the phase angles have been dropped for simplicity. Viewed as an AM/FM signal, the $a_n e^{jn\bar{\omega}_P t}$ is the carrier term and the term in the brackets is the (complex) amplitude modulation of the carrier. In cochlea modeling, the neural rate information is sensitive to the amplitude of the cochlea filter outputs near the amplitude envelope maxima. Note that for large $A_{AM}(t)$ modulus that $A_{AM}(t) \approx [1 + \frac{2a_{n-1}}{a_n}\cos(\bar{\omega}_P t)]$. That is, the modulus of this cochlear filter component has an amplitude variation tied to the pitch. This is also true if more than three harmonics are included in $S_{BPF}(t)$ (since the $|a_{n-k}| < |a_{n-l}| \vee l = -1, 0, -1$ and $k \neq -1, 0, -1$). A similar, though more complicated, analysis (including the phase angles of the harmonics) can show that the pitch nonstationarity (i.e. the case of $\beta(t) = 0$) is reflected in the amplitude envelope also. Thus the *amplitude envelope maxima* of these cochlea bandpass

[3] As an aside, the FMMM based results launched the PMMM effort whose purpose was to make the association at the harmonic level.

filter outputs track the non-periodicity of the input. This has led many scientists to investigate [118, 120] and some engineers to design speech models based on "time-synchrony" analysis. Such an engineering model is the Ensemble Interval Histogram (EIH) model of Ghitza [53, 54, 152].

A new fundamental frequency estimation algorithm, the AMPEX (auditory model-based pitch extractor) algorithm, uses this "time synchrony" model to extract pitch. In the paper describing this algorithm [198] it is compared to the SIFT algorithm of Markel (see Section 3.1.2) and the Sub-Harmonic Summation (SHS) algorithm of Hermes [70]. AMPEX was found to be superior to both SIFT and SHS and is considered by many to be the best pitch extraction algorithm [22].

4.3 PMMM Spread Spectrum Analogy Revisited

Yang, Wang, and Shamma [208] also suggest that the auditory representation used by the auditory system utilizes a non-uniform sampling of the cochlea filter bank outputs at the waveform extrema (i.e., the amplitude envelope maxima) and that their representation is "directly responsible for the enhancement effects which endow it with superior robustness in noise."

The PMMM warping described previously warps $s(k)$ in a manner that sharpens the spectrum of $s(k)$ for subsequent processing. Because PMMM processing is performed in discrete time, the interpolation process used in the de-warping can alternatively be viewed as a *non-uniform* sampling of $s(k)$ in a manner that sharpens the spectrum of $s(t)$. In the PMMM paradigm, the *non-uniform* sampling of $s(k)$ allows for the demodulation of a spread-spectrum like signal. Thus, the $\beta(t)$ processing previously described may be construed to have an auditory basis in its function, although it is clear that the auditory system does not perform the "warping" as the PMMM does.

As previously noted in the section on Spread Spectrum (Section 2.8), this time warping redundancy embodies the signal with a high immunity to interference –

for a demodulator able to demodulate the signal. This, again, is in agreement with Yang et al. statement above. No other work known to this author has pointed out the direct analogy of the quasi-periodicity of the speech signal to Spread Spectrum Communications Theory.

Additionally, new neurophysiology research in rate and temporal hair cell firings [33, 10] also support Yang et al., and therefore, the PMMM findings.

4.4 Comments on PMMM Ties to Auditory Research

Auditory theorists have recently found that coherent amplitude and constant difference log frequency modulation of signal components results in auditory system "fusion" of these components. The coherent time warping that the PMMM $\beta(t)$ uncovers introduces exactly this type of frequency modulation.

This coherent time modulation creates a spread-spectrum like signal that possesses a high immunity to interference – for a demodulator capable of demodulating it. The cochlea and Central Nervous System (CNS) processing is apparently capable of "demodulating" the speech signal. Most recent research suggests that the cochlea/CNS processing performs this task via a type of time-synchrony analysis based on the cochlea bandpass filter outputs. An analytic signal AM/FM analysis of a cochlea filter output suggests that such a timing extraction is possible and that association between signal components of the same sound may occur this way in humans. If true, the CNS processing occurs on a cochlea filter "component" level - not on the harmonic level.

The PMMM signal analysis, however, makes the time-warping association of PMMM harmonic components on the harmonic level. Therefore, the PMMM processing can associate components, sharpen the wanted harmonic components, and *remove* unwanted signal (noise) energy that is in between harmonics. This is a capability that none of the spectrum-based speech enhancement techniques of Section 1.5 possess. The previous comb filtering techniques possess this ability for only very stationary utterances.

Although this is true, the removal of low level noise energy in high signal energy regions, such as formants, may not be perceived due to masking [9, 48][4]. However, the noise removal from these regions may be of value for other applications (e.g., speech coding in noisy environments).

[4] Another way to view masking in the context of time-synchrony analysis above is to note that noise in such regions does not significantly modify the AM envelope (and thus the AM-derived time-synchrony) of these component regions.

Chapter 5

PMMM Applications and Numerical Results

The Pitch Mode Modulation model is a new long-term speech model and, as such, can be exploited in much the same way as other speech models have been exploited. As described in the last section, the PMMM can be construed to have an auditory basis. Or, stated differently, the most of the redundancy in speech is "captured" in the time warping function, $\beta(t)$.

The primary application that the model was developed for was speech enhancement for the purposes of speaker identification. The next section addresses this application. While developing the PMMM parameter estimation technique, it was found that speech synthesis from PMMM parameters for voiced and mixed-voice speech was trivial. Additionally, it was found that pitch and time-scale modification, an application that has previously taken much effort to accomplish [148], is also trivially accomplished. These applications are also discussed in detail in this chapter. Other applications of the PMMM are discussed in the next chapter in lesser detail.

5.1 Speaker Identification in Adverse Environments

The performance of most speech recognition and Speaker Identification (SI) systems, whose designs are predicated on assumptions about the ambient conditions, such as low background noise, degrades rapidly in the presence of noise or other signal distortions. It is important to effectively deal with the adverse noise conditions in real-world applications.

Previous research has shown that the performance of such systems is maximized when the system is tested under the noise conditions which approximate those that the system was trained with [152]. However, it is seldom realistic to expect that the noise condition of the training and testing condition be similar due to the varied nature of potential noise sources. It will be shown that PMMM-based enhancement will lessen the speaker identification sensitivity to mismatches between testing and training data.

5.1.1 PMMM Speech Enhancement Technique

The text independent speaker identification application described here uses the PMMM enhancement as a preprocessing step. The speech is first separated into 32 ms frames[1] and involves the following processing steps for a particular frame j:

1. Perform the iterative PMMM de-warping procedure of Section 2.5.4 until all quasi-harmonics are of uniform bandwidth.

2. Comb filter the de-warped signal with a comb filter whose passbands are centered on harmonics of the final pitch estimate, $\bar{\omega}_{1_j}$. The comb filter bandwidth is approximately 31 Hz.

3. Re-warp the resulting signal to re-introduce the quasi-periodicity of the original speech.

Diagrammatically, the result of this processing is denoted as $\hat{s}_F(n)$ of Figure 2.14.

After the PMMM enhancement, the speech is classified by a cepstrum-based speaker identification system. This system includes pre-emphasis by $H(z) = 1-0.95z^{-1}$, 12-order LPC-based cepstrum generation, and Learning Vector Quantizer (LVQ) clustering. A 30 ms Hamming window is applied to the signal every

[1] Recall from Section 2.5.1 that the FFTs span 2 x 32, or 64 ms, of Hamming windowed speech.

ten ms. Additional information on this SI system can be found in [5]. The specific SI system is a secondary concern as we are interested in the performance difference with and without the PMMM enhancement with the *same* SI system.

The speech corpus for this experiment is a subset of the DARPA TIMIT database that included 20 speakers, ten male and ten female, from the New England dialect. The system was trained and tested with approximately 12 to 15 seconds of speech for each speaker. 8 kHz sampling was used throughout. Three types of noise were considered: a white Gaussian noise, a babble noise and a low-pass type colored noise. The babble noise was created by summing the speech of ten male and ten female speakers of the TIMIT database. Each speaker segment in this sum had silence removed and a signal power normalization performed prior to the summation. The colored noise was shaped to match the spectral tilt of an average male speaker. It was created by filtering white noise by a cascade of two low-pass filters designed by the Hamming window technique with orders of three and five and cutoff frequencies of 50 and 1500 Hz, respectively.

5.1.2 Typical Speaker Identification System Training Results

This section presents the baseline performance of the SI system *without* PMMM enhancement.

Figure 5.1 presents the probability of correct identification, P_{correct}, against Signal to Noise Ratio (SNR), in decibels (dB), for the 20 speaker identification task for the three types of broadband noise considered. The calculation of the SNR was based on the long-term average of the signal energy with silence periods removed. When trained on clean speech, the performance in white noise degrades much more quickly than for the colored noise or the babble noise as the noise level of the respective noise sources are increased. 25 dB SNR of white noise, while subjectively annoying, does not impair human listeners in identifying speakers.

In order to improve the robustness to SNR with white noise, the same system was trained with 25 dB of white noise added to the training set. These results are

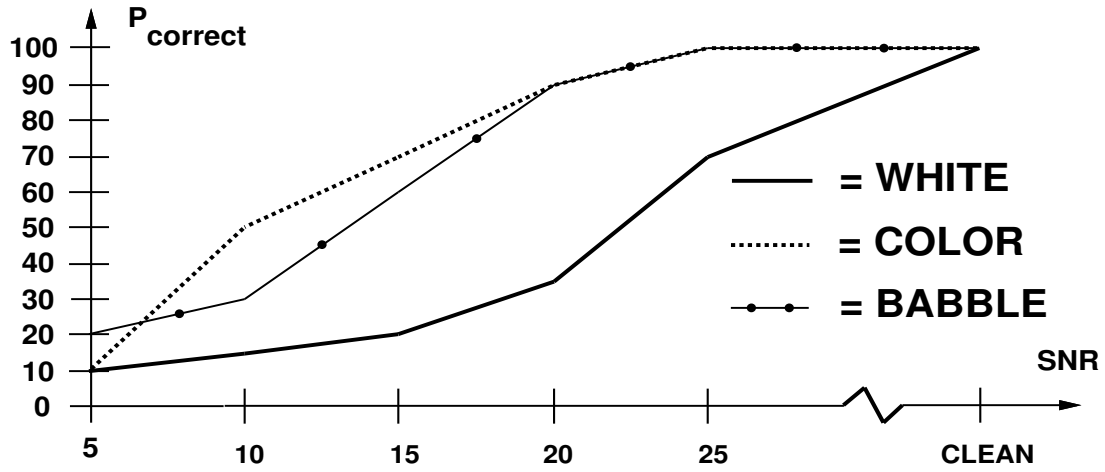

Figure 5.1: Training on Clean Speech (No enhancement)

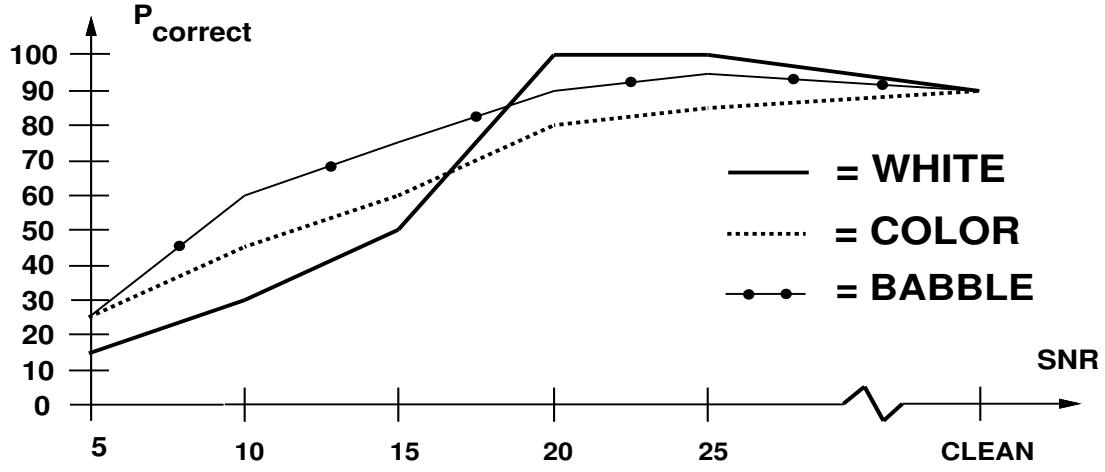

Figure 5.2: Training on 25 dB White Noise (No enhancement)

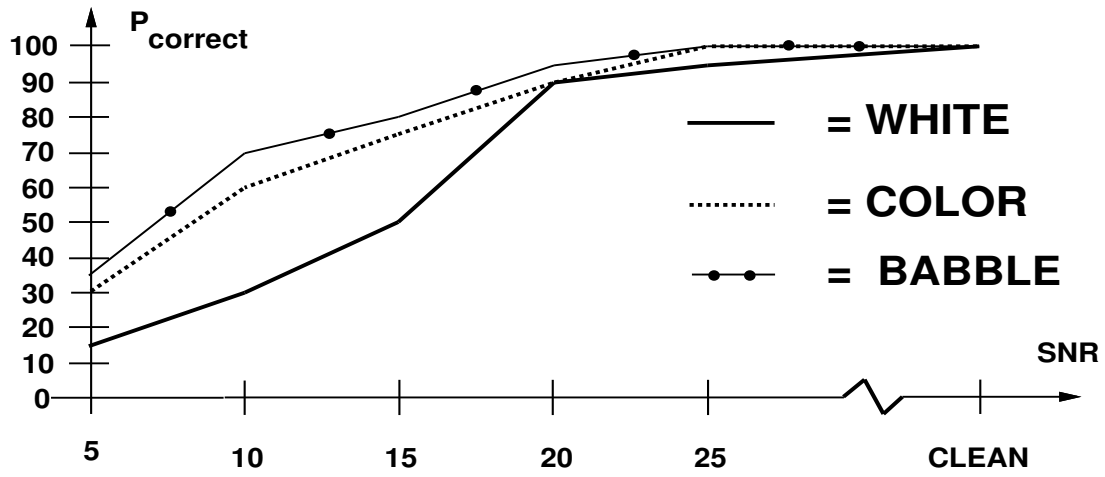

Figure 5.3: Training on Enhanced Clean Speech

shown in Figure 5.2. Note the white noise performance at and near 25 dB is 100% and the babble noise performance at low SNR also improves. Unfortunately, the previously good performance in colored noise and high SNR babble noise degrades. Also note that 100% correct identification in the clean condition (no noise) is no longer attained. When one optimizes the SI system for one specific noise type and noise level, the system "learns" that specific noise type and the performance with a different noise type or significantly different noise level will result in decreased performance. This performance characteristic has been noted in the literature previously [152].

The inadequacy of this type of training is the fact that training on specific noise types and noise level is seldom realistic in real-world application. This is due to the varied nature of possible noise sources.

5.1.3 PMMM-Based Enhancement Speaker Identification Results

Figure 5.3 presents the performance of the system trained with enhanced versions of the clean speech and tested with the enhanced versions of the testing set with their respective noise sources. In comparison to Figure 5.2, the performance of all the noise types improves, with the performance in white noise and babble noise

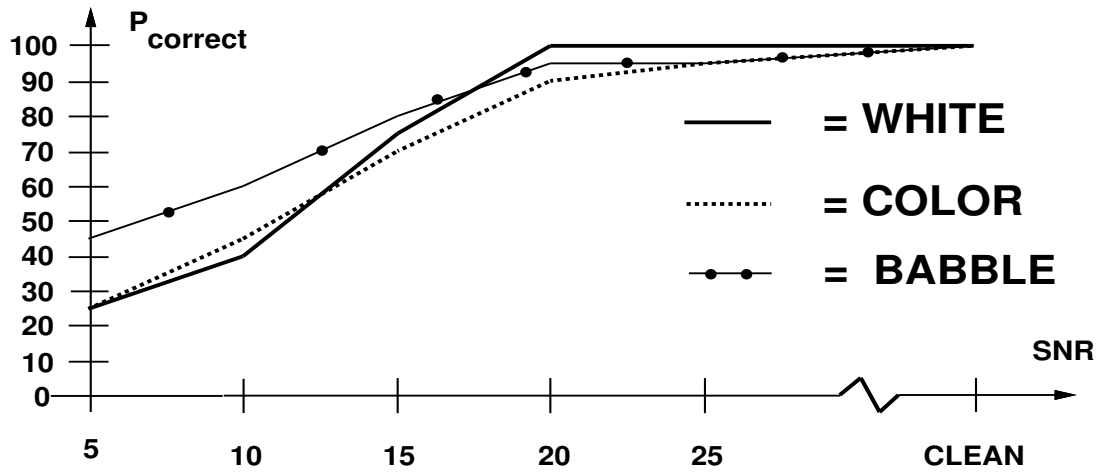

Figure 5.4: Training on 25 dB White Noise Enhanced Speech

improving dramatically. In addition, there is a 100% correct identification for clean speech. Clearly seen is a relative insensitivity to the actual noise type.

It is also possible to train on enhanced versions of the specific noise cases in a manner similar to the 25 dB white noise training of Figure 5.1. In this case, the performance under the particular noise source improves to 100%, but the other noise sources *do not* degrade as drastically, due to the PMMM enhancement. For example, Figure 5.4 presents the performance after training on the enhanced version of the training set with 25 dB of white noise. In comparison to Figure 5.2, there is again less sensitivity to the actual noise type and 100% correct identification for clean speech.

The best performance, overall, is when the training data is the clean enhanced speech. This is fortunate, since it is possible in many applications to obtain training data under controlled conditions. Examples of these applications are: voice verification for clearance into secure areas and home incarceration. However, when this is not possible, training on enhanced versions of noisy speech is still preferable to no enhancement, as a comparison of Figure 5.1 and Figure 5.4 shows. For example, the μ-law coding used in the North American telephone systems only guarantees about 28 dB SNR and the quantization error of these systems is known

to be relatively white prior to far-end equalization[113]. Figure 5.1 shows that training on such a system would not be adequate for the SI application without the aid of enhancement.

5.2 PMMM-Based Speech Synthesis

The PMMM parameterization of speech allows for the enhancement of speech via comb filtering, as in the previous section. As an alternative to comb filtering, the speech could be synthesized from PMMM parameters and used for enhancement. For speech synthesis, $a_i(t)$ estimates must be formed. Figure 2.14 shows the speech synthesis option as the signal $\hat{s}_S(k)$ and Section 2.5.5 describes, in detail, how the $a_i(k)$ estimates should be made. If the synthesis is for enhancement purposes, the $a_i(k)$ estimates should be formed with a relatively narrow bandwidth. If the synthesis is for speech coding in relatively low noise environments, a wider bandwidth could be used.

Specifically, speech synthesis from PMMM parameters on a frame by frame basis can be performed according to:

$$\hat{s}_{S_j}(k) = \sum_{i=1}^{N} a_{i,j}(k) \cos(i\ \bar{\omega}_{1_j}[k + \beta_j(k)]\ + K_{i,j}). \tag{5.1}$$

where j is the frame index and the $a_{i,j}(k)$ are estimated for each component i after completion of the final iteration of frame j. The constants $K_{i,j}$ are adjusted such that the cosine argument is smooth across frames - no attempt was made to recreate the relative phase terms across components (the $i\bar{\omega}_1\Theta_i$ terms in Equation 2). The resulting synthesized speech, $\hat{s}_S(k)$, sounds virtually identical to the comb filtered speech $\hat{s}_F(k)$ for the same noise impairment.

The $a_{i,j}(k)$ were observed to be slowly time-varying waveforms in the PMMM decomposition. To reduce the number of parameters needed for Equation 5.1, the $a_{i,j}(k)$ are parameterized by constants $A_{i,j}$ in a 32 ms frame synthesis. This effectively decimates the $a_{i,j}(k)$ to the frame rate of 31.25 Hz. The constants

$A_{i,j}$, are taken as the average $a_{i,j}(k)$ in a given frame. A new set of $a_{i,j}(k)$ are created for use in Equation 5.1 by linearly interpolating the constants $A_{i,j}$ across frames. The resulting synthesized speech was perceptually identical to the synthesis without the $a_{i,j}(k)$ decimation.

These observations suggest two findings for the PMMM decomposition of speech: that $a_i(t)$ can be represented using low frequency waveforms and that the initial phase offsets for the components are subjectively unimportant[2].

5.2.1 Time-Scale and Pitch Modification of Speech

From a speech modeling perspective, the spectral envelope information is coded by the $A_{i,j}$ of the previous section, and the excitation is coded by the $\bar{\omega}_{1_j}$ and the $\beta(k)_j$.

For time-scale modification of speech, the spectral envelope parameters $A_{i,j}$ remain unmodified except that their presentation in time is increased or decreased. The excitation parameters are modified in the following manner. The average frequency in a frame $\bar{\omega}_{i,j}$ is unchanged, but the $[t+\beta_j(k)]$ is stretched or compressed according to the new time scale. This is shown diagrammatically in Figure 5.5 for a time scale increase of a factor of two.

Previous attempts at time scale modification of speech using a harmonic synthesis [148] found that maintaining the original phase relationships among the sine waves was essential. Calculating, for the purposes of maintaining these relationships, was computationally intense and required time-warping of, so-called, excitation "phase onset times".

The simplicity of PMMM time-scale modification is made clear by considering an instant of time $t_a + \delta t$ in Figure 5.5. This instant of time becomes mapped to $\tau_a + \delta\tau$ in the uniform time scale τ and $t'_a + \delta t'$ in the t' time scale. This warping maintains the *same* relative phase relationships between the sinusoids at time $t_a + \delta t$ in t as in time $t'_a + \delta t'$ in the new time base t' *automatically*!

[2]The unimportance of phase has been noted for speech enhancement applications previously [200].

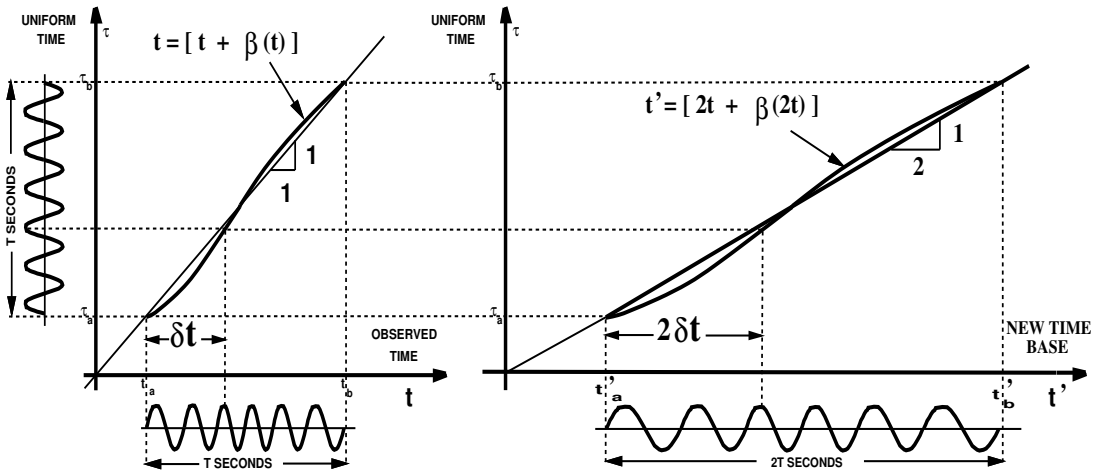

Figure 5.5: Time Scale Modification of Speech

For pitch change, the fundamental frequency must be changed and the amplitude estimates must change. The amplitude estimates must change such that the new values lie on the same spectral (vocal tract) envelope as the original estimates. This can be done by estimating the spectral envelope from the $A_{i,j}$ and then computing new $A_{i,j}$'s at integer multiples of the new fundamental frequency [139]. For the excitation part, there are two ways to accomplish the fundamental frequency change. One is simply to scale the average frame frequency $\bar{\omega}_{1_j}$ of each frame. The second is to form the kernel part of the cosine arguments (i.e. $\bar{\omega}_{1_j}[k+\beta_j(k)]$) such that the (discrete) instantaneous frequency of the fundamental is scaled by the wanted value. The second is subjectively more pleasing and is accomplished by taking the first difference (the derivative), scaling by the wanted value and cumulative summing (re-integrating) the result.

Time-scale modifications by factors of 2 and 1/2 have been coded and have been successfully demonstrated. The synthesis maintains the character of the person's voice for the voiced and mixed-voice segments. Time lengthening by more than a factor of approximately 1.5 results in some fricative and stop consonants sounding unnaturally long. Although not simulated, there is no reason why continually adaptive real-time time-scale modification from PMMM parameters could not easily be accomplished. Synthesis from the PMMM parameters

for this application would be trivial. In contrast, the previous work would require substantial real-time calculation to maintain the phase relationships of the sinusoids.

Pitch-scale modifications of up to 50% have been accomplished, although $\pm 30\%$ is typical. Previous work has had difficulty maintaining the mixed voice portions of these sounds, while the PMMM synthesis maintains the character of the speaker for these utterances.

Simultaneous time-scale and pitch change synthesis has been accomplished from PMMM parameters using both techniques simultaneously[3].

It should be noted here that the difficult part is obtaining the PMMM parameters. Once they have been obtained, time-scale and pitch change is easily accomplished with the PMMM parameterization. Thus, an important use of these PMMM parameterization properties might be in speech synthesis where time-scale and pitch changes are required for proper emphasis of synthetic speech.

[3]If the application is only time-scale or pitch modification (i.e., not speech enhancement), only one or two iterations of the PMMM parameter estimation algorithm would be required for adequate $a_{i,j}(k)$ estimates. Section 6.2 discusses this issue and the issue of required $\beta(k)$ accuracy for synthesis purposes.

Chapter 6

Other Potential Applications of the PMMM

This Chapter investigates other uses for PMMM processing. The first three sections, Co-Channel speaker investigation, low bit-rate harmonic coding, and "super-resolution" pitch analysis have been investigated during the course of this research. The last section, Vocal Physiology Abnormality Determination, is somewhat speculative.

6.1 Co-Channel Speaker or Sound Separation

Speaker or sound separation is a signal processing task that attempts to preserve acoustic components belonging to one speaker relatively intact while removing or recovering acoustic components of another speaker. Previous research in the speaker separation was previously mentioned in Section 4.1 in which the association of sounds to a specific speaker was made on the basis of the amplitude envelope of Formant Mode Modulation Model components. This separation is successful whenever a given speakers speech energy is dominant in a given spectral region. However, this technique is powerless whenever the goal is to recover speech components of one speaker which are lower in power than the dominant speaker.

Several two-speaker, speaker separation techniques have been developed for voiced speech [81, 137, 66, 67, 25, 122, 87, 147]. The essential premise of these techniques is that if one could track the individual pitch tracks for each speaker, then a harmonic association of both speaker's harmonics could be made. The speech for each individual could be obtained by comb filtering the combined

speech or by synthesis via model parameters. However, as noted by Quatieri in [147], "the further development of a pitch estimation algorithm, capable of handling summed waveforms of vastly different intensity levels, is 'critical' to sinusiodal speech separation and enhancement". As a result, "good" enhancement for his technique was only obtained for a target-to-interference ratio of 0 dB.

The PMMM reveals yet another reason for poor enhancement. This application requires high frequency resolution, which, in turn, requires long analysis frames. The harmonic structure of each person's speech is smeared by pitch quasi-periodicity, as was explained in the Spread Spectrum analogy of Section 2.8.

PMMM processing could be designed to "sharpen" one set of harmonics, filter out the harmonic energy, and then re-warp the signal to the original time base. As the present algorithm expects only one dominant harmonic sound, the $\beta(t)$ clustering algorithm would require further constraints to select (for clustering) only those components that are harmonics of the dominant speaker's pitch. A modification would be required to guide the $\beta(t)$ clustering based on this new information, as the present PMMM processing requires no *a priori* pitch information. Another difficulty is that in any given frame, the lower power speaker *may* be the speaker with the highest energy in that frame. Therefore, a pitch tracking algorithm for the highest power speaker would also be required to help segregate the speakers.

For the two speaker separation problem, the present embodiment of the PMMM enhancement works well to separate the dominant speaker when the dominant speaker to interfering speaker ratio is 5 dB or more. The PMMM processing in this case "locks onto" the dominant time-synchrony, that being the $\beta(t)$ of the dominant speaker. For an interfering source of a braodband type, such as the babble noise of Section 5.1.1, the present embodiment of the PMMM enhancement works well to about 0 dB SNR. Although this enhancement significantly

improves the performance of signal processing applications, such as the SI system, the wanted speech *before* the enhancement is still highly intelligible. The PMMM enhancement does not significantly add to the intelligibility for this case. It does, however, reduce the "fatigue factor" for listeners who would be required to listen to such speech for long periods of time.

To verify that PMMM-based separation is possible with negative signal to interference ratios, PMMM synthesis from the noisy speech was performed using PMMM parameters ($\beta_j(k)$ and $\bar{\omega}_{1_j}$) derived from clean (typical office environment) speech[1]. Separation from Babble noise was obtained down to -10 dB SNR and separation for the two speaker separation case was obtained down to -20 dB SNR. The best reported performance known to this author for the two speaker separation case is for "minimal" enhancement at -16 dB SNR [147].

6.2 Low Bit-Rate Harmonic Speech Coding

Section 5.2 described how to synthesize speech from PMMM parameters. The PMMM synthesis is essentially similar to the harmonic coders of Section 3.2 except that the fundamental frequency is parametrically allowed to vary *within* a frame. The benefits of this from a coding standpoint is that the bits that were used to transmit phase information and voiced/unvoiced information are not needed for the PMMM based coder in order to generate voiced or mixed voice sounds. The drawback is that we need to transmit a time warping function $\beta(t)$ in lieu of this information.

To answer the question of how many bits are needed to represent the $\beta(k)$ requires knowing what significant properties of the $\beta(k)$ function *need* to be transmitted from a perception standpoint. This is an important question, as the $\beta(k)$ functions have as many samples as the original speech and can take on very intricate shapes.

[1] Recall from Section 3.3.3 that the performance of many speech enhancement and speaker separation techniques are reported using parameters obtained from clean speech.

Several ensembles of $\beta(k)$ functions were collected for voiced speech and a Principal Component Analysis (PCA) [52][2] was performed on the $\beta(k)$ ensembles. For all such $\beta(k)$ ensembles studied, the principal component vectors (the eigenvectors of the autocorrelation matrix formed by the $\beta(k)$ ensembles), $\mathbf{v}_l$, approximated a sine wave series of integer multiples of half sine wave cycles. Therefore, the $\mathbf{v}_l$ were approximated by $\hat{\mathbf{v}}_l$ whose elements, $\hat{v}_l(k)$, were:

$$\hat{v}_l(k) \;=\; \sin(l\,\frac{\pi}{N}\,k),\;\; k=0,1,...(N-1),\; l=1,2,...N, \tag{6.1}$$

where N is the frame length in samples. These basis functions make intuitive sense, as the $\beta(t)$ functions are defined to begin and end at zero, as do the sine wave series in Equation 6.1. Each individual $\beta(k)$ can then be represented by a linear sum of the basis functions $\hat{\mathbf{v}}_l$ or:

$$\beta_j(k) \;=\; \sum_{l=1}^{N} \lambda_{l_j}\,\hat{v}_l(k) \tag{6.2}$$

where, typically, $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_N|$. Comparisons of synthesized speech using either the original $beta(k)$ or the $\beta(k)$ approximated by the first *two* basis elements of Equation 6.2 were found to be perceptually indistinguishable[3]. This implies that although significant detail is required in the $\beta(k)$ function to sharpen the harmonics, only the rough detail provided by the first two terms of the sine wave series is required for perceptual similarity[4].

The implication is that only three coefficients, $\bar{\omega}_{1_j}$, λ_{l_j} and λ_{2_j} are required for frame j to encode the entire excitation. Since λ_{2_j} is usually less than λ_{1_j}, but still in the same dynamic range, both λ's can be quantized with 11 bits or less[5]. The coding for the entire excitation, assuming 7 bits for $\bar{\omega}_{1_j}$ (generous) and a 32

[2]Also known as the Karhunen-Loève Transform (KLT) or the Hotelling Transform.

[3]The synthesis used 32 ms frames.

[4]This is the rationale for stating that only one or two PMMM iterations were necessary for the time-scale and pitch modification application of Section 5.2.1.

[5]This estimate is based on the authors experience with the precision required for $\beta(k)$ approximation.

ms frame size, requires less than 600 bps.

The spectral envelope information is contained in the $A_{i,j}$, in a manner completely analogous to the other harmonic coders mentioned. A simple coding of the spectral envelope would code approximately 16 of the most significant $A_{i,j}$'s at an average quantization of five bits per coefficient to yield approximately 2500 bps.

Thus, without serious attempts at reducing the bit rate, a 3100 bps coding is achievable with the PMMM. If we instead use Line Spectral Pairs (LSP) to encode the spectral envelope, recent speech research suggests that this coding could be made at approximately 24 bits/frame [132] to 32 bits/frame [184]. This would result in 750 to 1000 bps for the spectral envelope or 1350 to 1600 bps for the entire PMMM based coding.

6.3 Robust Technique for "Super-Resolution" Pitch Extraction

The PMMM fundamental frequency estimation technique of Section 2.5.3 demonstrated that PMMM processing produces a precise and robust fundamental frequency estimate for the frame j, $\bar{\omega}_{1_j}$.

A PMMM-based *instantaneous* pitch estimate can be formalized by taking the derivative of the phase function of the PMMM model of the first harmonic component, $\phi_{1_j}(t) = \bar{\omega}_{1_j}[t + \beta(t) + \Theta_1]$, resulting in:

$$\omega_{1_j}(t) = \bar{\omega}_{1_j}[1 + \frac{d\beta_j(t)}{dt}]. \tag{6.3}$$

So-called "super-resolution" pitch determination [112, 123] has sought to accomplish robust pitch determination in noise using a time warping approach that takes mild amplitude non-stationarity into account. In [112] the authors state that pitch variations in the range of 2-10% between two successive pitch periods are possible and that the glottal pulses are not uniform in terms of their amplitudes (volumes). Their fundamental premise is that a super resolution pitch

extractor must account for both amplitude and pitch non-stationarity and that their model incorporates this.

As PMMM processing accounts for both pitch and amplitude non-stationarity, the instantaneous pitch estimate afforded by PMMM processing may well outperform these and other "super-resolution" pitch tracking techniques. No detailed comparison of the PMMM instantaneous pitch estimates to these other pitch tracking techniques was performed. Only one iteration would probably be necessary for this application, as the harmonics out to approximately 2.5 kHz are separated enough for analytic signal generation after one iteration.

6.4 Vocal Physiology Abnormality Determination

Recent surgical techniques have made major progress in improving the wanted outcomes from vocal tract and vocal fold surgery. Present signal processing techniques are inadequate to measure the harmonic-to-noise ratio (clarity of voice), the perturbations of cycle-to-cycle amplitude ("shimmer") and cycle-to-cycle frequency ("jitter") for diagnostic purposes [166]. For example, in studies of fatigue in trained singers, the difference between a rested and tired voice is usually obvious to the ear, but these changes cannot be detected consistently with the test equipment available to laryngologists.

The PMMM has demonstrated an ability to estimate the detailed nature of the aperiodicity of voiced sounds. Experimentation with PMMM synthesis of the previous section revealed that if the "estimated" $\beta(t)$ were replaced with a larger form of aperiodicity, say $2\beta(t)$, that the synthesis had a more "harsh" or more hoarse quality to it. Replacing the $\beta(t)$ with a lower amplitude $\beta(t)$ results in the synthesis sounding more machine like. The time warping function $\beta(t)$ captures important aspects of the timbre of a person's voice.

The implication of this is that it may be possible to characterize the "raspiness" of a given persons voice while voicing a known phoneme. The phoneme /a/ as in father was used in [166]. Since the PMMM processing is the first long-term speech

model known to this author that does not introduce constraints on the amplitude or frequency stationarity in its processing, it may be of value as a diagnostic tool for laryngologists.

Chapter 7

Summary, Innovations, and Future Research

7.1 The Problem and Solution

In this thesis, a new speech model, the Pitch Mode Modulation Model, was formulated and applications of it for speech enhancement and general speech processing were presented.

Although speech signals are known to be highly redundant, they are also non-stationary. Classical *short-term* speech models require that the model parameters be extracted from short signal segments, whose short duration make the estimation of the parameters difficult and sometimes obscures the very redundancy the model was based on.

Likewise, virtually all speech enhancement techniques critically depend upon some form of stationarity, either in the speech signal, the noise signal, or both. Due to the highly disparate statistical nature of noise signals and the non-stationarity of speech signals, most enhancement techniques are *short-term* techniques that focus on only a specific form of noise impairment.

In the PMMM, the components of speech were found to vary in time synchrony with a specific time warping function, called the beta function. The PMMM exploits the time-synchronous redundancy in speech to time-warp the speech waveform such that the quasi-periodic signal becomes periodic. After this *nonlinear* processing step, the speech signal can be enhanced by standard *linear* techniques (e.g. comb filtering and re-warping) or by synthesis from model parameters. Thus, processing energy is focused on the known speech signal properties, not on specific characteristics of the noise.

The PMMM is a *long-term* speech model that is able to model the non-stationarity found in normal voiced and mixed-voice speech. Unlike other long-term speech, it does not artificially constrain the parameters that model the non-stationarity to possess a specific form (e.g., a power series). It, therefore, can be used to model long segments of voiced speech necessary for many speech enhancement and speaker separation techniques.

7.2 Innovations of this Research

The innovations of this research are that:

1. The PMMM is the first long-term model able to associate, on the harmonic-level and in noisy environments, quasi-harmonic speech components and dis-associate them from non-speech components.
2. No specific parameterization of the time warping function (e.g., a time series) was used as a constraint in modeling the frequency non-stationarity of speech.
3. No specific parameterization of the amplitude of the PMMM components (e.g. constants) was used as a constraint in modeling the amplitude non-stationarity of speech.
4. There is sensitivity of the PMMM technique to *a priori* pitch estimates, as none are used.
5. The PMMM is the first mathematical formulation that describes the spectral effect of a coherent time warping of PMMM components using phase modulation principles.
6. A robust pitch estimate is obtained *after* PMMM processing, using only component energy *known* to be correlated (in phase modulation characteristic) to the other speech components.
7. The PMMM, in contrast to classical (short-term) speech models, has no need to characterize the vocal tract and excitation parameters separately.

The time warping function captures the frequency non-stationarity and the $a_i(t)$ the (time-varying) PMMM component amplitudes.

8. The PMMM is the first technique to obtain and qualitatively describe a spread spectrum-like processing gain for amplitude estimation of PMMM components in noise.

7.3 Proposed Areas of Continuing Research

The PMMM is novel from many speech modeling aspects. As such, it is expected that algorithmic improvements to the estimation of PMMM parameters for any specific application are possible. Modification of some procedures have already been mentioned in Chapters 5 and 6 for specific application areas. The major area for improvement is the fast and accurate estimation of the time warping function $\beta(t)$. This is because all PMMM processing improvement is predicated on accurate $\beta(t)$ estimation and this is, in its present embodiment, computationally intensive.

Knowing that the PMMM form exists, it may be possible to accomplish the same ends (i.e., estimating a time warping function) through a different means. For example, the beta function extraction methodology of Section 2.5 outlined the iterative method used in this research to estimate $\beta(t)$. To obtain robustness in noise, this estimation process was "slowed down" to insure that each iterative $\beta(t)$ estimate approached the wanted $\beta(t)$ in an appropriate progression. This is because, in the absence of noise, the optimum $\beta(t)$ can be formulated as the following *nonlinear* optimization problem for frame j:

Find $\beta(t)$ *to minimize :*

$$\mathcal{E}_j = \sum_{k=0}^{N-1} \|\check{s}_j(k) - s_j(k)\|^{\mathcal{L}} \tag{7.1}$$

Subject to:

$$\tilde{s}_j(k) = \sum_{i=1}^{M} a_{i_j}(k)\cos(\phi_{i_j}(k)), \tag{7.2}$$

and

$$\phi_{i_j}(k) = \bar{\omega}_{1_j}[t + \beta_j(k) + \Theta_{i_j}]\,, \tag{7.3}$$

where $\|\cdot\|^{\mathcal{L}}$ is some norm of the error, N is the frame length in samples, and M is the number of quasi-harmonics in the decomposition.

The Principal Component Analysis technique used in the low-bit rate harmonic speech coding section (Section 6.2) suggests that an iterative decomposition of $\beta(k)$ into a sine wave series is possible. It may be possible to design a $\beta(k)$ estimation technique using such a decomposition. Or perhaps an EM technique might be employed to devise such a method [35, 23, 43]. If phase estimates of the harmonics are not calculated in such a method, another metric will need to be developed to determine if the time warping function estimate, $\beta(k)$, is improving. Such a measure might be a spectral sharpening measure, similar in spirit to one used in Almeida's model [2].

Another possible improvement might be the use of the PMMM processing in existing speech enhancement techniques. For example, the single-channel adaptive noise canceling arrangement of Section 1.5.4.1 was predicated on the pitch period to pitch period stationarity of the speech signal. PMMM-like processing could be designed into the system in order to make the signal and the reference more stationary.

References

[1] J. B. Allen, "Cochlear Modeling", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–33, no. 2, pp. 3–29, January 1985.

[2] L. B. Almeida and J. M. Tribolet, "Nonstationary Spectral Modeling of Voiced Speech", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–31, no. 3, pp. 664–677, June 1983.

[3] K. T. Assaleh and R. J. Mammone, "Spectral-Temporal Decomposition of Multi-component Signals", *IEEE SP Workshop on Statistical Signal and Array Processing*, Victoria, Canada, October 1992.

[4] K. T. Assaleh and R. J. Mammone, "Spectral-temporal decomposition of Multi-component signals", *IEEE 6th SP Workshop on SSAP*, Victoria, Canada, 1992.

[5] K. T. Assaleh, R. J. Mammone, and J. L. Flanagan, "Speech Recognition Using the Modulation Model", *Proc. ICASSP–93*, vol. 2, pp. 664–667, 1993.

[6] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *J. Acoust. Soc. Am.*, vol. 50, no. 2, pt. 2, pp. 637–655, August 1971.

[7] B. S. Atal and J. R. Remde, "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates", *Proc. ICASSP–82*, pp. 614–617, 1982.

[8] E. Bedrosian, "A product theorem for Hilbert transform", *Proc. IEEE* (letters), vol. 51, no. 5, pp. 868–869, May 1963.

[9] L. L. Beranek, *Acoustics,* New York, NY: McGraw-Hill, 1954.

[10] C. C. Blackburn and M. B. Sachs, "The Representation of Steady State Vowel Sound /EH/ in the Discharge Patterns", *Journal of Neurophysiology*, vol. 63, pp. 1191–1212, 1990.

[11] R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra*, New York, NY: Dover Publications, 1958.

[12] S. F. Boll, "Suppression of Noise in Speech Using the SABER Method", *Proc. ICASSP–78*, pp. 606–609, 1978.

[13] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–27, no. 2, pp. 113–120, April 1979.

[14] S. F. Boll and D. C. Pulsipher, "Suppression of Acoustic Noise in Speech Using Two Microphone Adaptive Noise Cancellation", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–28, no. 6, pp. 752–753, December 1980.

[15] B. Boashash, *Time-Frequency Signal Analysis – Methods and Applications*, New York, NY: Longman Cheshire, 1991.

[16] B. Boashash, "Estimating and Interpreting the Instantaneous Frequency of a Signal – Part 2: Algorithms and Applications", *Proc. IEEE*, vol. 80, no. 4, pp. 540–568, 1992.

[17] B. Boashash, editor, *Time-Frequency Signal Analysis, Methods and Application*, New York, NY: John Wiley & Sons, 1992.

[18] A. S. Bregman, *Auditory Scene Analysis, The Perceptual Organization of Sound*, Cambridge, MA: MIT Press, 1990.

[19] J. P. Burg, "Maximum Entropy Spectral Analysis", *Proc. of 37th Meeting of the Society of Exploratory Geophysics*, 1967.

[20] J. P. Burg, "A New Analysis Technique for Time Series Data", *Proc. NATO Advanced Study Institute on Signal Processing*, Enschede, Netherlands, 1968.

[21] I. S. Burnett and R. J. Holbeche, "A Mixed Prototype Waveform/CELP Coder for Sub 3Kb/s", *Proc. ICASSP–93*, vol. 2, pp. 175–178, 1993.

[22] J. P. Campbell, Jr., "Features and Measures for Speaker Recognition", Ph.D. Thesis, Oklahoma State University, December 1992.

[23] D. Chazan, Y. Stettiner, and D. Malah, "Optimal Multi-Pitch Estimation Using the EM Algorithm for Co-Channel Speech Separation", *Proc. ICASSP–93*, vol. 2, pp. 728–731, 1993.

[24] D. Chien, "Automated Two Speaker Separation with Variable Frame Size Transform", M.S. Thesis, East Texas State University, 1988.

[25] D. G. Childers and C. K. Lee, "Co-Channel Speech Separation", *Proc. ICASSP–87*, vol. 1, pp. 181–184, April 1987.

[26] J. M. Chowning, "The Synthesis of Complex Audio Spectra by Means of Frequency Modulation", *J. Audio Engineering Society*, vol. 21, no. 7, pp. 526–534, September 1973.

[27] J. M. Chowning, "Frequency Modulation Synthesis of the Singing Voice", Chapter 6 in *Current Directions in Computer Music*, M. Mathews and J. Pierce, editors, MIT Press, Cambridge, MA, 1991.

[28] P. L. Chu, "Quadrature Mirror Filter Design for an Arbitrary Number of Equal Bandwidth Channels", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–33, no. 1, pp. 203-218, February 1985.

[29] C. K. Chui, editor, *Wavelets: A Tutorial in Theory and Applications*, San Diego, CA: Academic Press, Inc., 1992.

[30] L. Cohen, "What is a multicomponent signal?", *Proc. ICASSP–92*, San Francisco, CA, vol. 5, pp. V-113 – V-116, 1992.

[31] C. E. Cook, et al., editors, *Spread Spectrum Communications*, New York, NY: IEEE Press, 1983.

[32] R. A. Curtis and R. J. Niederjohn, "An Investigation of Several Frequency-Domain Processing Methods for Enhancing the Intelligibility of Speech in Wideband Random Noise", *Proc. ICASSP–78*, pp. 602–605, 1978.

[33] A. de Cheveeigne, "Separation of Concurrent Harmonic Sounds: Fundamental Frequency Estimation and a Cancellation Model of Auditory Processing", *J. Acoust. Soc. Am.*, in press, release late 1993.

[34] J. R. Deller Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, New York, NY: Macmillan Publishing Company, 1993.

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society*, vol. 39, 1–88, 1977.

[36] C. DiMonte and K. S. Arum, "Tracking the Frequencies of Superimposed Time-Varying Harmonics", *Proc. ICASSP–90*, vol. E, pp. 2539–2542, 1990.

[37] G. R. Doddington, "Speaker Recognition - Identifying People by their Voices", *Proc. IEEE*, vol. 73, no. 11, pp. 1651–1663, November 1985.

[38] E. Dorkena and S. H. Nawab, "Time-Frequency Analysis of Non-Stationary Harmonic Sounds", *Proc. ICASSP–93*, vol. 3, pp. 249–252, 1993.

[39] D. F. Elliott, editor, *Handbook of Digital Signal Processing, Engineering Applications*, San Diego, CA: Academic Press, Inc., 1987.

[40] Y. Ephraim, "Statistical-Model-Based Speech Enhancement Systems", *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1554, October 1992.

[41] C. G. M. Fant, "On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies", In *For Roman Jakobson*, The Hague, The Netherlands: Mouton, 1956.

[42] C. G. M. Fant, *Acoustic Theory of Speech Production*, The Hague, The Netherlands: Mouton, 1960.

[43] M. Feder and E. Weinstein, "Parameter Estimation of Superimposed Signals Using the EM Algorithm", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, no. 4, pp. 477–489, April 1988.

[44] A. B. Fineberg and R. J. Mammone, "A method for instantaneous frequency tracking of multiple narrowband signals", *Signal Processing*, vol. 29, no. 1, pp. 29–44, October 1992.

[45] A. B. Fineberg, R. J. Mammone, and J. L. Flanagan, "An Application of the Modulation Model to Speech Recognition", *Proc. ICASSP–92*, vol. S1, pp. I-541–I-545.

[46] J. L. Flanagan, "Parametric coding of speech spectra", *J. Acoust. Soc. Am.*, vol. 68, no. 2, pp. 412–419, August 1980.

[47] J. L. Flanagan and S. W. Christensen, "Computer studies on parametric coding of speech spectra", *J. Acoust. Soc. Am.*, vol. 68, no. 2, pp. 420–430, August 1980.

[48] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, Third Printing, New York, NY: Springer-Verlag, 1983.

[49] S. Furui and M.M. Sondhi, editors, *Advances in Speech Signal Processing"*, New York, NY: Marcel Dekker, Inc., 1989.

[50] S. Furui, *Digital Speech Processing, Synthesis and Recognition*, New York, NY: Marcel Dekker, Inc., 1989.

[51] S. A. Gelfand, *Hearing: An Introduction to Psychological and Physiological Acoustics*, Second Edition, New York, NY: Marcel Dekker, Inc., 1990.

[52] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Boston, MA.: Kluwer Academic Publishers, 1992.

[53] O. Ghitza, "A Measure of In-Synchrony Regions in the Auditory Nerve Firing Patters as a Basis for Speech Vocoding", *Proc. ICASSP–85*, vol. 2, pp. 505–508, 1985.

[54] O. Ghitza, "Temporal Non-Place Information in the Auditory-Nerve Firing Patterns as a Front-End for Speech Recognition in a Noisy Environment", *J. Phonetics*, vol. 16, pp. 109–124, 1988.

[55] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain", *J. Acoust. Soc. Am.*, vol. 46, no. 2., pt. 2, pp. 442–448, 1969.

[56] J. T. Graf and N. Hubing, "Dynamic Time Warping Comb Filter for the Enhancement of Speech Degraded by White Gaussian Noise", *Proc. ICASSP–93*, vol. 2, pp. 339–342, 1993.

[57] R. M. Gray and L. D. Davidson, *Random Processes: A Mathematical Approach for Engineers*, Englewood Cliffs, NJ: Prentice-Hall, 1986.

[58] D. W. Griffin, *The Multi-Band Excitation Vocoder*, unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA, p. 23, 1987.

[59] D. W. Griffin and J. S. Lim, "The Multiband Excitation Vocoder", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 8, pp. 1223–1236, August 1988.

[60] Y. H. Gu and W. M. G. van Bokhoven, "Co-Channel Speech Separation Using Non-linear Adaptive Filtering", *Proc. ICASSP–91*, pp. 949–952, 1991.

[61] S. Handel, *Listening: An Introduction to the Perception of Auditory Events*, Cambridge, MA: MIT Press, 1989.

[62] J. H. L. Hansen and M. A. Clements, "Enhancement of Speech Degraded by Non-White Additive Noise", Final Technical Report submitted to Lockheed Corp., DSPL–85–6, Georgia Institute of Technology, Atlanta, GA, August, 1985.

[63] J. H. L. Hansen and M. A. Clements, "Iterative Speech Enhancement with Spectral Constraints", *Proc. ICASSP–87*, vol. 1, pp. 189–192, 1987.

[64] J. H. L. Hansen and M. A. Clements, "Constrained Iterative Speech Enhancement With Application to Automatic Speech Recognition", *Proc. ICASSP–88*, pp. 561–564, 1988.

[65] J. H. L. Hansen and M. A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition", *IEEE Trans. Signal Processing*, vol. 39, no. 4, pp. 795–805, April 1991.

[66] B. A. Hanson, "Processing Techniques for Intelligibility Improvement to Speech with Co-Channel Interference", *Speech Technology Lab, Final Report*, F30602–81–C–0226, August 1983.

[67] B. A. Hanson, "The Harmonic Magnitude Suppression (EMS) Technique for Intelligibility Enhancement in the Presence of Interfering Speech", *Proc. ICASSP–84*, pp. 18A.5.1–18A.5.4, 1984.

[68] S. Haykin, *Adaptive Filter Theory,* Englewood Cliffs, NJ: Prentice-Hall, 1986.

[69] S. Haykin, *Digital Communications*, New York, NY: John Wiley & Sons, 1988.

[70] D. J. Hermes, "Measurement of Pitch by Subharmonic Summation", *J. Acoust. Soc. Am.*, vol. 83, no. 1, pp. 257–264, January 1988.

[71] D. J. Hermes and J. C. van Gestel, "The frequency scale of speech intonation", *J. Acoust. Soc. Am.*, vol. 90, no. 1, pp. 97–102, July 1991.

[72] W. Hess, *Pitch Determination of Speech Signals*, New York, NY: Springer-Verlag, 1983.

[73] M. L. Honig and D. G. Messerschmitt, *Adaptive Filters: Structures, Algorithms and Applications*, Boston, MA: Kluwer Academic Publishers, 1984.

[74] A. N. Ince, editor, *Digital Speech Processing: Speech Coding, Synthesis and Recognition*, Boston, MA: Kluwer Academic Publishing, 1992.

[75] F. Itakura, "Line spectrum representation of linear prediction coefficients of speech signals", *J. Acoust. Soc. Am.*, vol. 57, p. S-35, 1975.

[76] R. W. Johnson and J. E. Shore, "Minimum Cross-Entropy Spectral Analysis of Multiple Signals", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–31, pp. 574–582, 1983.

[77] G. Jones and B. Boashash, "Instantaneous Frequency, Instantaneous Bandwidth and the Analysis of Multicomponent Signals", *Proc. ICASSP–90*, vol. E, pp. 2467–2470, 1990.

[78] B-H. Juang, "Speech Recognition in Adverse Environments", *Computer Speech and Language*, pp. 275–294, 1991.

[79] J. M. Kates, "A Time-Domain Digital Cochlear Model", *IEEE Trans. Signal Processing*, vol. 39, no. 12, pp. 2573–2592, December 1991.

[80] S. Kay, "Statistically/Computationally Efficient Frequency Estimation", *Proc. ICASSP–88*, pp. 2292–2295, 1988.

[81] J. F. Kaiser, "Reproducing the Cocktail Party Effect", 59th Meeting of the Acoustical Society of America, Providence, Rhode Island, June 9-11, 1960.

[82] G. E. Kopec and M. A. Bush, "An LPC-Based Spectral Similarity Measure for Speech Recognition in the Presence of Co-Channel Speech Interference", *Proc. ICASSP–89*, pp. 270–273, 1989.

[83] R. Kumaresan and D. W. Tufts, "Estimating the Parameters of Exponentially Damped Sinusoids and Pole-Zero Modelling in Noise", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–30, no. 6, pp. 833–840, December 1982.

[84] R. Kumaresan, C. S. Ramalingam, and A. G. Sadasiv, "On Accurately Tracking the Harmonic Components' Parameters in Voiced-Speech Segments and Subsequent Modeling by a Transfer Function", *Proc. 1992 Circuits, Systems and Computers Conference*, pp. 1645–1649, October 26–28, 1992.

[85] R. Kumaresan and C. S. Ramalingam, "Formant Tones: Weak Components at Formant Frequencies Actively Generated in the Vocal Tract", *Proc. ICASSP–93*, vol. 4, pp. 224–227, 1993.

[86] B. P. Lathi, *Modern Digital and Analog Communication Systems*, New York, NY: CBS College Publishing, 1983.

[87] C. K. Lee and D. G. Childers, "Cochannel speech separation", *J. Acoust. Soc. Am.*, vol. 83, no. 1, pp. 274–280, January 1988.

[88] E. A. Lee and D. G. Messerschmitt, *Digital Communications*, Boston, MA: Kluwer Academic, 1988.

[89] J. P. Lefevre and O. Passien, "Efficient Algorithms for Obtaining Multipulse Excitation for LPC Coders", *Proc. ICASSP–85*, vol. 2, pp. 957–960, 1985.

[90] J. S. Lim and A. V. Oppenheim, "All-Pole Modeling of Degraded Speech", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–26, pp. 197–210, June 1978.

[91] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", *Proc. IEEE*, vol. 67, pp. 1586–1604, December 1979.

[92] J. S. Lim, *Speech Enhancement*, Englewood Cliffs, NJ: Prentice-Hall, 1983.

[93] J. S. Lim and D. W. Griffin, "A New Model-Based Speech Analysis/Synthesis System", *NSA Speech Research Symposium*, vol. 2, San Diego, CA, Jan. 1985.

[94] R. F. Lyon and C. Mead, "An Analog Electronic Cochlea", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 7, pp. 1119–1134, 1988.

[95] C. Ma, Y. Kamp and L. F. Willems, "Robust signal selection for linear prediction analysis of voiced speech", *Speech Communication*, vol. 12, no. 1, pp. 69–81, March 1993.

[96] J. Makhoul, "Linear Prediction: A Tutorial Review", *Proc. IEEE*, vol. 63, pp. 561–580, 1975.

[97] J. Makhoul, "Stable and Efficient Lattice Methods for Linear Prediction", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–25, no. 5, pp. 423–428, October 1977.

[98] D. Malah and R. V. Cox, "Time-Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–27, pp. 121–133, Apr. 1979.

[99] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech Nonlinearities, Modulations, and Energy Operators", *Proc. ICASSP–91*, Toronto, Canada, pp. 421–424, 1991.

[100] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On separating amplitude from frequency modulations using energy operators", *Proc. ICASSP–92*, San Francisco, CA, pp. II-1–II-4, 1992.

[101] J. D. Markel, "Digital Inverse Filtering — A New Tool for Automatic Formant Estimation", *IEEE Trans. Audio Electroacoust.*, vol. AU–20, no. 2, pp. 129–137, June 1972.

[102] J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation", *IEEE Trans. Audio Electroacoust.*, vol. AU–20, no. 5, pp. 367–377, December 1972.

[103] J. D. Markel and A. H. Gray, Jr., "On Autocorrelation Equations as Applied to Speech Analysis", *IEEE Trans. Audio Electroacoust.*, vol. AU–21, pp. 69–79, April 1973.

[104] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, New York, NY: Springer-Verlag, 1976.

[105] J. S. Marques and L. B. Almeida, "Frequency-Varying Sinusoidal Modeling of Speech", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–37, no. 5, pp. 763–765, May 1989.

[106] S. McAdams, *Spectral Fusion, Spectral Parsing, and the Formation of Auditory Images*, unpublished doctoral dissertation, Stanford University, page 200, 1984.

[107] R. J. McAulay and M. L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–28, pp. 137–145, April 1980.

[108] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on Sinusoidal Representation", *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP–34, no. 4, pp. 744–754, August 1986.

[109] R. J. McAulay, T. M. Parks, T. F. Quatieri, and M. Sabin, "Sine-wave Amplitude Coding at Low Data Rates", *IEEE Workshop in Speech Coding*, Vancouver, B.C., Canada, September 1989.

[110] R. J. McAulay and T. F. Quatieri, "Pitch Estimation and Voicing Detection Based on a Sinusoidal Speech Model", *Proc. ICASSP–90*, pp. 249–252, 1990.

[111] R. J. McAulay and T. F. Quatieri, "The Application of Subband Coding to Improve Quality and Robustness of the Sinusoidal Transform Coder", *Proc. ICASSP–93*, vol. 2, pp. 439–442, 1993.

[112] Y. Medan, E. Yair, and D. Chazan, "Super Resolution Pitch Determination of Speech Signals", *IEEE Trans. Signal Processing*, vol. 39, no. 1, pp. 40–48, January 1991.

[113] Members of the Technical Staff of Bell Laboratories, *Transmission Systems for Communications* ("Blue Book"), Holmdel, NJ: Bell Laboratories Publications, 1982.

[114] J. M. Mendel, *Optimal Seismic Deconvolution*, New York, NY: Academic Press, 1983.

[115] K. Min, D. Chien, S. Li, and C. Jones, "Automated Two Speaker Separation System", *Proc. ICASSP–88*, pp. 537–540, 1988.

[116] T. B. Minde, T. Wigren, J. Ahlberg, and H. Hermansson, "Techniques for Low Bit Rate Speech Coding Using Long Analysis Frames", *Proc. ICASSP–93*, vol. 2, pp. 604–607, 1993.

[117] O. M. M. Mitchell, C. A. Ross, and G. H. Yates, "Signal Processing for a Cocktail Party Effect", *J. Acoust. Soc. Am.*, vol. 50, no. 2, pt. 2, pp. 656–660, August 1971.

[118] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, Second Edition, London, England:Academic Press, 1982.

[119] B. C. J. Moore and B. R. Glasberg, "Formulae Describing Frequency Selectivity as a Function of Frequency and Level, and their Use in Calculating Excitation Patterns", *Hearing Research*, vol. 28, pp. 209–225, 1987.

[120] B. C. J. Moore and R. D. Patterson, editors, *Auditory Frequency Selectivity*, New York, NY: Plenum Press, 1986.

[121] B. R. Musicus and J. S. Lim, "Maximum Likelihood Parameter Estimation on Noisy Data", *Proc. ICASSP–79*, pp. 224–227, April 1979.

[122] J. A. Naylor and S. F. Boll, "Techniques for Suppression of an Interfering Talker in Co-channel Speech", *Proc. ICASSP–87*, pp. 6.12.1–6.12.4, 1987.

[123] H. Ney, "Time Warping Approach to Fundamental Period Estimation", *IEEE Trans. Systems, Man, Cybernetics*, vol. SMC-12, no. 3, pp. 383–388, May/June 1982.

[124] A. M. Noll, "Cepstrum pitch determination", *J. Acoust. Soc. Am.*, vol. 41, pp. 293–309, 1976.

[125] A. H. Nuttall and E. Bedrosian, "On the Quadrature Approximation to the Hilbert Transform of Modulated Signals", *Proc. IEEE,* (letters), vol. 54, no. 10, pp. 1458–1459, October 1966.

[126] A. V. Oppenheim, "A Speech Analysis-Synthesis System Based on Homomorphic Filtering", *J. Acoust. Soc. Am.*, vol. 45, pp. 458–465, February 1969.

[127] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1989.

[128] S. J. Orfanidis, *Optimum Signal Processing: An Introduction*, New York, NY: McGraw-Hill, 1990.

[129] D. O'Shaughnessy, *Speech Communication: Human and Machine*, Reading, MA: Addison-Wesley, 1987.

[130] D. O'Shaughnessy, "Enhancing Speech Degraded by Additive Noise or Interfering Speakers", *IEEE Communications Magazine*, pp. 46–52, February 1989.

[131] K. K. Paliwal and A. Basu, "A Speech Enhancement Method Based on Kalman Filtering", *Proc. ICASSP–87*, vol. 1, pp. 177–180, 1987.

[132] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame", *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 1, pp. 3–14, January 1993.

[133] A. Papoulis, *Signal Analysis*, New York, NY: McGraw-Hill, 1977.

[134] A. Papoulis, "Random Modulation: A Review", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, no. 1, pt. 1, pp. 96–105, February 1983.

[135] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, New York, NY: McGraw-Hill, 1984.

[136] T. W. Parks and B. A. Weisburn, "Classification of Whale and Ice Sounds With a Cochlear Model", *Proc. ICASSP–92*, vol. 2, pp. II–481 to II–484, 1992.

[137] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection", *J. Acoust. Soc. Am.*, vol. 60, pp. 911–918, 1976.

[138] S. Parthasarathy and D. W. Tufts, "Excitation-Synchronous Modeling of Voiced Speech", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–35, no. 9, pp. 1241–1249, September 1987.

[139] D. B. Paul, "The Spectral Envelope Estimation Vocoder", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–29, no. 4, pp. 786–794, August 1981.

[140] Y. M. Perlmutter, L. D. Braida, R. H. Frazier, et al., "Evaluation of a Speech Enhancement System", *Proc. ICASSP–77*, pp. 212–215, 1977.

[141] T. L. Peterson and S. T. Boll, "Acoustic Noise Suppression in the Context of a Perceptual Model", *Proc. ICASSP–81*, pp. 1086–1088, 1981.

[142] R. H. Pettit, *ECM and ECCM Techniques for Digital Communications Systems*, Belmont, CA: Lifetime Learning Publications, 1982.

[143] L. A. Pipes, *Applied Mathematics for Engineers and Physicists*, New York, NY: McGraw-Hill, 1958.

[144] R. D. Preuss, "A Frequency Domain Noise Cancellation Preprocessor of Narrow-band Speech Communications Systems", *Proc. ICASSP–79*, pp. 212–215, 1979.

[145] J. G. Proakis, *Digital Communications*, New York, NY: McGraw-Hill, 1983.

[146] J. G. Proakis and D. G. Manolakis, *Introduction to Digital Signal Processing*, New York, NY: Macmillan Publishing Company, 1988.

[147] T. F. Quatieri and R. G. Danisewicz, "An Approach to Co-Channel Talker Interference Suppression Using a Sinusoidal Model for Speech", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 1, pp. 56–69, January 1990.

[148] T. F. Quatieri and R. J. McAulay, "Shape Invariant Time-Scale and Pitch Modification of Speech", *IEEE Trans. Signal Processing*, vol. 40, no. 3, pp. 497–510, March 1992.

[149] L. R. Rabiner and R. W. Schafer, "On the Behavior of Minimax FIR Digital Hilbert Transformers", *Bell Syst. Tech. Jour.*, vol. 53, no. 2, pp. 363–390, 1974.

[150] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the End Points of Isolated Utterances", *Bell Syst. Tech. J.*, vol. 54, no. 2, pp. 297–315, 1975.

[151] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ: Prentice-Hall, 1978.

[152] L. R. Rabiner and B-H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[153] R. P. Ramachandran, M. Zilovic, and R. J. Mammone, "Alternative Linear Predictive Analysis Techniques", Submitted to: *IEEE Intl. Symp. on Circuits and Systems,* London, England, June 1994.

[154] M. A. Ramalho, "Modulation Model-Based Speaker/Sound Separation", Chapter in "1991/1992 Annual Report for the Project A Modulation Model for Speaker Identification", CAIP Research Center/Rutgers University report to the Rome Laboratory of the US Air Force, October, 1992.

[155] M. A. Ramalho and R. J. Mammone, "New Speech Enhancement Techniques Using the Pitch Mode Modulation Model", Rutgers University, Piscataway, NJ, CAIP Technical Report, CAIP–TR–155, February 12, 1993.

[156] M. A. Ramalho and R. J. Mammone, "New Speech Enhancement Techniques Using the Pitch Mode Modulation Model", *Proc. of the* 36^{th} *Midwest Symposium on Circuits and Systems*, Detroit, MI, August 16–18, 1993.

[157] R. W. Ramirez, *The FFT: Fundamentals and Concepts*, Englewood Cliffs, NJ: Prentice-Hall, 1985.

[158] N. S. Reddy and M. N. Swamy, "High-Resolution Formant Extraction from Linear-Prediction Phase Spectra", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–32, no. 6, pp. 1136–1144, 1984.

[159] U. L. Rhode and T. T. N. Bucher, *Communications Receivers: Principles and Design*, New York, NY: McGraw-Hill, 1988.

[160] S. Rice, "Envelopes of Narrow-Band Signals", *Proc. IEEE*, vol. 70, no. 7, pp. 692–699, July 1982.

[161] O. Rioul and M. Vetterli, "Wavelets and Signal Processing", *IEEE SP Magazine*, vol. 8, no. 4, pp. 14–38, October 1991.

[162] M. A. Richards, "Helium Speech Enhancement Using the Short-Time Fourier Transform", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–30, no. 6, pp. 841–853, December 1982.

[163] A. W. Rihaczek and E. Bedrosian, "Hilbert Transform and the Complex Representation of Real Signals", *Proc. IEEE*, (letters), vol. 54, no. 3, pp. 434–435, March 1966.

[164] M. J. Ross et al., "Average Magnitude Difference Function Pitch Extractor", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–25, no. 1, pp. 24–33, Feb. 1977.

[165] M. Sambur, "Adaptive Noise Cancelling for Speech Signals", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–26, no. 5, pp. 419–423, October 1978.

[166] R. T. Sataloff, "The Human Voice", *Scientific American*, vol. 267, no. 6, pp. 108–115, December 1992.

[167] B. Sayers and E. C. Cherry, "Mechanism of binaural fusion in the hearing of speech", *J. Acoust. Soc. Am.*, vol. 29, pp. 973-987, 1957.

[168] R. W. Schafer and L. R. Rabiner, "System for automatic analysis of voiced speech", *J. Acoust. Soc. Am.*, vol. 47, no. 2, pt. 2, pp. 634–648, February 1970.

[169] J. Schroeder and R. Yarlagadda, "Linear Predictive Spectral Estimation via the $\mathcal{L}_\infty$ Norm", *Signal Processing*, vol. 17, no. 1, pp. 19–29, May 1989.

[170] M. R. Schroeder, "Recognition of Complex Acoustic Signals", *Life Science Research Reports*, vol. 55, pp. 323–328, November 1962.

[171] M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates", *Proc. ICASSP–85*, pp. 937–940, 1985.

[172] R. T. Schumacher, "Analysis of Aperiodicities in Nearly Periodic Waveforms", *J. Acoust. Soc. Am.*, vol. 91, no. 1, pp. 438–444, January 1992.

[173] M. Schwartz, *Information, Transmission, Modulation and Noise*, New York, NY: McGraw-Hill, 1980.

[174] H. E. Secker-Walker and C. L. Searle, "Time-domain analysis of auditory-nerve-fiber firing rates", *J. Acoust. Soc. Am.*, vol. 88, no. 3, pp. 1427–1436, September 1990.

[175] S. Seneff, "Pitch and Spectral Estimation of Speech Based on Auditory Synchrony Model", *Proc. ICASSP–84*, vol. 3, pp. 36.2.1–36.2.4, 1984.

[176] S. A. Shamma, "Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve", *J. Acoust. Soc. Am.*, vol. 78, no. 5, pp. 1612–1621, November 1985.

[177] S. A. Shamma, "Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve", *J. Acoust. Soc. Am.*, vol. 78, no. 5, pp. 1622–1630, November 1985.

[178] M. Sharma, "A New Method of Component Separation Without Filtering", Chapter in "1991/1992 Annual Report for the project A Modulation Model for Speaker Identification", CAIP Research Center/Rutgers University report to the Rome Laboratory of the US Air Force, October, 1992.

[179] M. Sharma, "A Modulation Model Based Approach to Keyword Spotting", Masters Thesis, Rutgers University, New Brunswick, NJ, May 1993.

[180] M. Sharma, "Modulation Model Based Features for Keyword Spotting", *Proc. of IEEE Workshop on Automatic Speech Recognition*, December 1993.

[181] J. E. Shore, "Minimum Cross-Entropy Spectral Analysis", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–29, no. 2, pp. 230–237, April 1981.

[182] S. Singhal and B. S. Atal, "Improving Performance of Mulit-Pulse LPC Coders at Low Bit Rates", *Proc. ICASSP–84*, vol. 1, pp. 131–134, 1984.

[183] F. K. Soong, and A. E. Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", *Proc. ICASSP–88*, pp. 871–879, 1988.

[184] F. K. Soong and B-H. Juang, "Optimal Quantization of LSP Parameters", *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 1, pp. 15–24, January 1993.

[185] J. Sorensen, M. Savic, H. Gao, and V. J. Amuso, "A Tracking Multi-speaker Model for Speech Separation Applications", *Proc. of Rutgers University Speech Research Symposium XII*, Rutgers University, Piscataway, NJ, 1992.

[186] D. Starer and A. Nehorai, "High-Order Polynomial Root Tracking Algorithm", *Proc. ICASSP–92*, San Francisco, CA, vol. 4, pp. 465–468, 1992.

[187] H. Stark and J. W. Woods, *Probability, Random Processes, and Estimation Theory for Engineers*, Englewood Cliffs, NJ: Prentice-Hall, 1986.

[188] Y. Stettiner, D. Malah, and D. Chazan, "Estimation of a Long-Term Model for Accurate Representation of Voiced Speech", *Proc. ICASSP–93*, vol. 2, pp. 534–537, 1993.

[189] F. G. Stremler, *Introduction to Communications Systems*, Reading, MA: Addison-Wesley Publishers, Series in Electrical Engineering, 1982.

[190] M. Sun and R. J. Sclabassi, "Discrete-Time Instantaneous Frequency and Its Computation", *IEEE Trans. Signal Processing*, vol. 41, no. 5, pp. 1867–1880, May 1993.

[191] H. Taub and D. L. Schilling, *Principles of Communications Systems*, New York, NY: McGraw-Hill, 1986.

[192] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1992.

[193] J. R. Treichler, C. R. Johnson, Jr., and M. G. Larimore, *Theory and Design of Adaptive Filters*, New York, NY: John Wiley & Sons, 1987.

[194] K. Tröndle and G. Söder, *Optimization of Digital Transmission Systems*, Norwood MA: Artech House, 1987.

[195] C. K. Un and K. Y. Choi, "Improving LPC Analysis of Noisy Speech by Autocorrelation Subtraction Method", *Proc. ICASSP–81*, pp. 1082–1085, 1981.

[196] P. P. Vaidyanathan, "Quadrature Mirror Filter Banks, M-Band Extensions and Perfect-Reconstruction Techniques", *IEEE ASSP Magazine*, vol. 4, no. 3, pp. 4-20, July 1987.

[197] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[198] L. M. Van Immerseel and J-P. Martens, "Pitch and voiced/unvoiced determination with an auditory model", *J. Acoust. Soc. Am.*, vol. 91, no. 6, pp. 3511–3526, June 1992.

[199] M. Vetterli and C. Herley, "Wavelets and Filter Banks: Theory and Design", Columbia University Center of Telecommunications Research Report, 1992.

[200] D. Wang, "The Unimportance of Phase in Speech Enhancement", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–30, no. 4, pp. 679–681, 1982.

[201] F-M Wang, et al., "Frequency Domain Adaptive Postfiltering for Enhancement of Noisy Speech", *Speech Communication*, vol. 12, no. 1, pp. 41–56, March 1993.

[202] K. Wang, S. A. Shamma, and W. J. Byrne, "Noise robustness in the auditory representations of speech signals", *Proc. ICASSP–93*, vol. 2, pp. 335–338, 1993.

[203] R. J. Webster, "Spectral Line Profiles Generated by Deterministic Frequency Modulation", *IEEE Trans. Signal Processing*, vol. 39, no. 4, pp. 1012–1017, April 1991.

[204] M. Weiss, E. Aschkenasy, and T. Parsons, "Processing Speech Signals to Attenuate Interference", *IEEE Symp. on Speech Recognition*, pp. 292–295, 1974.

[205] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1985.

[206] J. D. Wise, J. A. Caprio, and T. W. Parks, "Maximum Likelihood Pitch Estimation", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–24, no. 5, pp. 418–423, 1976.

[207] J. M. Wozencraft and I. M. Jacobs, *Principles of Communications Engineering*, New York, NY: John Wiley & Sons, 1965.

[208] X. Yang, K. Wang and S. A. Shamma, "Auditory Representations of Acoustic Signals", *IEEE Trans. Inform. Theory*, pp. 824–839, March 1992.

[209] B. Yegnanarayana, "Formant extraction from linear-prediction phase spectra", *J. Acoust. Soc. Am.*, vol. 63, no. 5, pp. 1638–1640, May 1978.

[210] W. A. Yost and C. S. Watson editors, *Auditory Processing of Complex Sounds*, Hillsdale, NJ: Lawrence Erlbaum Associates, 1987.

[211] G. Zweig, R. Lipes, and J. R. Pierce, "The cochlear compromise", *J. Acoust. Soc. Am.*, vol. 59, no. 4, pp. 975–982, April 1976.

Vita

Michael A. Ramalho

1975	Graduated from Middletown Township High School, Middletown, New Jersey.
1975-79	B.S. in Electrical Engineering, Rutgers University, College of Engineering, New Brunswick, New Jersey.
1979-83	Member of the Technical Staff, Bell Laboratories, Whippany, New Jersey.
1983	M.Eng. in Electrical Engineering, Cornell University, Ithaca, New York.
1984-86	Member of the Technical Staff, Bellcore, Morristown, New Jersey.
1986-91	District Manager, Bellcore, Morristown, New Jersey.
1991-92	Director, Bellcore, Morristown, New Jersey.
1992	Technical Program Vice-Chairman, ICC/Supercom '92, Chicago, Illinois.
1994	Ph.D. in Electrical Engineering, Rutgers University, Graduate School, New Brunswick, New Jersey.