Noise Suppression with Non-Air-Acoustic Sensors

by

David P. Messing

B.S. Electrical Engineering and Computer Science
U.C. Berkeley, 2001

SUBMITED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN ELECTRIAL ENGINEERING AND COMPUTER SCIENCE
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2003

Signature of Author:_____
Department of Electrical Engineering and Computer Science
August 28, 2003

Certified by:_____
Dr. Thomas F. Quatieri
Senior Technical Staff
MIT Lincoln Laboratory
Thesis Supervisor

Accepted by:_____
Dr. Arthur C. Smith
Professor of Electrical Engineering
Chairman, Departmental Committee on Graduate Students

Noise Suppression with Non-Air-Acoustic Sensors

by

David P. Messing

Submitted to the Department of Electrical Engineering on August 29, 2003 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Electrical Engineering and Computer Science

Abstract

Nonacoustic sensors such as the general electromagnetic motion sensor (GEMS), the physiological microphone (P-Mic), and the electroglottograph (EGG) offer multimodal approaches to speech processing and speaker and speech recognition. These sensors provide measurements of functions of the glottal excitation and, more generally, of the vocal tract articulator movements that are relatively immune to acoustic disturbances and can supplement the acoustic speech waveform. This thesis describes an approach to speech enhancement that exploits these nonacoustic sensors according to their capability in representing specific speech characteristics in different frequency bands. Frequency-domain sensor phase, as well as magnitude, is found to contribute to signal enhancement. Testing involves the time-synchronous multi-sensor DARPA Advanced Speech Encoding corpus collected in a variety of harsh acoustic noise environments. The enhancement approach is illustrated with examples that indicate its applicability as a pre-processor to low-rate vocoding and speaker authentication, and for enhanced listening from degraded speech.

Thesis Supervisor: Dr. Thomas F. Quatieri
Title: Senior Technical Staff, MIT Lincoln Laboratory

Acknowledgements

I thank all the members of the Information Systems Technology group of MIT Lincoln Laboratory for their support, good humor, and professional attitude. In particular, I thank my advisor Thomas Quatieri for the long hours he spent meeting with me, his light-hearted and humorous yet motivational attitude, his guidance, and instruction.

Special thanks also go to my office mate Nick Malyska for advice, numerous stimulating discussions, and his willingness to drive me between MIT and Lincoln at random hours.

I thank Kevin Brady, Joe Campbell, Bill Campbell, Michael Brandstein, Carl Quillen, Bob Dunn, Wade Shen, and Doug Sturim for help with random questions I had, advice, discussions, and enthusiastic support.

I thank our systems administrators Scott Briere and Steve Huff for keeping the machines and software I used up and running, and also helping me with printing issues.

I thank Misha Zitser for the late-night thesis writing companionship and Barry Jacobson for the bananas that allowed the two of us to function late at night.

I thank Cliff Weinstein and Marc Zissman for their support and funding of this project.

I thank my parents and two brothers for their support of my work and involvement at MIT.

Table of Contents

1 Introduction 7

2 Background and Framework 8

2.1 Speech 8

2.1.1 Voiced Speech 8

2.1.2 Unvoiced Speech 8

2.2 Noise 9

2.2.1 Stationary Noises 9

2.2.2 Non-stationary Noises 9

2.3 Noise Suppression 9

2.3.1 Motivation 9

2.3.2 Current Noise Suppression Algorithms 10

2.3.3 Baseline Algorithm Description 11

2.4 Sensor descriptions 15

2.4.1 GEMS sensor 15

2.4.2 PMIC sensor 15

2.4.3 EGG sensor 16

2.4.4 Resident Microphones 16

2.4.5 B&K microphone 16

2.5 Corpus and Environment Descriptions 17

2.5.1 Lawrence Livermore National Laboratory Corpus 17

2.5.2 Advanced Speech Encoding Corpus 17

2.6 Summary 18

3 Speech Activity Detection 19

3.1 Traditional Implementation and Problems 19

3.2 Multi-Sensor Approach 19

3.3 Speech Class Detection 20

3.3.1 Motivation 20

3.3.2 Classes of Speech 21

3.3.3 Detection of Classes 23

3.3.4 Data 24

3.4 Conclusion 24

4 Performance Bounds of Wiener Filtering 26

4.1 Background 26

4.2 Baseline System 26

4.3 Wiener Filter Magnitude Performance 29

4.3.1 Ideal Wiener Filter 29

4.3.2 Ideal Magnitude Test 30

4.4 Phase Performance 31

4.4.1 Original Magnitude and Clean Phase 32

4.4.2 Estimated Magnitude and Clean Phase 33

4.5 Conclusion 33

5 Magnitude Estimation of Speech Spectra 35
5.1 Baseline Algorithm 35
5.1.1 Algorithm Details 35
5.1.2 Baseline Algorithm Performance Examples 36
5.2 Processing Based on Different Speech Classes Using One Sensor 38
5.2.1 Division of Speech Classes 38
5.2.2 Filtering of Each Speech Class 40
5.3 Magnitude Estimation with Two Sensors 43
5.3.1 Two-Sensor Processing with Two Speech Class 43
5.3.2 Two Sensor Processing with Multiple Speech Classes 45
5.4 Conclusion 45

6 Phase Estimation 47
6.1 GEMS Phase 47
6.2 Other Sensor Phases 49
6.3 Synthetic Phase 49
6.4 Conclusion 55

7 Concluding Thoughts 57

A Detector Development 59
A.1 Metrics Used 59
A1.1 Mean-Square-Error (MSE) 59
A1.2 Segmental Signal to Noise Ratio (SSNR) 60
A.2 GEMS Dection 60
A.2.1 Inputs Used 60
A.2.2 Experiments 61
A.2.3 Conclusion of GEMS Detection 67
A.3 PMIC and GEMS Detection 67
A.3.1 Inputs Used 67
A.3.2 Experiments 68
A.3.3 PMIC and GEMS Detection Conclusion 71
A.4 Detection Conclusion 73

Chatper 1: Introduction and Organization

Linear filtering-based algorithms for additive noise reduction include spectral subtraction, Wiener filtering, and their adaptive renditions [NRC, 1989]. Nonlinear techniques have also arisen, for example, wavelet-based noise reduction systems [Donaho and Johnson, 1994] and suppression filters based on auditory models [Hanson and Nandkumar, 1995]. Although promising, these methods suffer from a variety of limitations such as requiring estimates of the speech spectrum and speech activity detection from a noisy acoustic waveform, distortion of transient and modulation signal components, and the lack of a phase estimation methodology.

In this thesis, we present an alternative approach to noise suppression that capitalizes on recent developments in nonacoustic sensors that are relatively immune to acoustic background noise, and thus provide the potential for robust measurement of speech characteristics. Early work in this area involved the use of the general electromagnetic motion sensor (GEMS) [Ng et al, 2000] [Burnett et al, 1999]. This thesis effort focuses on the GEMS but also investigates the physiological microphone (P-Mic) [Scanlon, 1998], and the electroglottograph (EGG) [Rothenberg, 1992]. These sensors can directly measure functions of the speech glottal excitation and, to a lesser extent, attributes of vocal tract articulator movements.

This thesis is divided into the five main chapters, chapters 2 – 6. Chapter 2 lays the framework of this work. It formulates the enhancement problem of interest and reviews a specific noise reduction algorithm based on an adaptive Wiener filter [Quatieri and Dunn, 2002]. It also describes the GEMS, P-Mic and EGG nonacoustic sensors, as well as the Lawrence Livermore Laboratory corpus and DARPA Advanced Speech Encoding Pilot Speech Corpus recorded in a variety of harsh noise environments. Chapter 3 presents an approach to speech activity detection based on different sensor modalities. Chapter 4 explores the limitations of the baseline reduction algorithm described in Chapter 2, and describes the affects of phase in segmented processing of speech. Chapter 5 provides a multimodal frame-based speech spectral magnitude enhancement scheme that utilizes the GEMS, P-Mic, and acoustic sensors in different frequency bands. Chapter 6 develops a multimodal spectral phase estimate integrated with the multi-band system of Chapter 5, and shows how it aids in speech enhancement and applications such as speech encoding. Finally, in Chapter 7, we summarize and discuss possible future research directions.

Chapter 2: Background and Framework

This chapter provides background for the major topics to be covered in this thesis. It begins with examining the characteristics of speech and discussing key differences between different speech sounds. Next, it discusses properties of stationary or non-stationary background noise and illustrates the detrimental effects of noise on speech and intelligibility. This chapter then reviews algorithms that have been used for speech enhancement, with an emphasis on a Wiener filter-based algorithm that had previously been developed at Lincoln Laboratory. The chapter concludes by discussing the corpora that were collected to conduct experiments that are discussed in subsequent chapters.

2.1 Speech

Speech is a very important part of everyday human life. It is essential for communications, exchanging ideas, and interacting. In life threatening situations, being able to understand speech can mean the difference between life and death. Much research has been conducted on speech production and speech modeling which has led to understanding of various speech sound classes. Two categories of speech that we shall focus on throughout this work are voiced speech and unvoiced speech.

2.1.1 Voiced Speech

Voiced speech is speech that requires the use of the vocal folds. Typical examples of voiced speech include vowels such as /a/, /Λ/ (as in "up"), /e/, /i/ (as in "eve"), /I/ (as in "it"), /o/, /u/ (as in "boot"), and /U/ (as in "foot"), and voiced consonants. Voiced consonants can further be divided into voiced fricatives such as /v/, /z/, voiced affricates such as /d3/ (as in the "j" in "just"), voiced plosives such as a /d/, /b/, /g/, glides such as /w/ and /y/, liquids such as /r/ and /l/, and nasals such as /m/ and /n/. Each of these subclasses of voiced speech are produced differently, yet they all share some similarities. In all of these examples and all voiced speech in general, air passes through the throat while the vocal folds vibrate, causing added excitation of the vocal tract and resulting in a vibrant and harmonic sounding speech.

2.1.2 Unvoiced Speech

Unvoiced speech is speech that does not involve vocal fold vibration and excitation. Typical examples of unvoiced speech include unvoiced fricatives such as /f/, /s/, /ʃ/ (as in the "sh" in "shag"), and /θ/ (as in the "th" in "thin"), whispers such as /h/, unvoiced affricates such as /tʃ/ (as in the "ch" in "choose"), and plosives such as /p/, /k/, and /t/. In all these sounds, a constriction in the vocal tract causes a build up of air which, when released, yields a "noisy" burst-like sound which is interpreted by humans as the onset of the consonant. Although shorter in duration than many voiced vowels, these unvoiced consonants are just as important, if not more, for speech intelligibility.

2.2 Noise

Noise hinders speech intelligibility. This makes it difficult to for a human or a machine to understand what is spoken and who spoke it. It is a very serious problem for people who work in high-noise environments such as a jet, helicopter, factory, or ambulance, and is also a factor in lower noise environments such as a crowded room with interfering conversations. This section focuses on two main types of noise - stationary noise and non-stationary noise.

2.2.1 Stationary Noises

The second order statistics of stationary noise does not change over time. Examples of stationary noise that approximately meet this condition are fan and engine noise. Also noise caused by a large number of conversations can often also be modeled as stationary noise.

2.2.2 Non-stationary Noises

The second order statistics of non-stationary noise change over time, often quite rapidly. Examples that meet this condition are very common and include gunfire, tank tread sounds, car horns, and other sudden noises. In the tests we conducted non-stationary noise was present in several of the environments - in particular the M2H tank environment and the urban warfare environment which had severe of gunfire (see below for further details of these environments). Often non-stationary and stationary noise is present together. For example a person may be driving a car at a constant speed while talking over a phone. In this case the car engine is a stationary noise source. If someone honks a car horn or police car sounds its siren briefly, a non-stationary noise is briefly introduced. Because of examples like this, often a challenge of real-world noise suppression algorithms is to be able to suppress non-stationary and stationary noises that are intermixed.

2.3 Noise Suppression

This section discusses suppression of noise and enhancement of speech. It begins elaborating on the motivation for noise suppression. Then it continues by discussing current methods employed to reduce noise, focusing on Wiener filtering. Then it describes in detail a Wiener filter algorithm previously developed at Lincoln Laboratory. It concludes by illuminating some of the issues that were focused on and resolved in subsequent chapters of this thesis.

2.3.1 Motivation

Noise suppression is essential in speech processing tasks such as speech encoding and speaker and speech recognition. Objectives of noise suppression algorithms include

making speech or encoded speech more intelligible to humans and more recognizable by machines. Traditionally, noise suppression of speech is accomplished in several stages. First, the sound source of interest must be recorded. Usually this is done by means of microphones that detect pressure of acoustic sounds. Consequently, when little acoustic noise is present these microphones record primarily the speech signal of interest. When background noise is present, however, both the desired speech and the background noise are recorded, reducing the intelligibility of the speech signal. This addition of noise can seriously affect tasks involving speech processing such as speaker identification and speech encoding. Certain speech enhancement algorithms aim to mitigate the ill effects of background noise and hopefully make speech more intelligible to both humans and computers. Although these techniques show promise, no useful intelligibility gains have been demonstrated on the degraded speech waveform itself, and limited intelligibility gains have been seen on encoded speech. One possible reason for these limitations is the inability of current suppression algorithms to capture fine temporal structure, either because of temporal blurring of speech events or because no attempt is made to estimate speech phase which contains important timing information, particularly in severe noise conditions.

The objective of this thesis is to aid noise suppression algorithms based on an acoustic signal by using non-air-acoustic measurements. A benefit of such an approach is that non-air-acoustic sensors are, ideally, unaffected by acoustic background noise.

2.3.2 Current Noise Suppression Algorithms

Many current algorithms used for noise suppression involve linear filtering to obtain a linear least squares estimate of the desired signal. Linear filtering algorithms for additive noise reduction include spectral subtraction, Wiener filtering, and their adaptive renditions [Ephraim and Malah, 1984] [Lim and Oppenhiem, 1979]. Because these algorithms use averages of signal and noise spectra, they can distort transient and time-varying speech components, such as plosive bursts, formant transitions, and vowel onsets that are essential for maintaining or increasing intelligibility while suppressing noise.

More recently, many nonlinear techniques have arisen including wavelet-based noise reduction systems that apply thresholding techniques to wavelet coefficients [Donahue, 1994] [Pinter, 1996]. Other nonlinear filtering methods include Occam filters based on data compression [Natarajan, 1995] and Teager and related quadratic energy-based estimation algorithms [Fang and Atlas, 1995], [Jabloun and Cetin, 1999]. These recent methods provide different approaches to preserving fine temporal structure of the speech signal while suppressing additive noise. Although these nonlinear techniques show promise, no useful intelligibility gains have been demonstrated.

Another approach to recovering the desired signal is to find a linear filter $h[n]$ such that the sequence $\hat{x}[n] = y[n] * h[n]$ minimizes the expected value of $(\hat{x}[n] - x[n])^2$ (here $\hat{x}[n]$ is an estimate of the clean speech signal, $x[n]$ is the clean speech, and $y[n]$ is the noisy input speech). The solution to this optimization problem in the frequency domain is given by

$$W(\omega) = \frac{S_x(\omega)}{S_x(\omega) + S_b(\omega)}$$

which is referred to as the Wiener filter [NRC, 1989]. Another solution to this problem in the time domain is referred to as a Kalman filter [Rhodes, 1971]. When the signals $x[n]$ and $b[n]$ meet the conditions under which the Wiener filter is derived, i.e., uncorrelated and stationary, then the Wiener filter can provide significant enhancement without severe object or background distortion. Effectively, this estimate is a Linear Least-Squares Estimate (LLSE) of the speech. To create a Wiener filter, the spectrum of the speech and noise must be known, measured, or at least estimated. The required spectral densities can be estimated by averaging over multiple frames when sample functions of $x[n]$ and $b[n]$ are provided. Wiener filtering is optimal in the mean-square error sense if the speech and noise are stationary. Typically, however, the signal components are nonstationary. When the object, for example, is on order of a few milliseconds in duration, its spectrum is difficult to measure, requiring an average to be essentially instantaneous.

A Kalman filter is very similar to a Wiener filter in that it also is designed to yield a Linear Least-Squares Estimate of the spoken speech. The main difference between the Wiener and Kalman filtering is that the Kalman filter relies on a measurement error which is updated and used to obtain the desired signal estimate. Unlike a Wiener filter, it does this estimation in the time domain.

2.3.3 Baseline Algorithm Description

As a starting point, an algorithm previously designed by the speech and information technology group at Lincoln Laboratory was used as a baseline for work involving this thesis [Quatieri and Dunn, 2002]. This algorithm implemented an acoustic-based Wiener filter. It divided the acoustic speech input into overlapping segments of 12 ms duration and conducted processing on each segment separately. Each segment was windowed with a triangular window and offset by 2 ms from the previous segment. The two main components of this processing are speech activity detection and filtering.

Speech Activity Detection

Speech activity detection was used to improve adaptability and was used by the filtering stage of the processing on each speech segment. The acoustic based speech detector was a basic energy based detector. A block diagram of the detector is shown below in figure 2.1. The detector operated by taking the input acoustic speech $x(t)$ and computing its short-time Fourier Transform (STFT), yielding a frequency domain representation of the speech over each 12 ms window. Then it computed the energy of each frequency component and summed over all frequencies, thus obtaining a measure of the energy present in the acoustic signal for each 12 ms segment of speech. Finally, this energy was compared to a threshold ß. If the energy of the segment is greater than ß, it is declared as containing speech. Otherwise, the segment is declared as containing only noise.

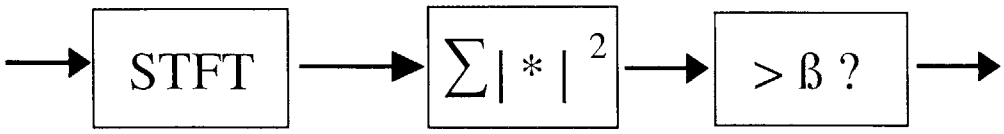


Figure 2.1: Block Diagram of the Speech Detector

In moderate noises this acoustic based speech activity detector was stable, not erroneously flipping on and off in a contiguous speech or background region, and accurate (when compared to the actual regions were speech was present). However, when a large amount of noise is present (like that experienced in the cockpit of a jet fighter, an Abrams tank, or even a very crowded room), the original Lincoln Laboratory speech detector did not perform well: it was very unstable, flipping on and off constantly in a contiguous speech or background region, and inaccurate (when compared side by side to the actual speech waveform). These effects are important in themselves because more accurate detection allows a greater flexibility to attenuate noise. For example, over a large segment of no speech, the entire segment can be drastically attenuated thus completely suppressing the noise -- or at least suppressing it much more than would be able if speech detection were not possible. More generally however, the speech filtering stage uses the detection results to attenuate speech spectral bands dominated by noise. The baseline algorithm accomplished this by means of a Wiener filter.

Wiener Filtering

Wiener filtering is used to enhance speech segments and requires the use of the detection described above. In this baseline algorithm, the noisy speech signal $y[n]$ is short-time processed at frame interval L samples and an estimate of the Wiener filter on frame $k-1$, is denoted by $W(k-1,\omega)$. The background spectral density, $S_b(\omega)$, is estimated by averaging spectra over a given background region. This averaging was one aspect that required the use of speech detection (see above). Assuming that the clean speech signal $x[n]$ is nonstationary, one approach to obtaining an estimate of its time-varying spectral density on the k^{th} frame uses the Wiener filter $W(k-1,\omega)$ to enhance the current frame. This operation yields an enhanced clean speech spectral estimate $\hat{X}(k,\omega) = W(k-1,\omega)Y(k,\omega)$ which can then be used to update the Wiener filter for the next frame:

$$W(k,\omega) = \frac{|\hat{X}(k,\omega)|^2}{|\hat{X}(k,\omega)|^2 + \alpha S_b(\omega)} \qquad (1)$$

where the parameter α has been added to control the degree of suppression.

To slow down the rapid frame-to-frame movement of the object spectral density estimate, temporal smoothing is applied to the spectrum in equation (1), thus modifying the Wiener filter $W(k,\omega)$ (the final modified Wiener filter is given in equation 3 below). To avoid blurring rapidly-varying and short-duration sounds with this temporal smoothing, the time constant that controls the smoothing was made a time-varying parameter. In particular, this time constant is selected to reflect the degree of stationarity of the waveform whereby when the spectrum is changing rapidly, little temporal smoothing is introduced because a near-instantaneous spectrum is required for the Wiener filter. On the other hand, when the spectrum is stationary, as in regions without speech or in steady object regions, an increased smoothing of the spectrum improves the spectral estimate for the Wiener filter. Although this filter adaptation results in relatively more noise in non-stationary regions, there is evidence that, perceptually, noise is masked by rapid spectral changes and accentuated in otherwise stationary regions [Knagenhjelm and Klein, 1995] [Moore, 1988].

The degree of stationarity is obtained through a spectral derivative defined for each frame as the mean-squared difference between two consecutive short-time spectral measurements of $y[n]$. This spectral derivative is used to control a time-varying time constant that is denoted by $\tau(k)$. The resulting smooth object spectrum is given by

$$\hat{S}_x(k,\omega) = \tau(k)\hat{S}_x(k-1,\omega) + [1-\tau(k)] \mid \hat{X}(k,\omega) \mid^2 \quad (2)$$

which is then used to make the final Wiener filter given by

$$W(k,\omega) = \frac{\hat{S}_x(k,\omega)}{\hat{S}_x(k,\omega) + \alpha S_b(\omega)} \quad (3)$$

The use of spectral change in the Wiener filter adaptation, as well as a number of refinements to the adaptivity, including iterative re-filtering and background adaptation, helps avoid blurring of temporal fine structure in rapidly-varying and short-duration signals and improves performance with a changing background [Quatieri and Baxter, 1997]. The baseline noise-supression algorithm is illustrated in figure 2.2.

The inclusion of background adaptation requires that we perform speech activity detection to determine which frames in a signal contain speech and background and which frames contain background only. This was one reason that detection is so important. Also, with the use of speech activity detection, we can further improve adaptivity by providing distinct Wiener filters, one during background and one during object. An object spectral estimate during background regions is derived, as well as a distinct object spectral estimate during object regions, thus alleviating the need to re-adapt across object/background boundaries. For each state, speech and background, we have a distinct Wiener filter, given respectively by

$$W_s(k,\omega) = \frac{\hat{S}_x^s(k,\omega)}{\hat{S}_x^s + \alpha S_b(k,\omega)}, \quad W_b(k,\omega) = \frac{\hat{S}_x^b(k,\omega)}{\hat{S}_x^b + \alpha S_b(k,\omega)}$$

where $\hat{S}_x^s(k,\omega)$ and $\hat{S}_x^b(k,\omega)$ are object estimates during the speech state and the background state, and $S_b(k,\omega)$ is a continuously changing background spectral density

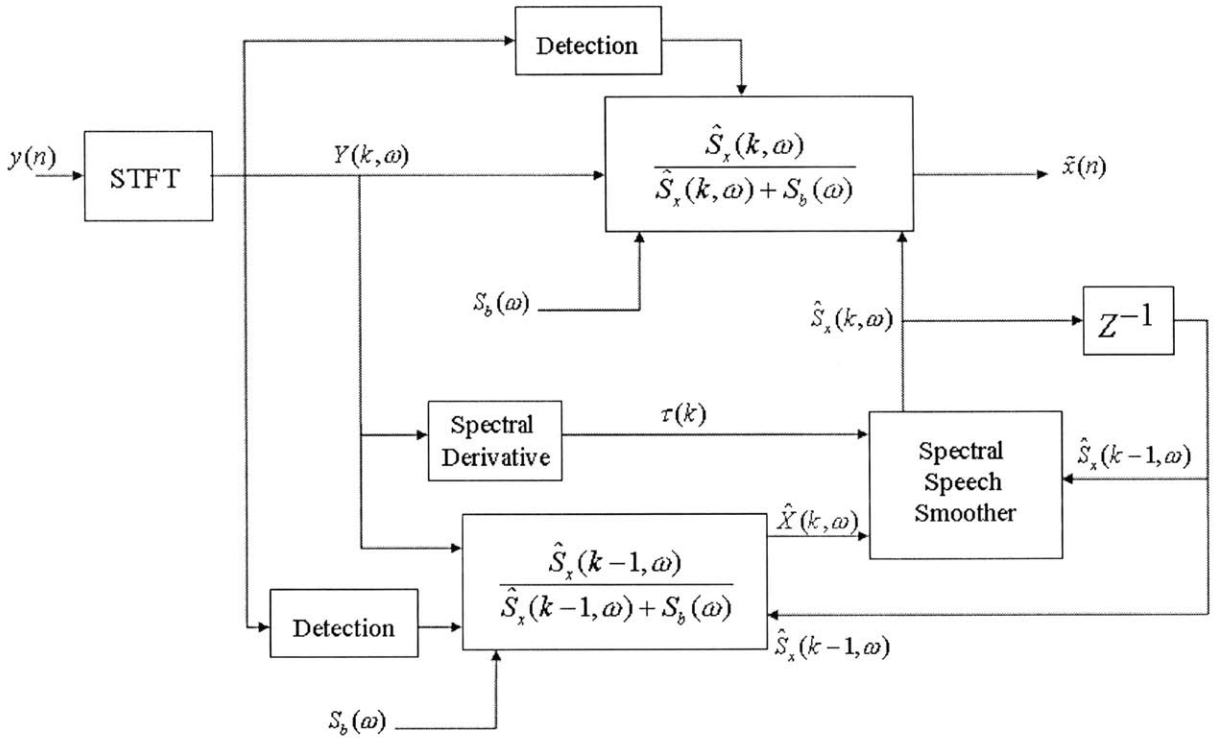


Figure 2.2: Noise reduction algorithm based on spectral change

estimate. Time constants used in spectral smoothing to obtain object estimates are a function of spectral change as in Equation (2). These refinements to the suppression algorithm provide better control over the background level, and improved fidelity of the object and background. Details of the refinements can be found in [Quatieri and Baxter, 1997] and [Quatieri and Dunn, 2002].

2.4 Sensor descriptions

In this thesis, several sensors were examined for use in noise suppression and detection. The goal of this was to find how they could be collectively used to aid speech enhancement and speech coding. This section describes the several sensors used in experiments and available in the collected corpora.

2.4.1 GEMS sensor

One of the non-air-acoustic sensors of interest in this thesis is the general electromagnetic motion sensor. The general electromagnetic motion sensor (GEMS) measures tissue movement during voiced speech, i.e., speech involving vocal chord vibrations [Burnett et al, 1999]. An antenna is typically strapped or taped on the throat at the laryngeal notch, but also can be attached at other facial locations. This sensor emits an electromagnetic signal that penetrates the skin and reflects off the speech production anatomy such as the tracheal wall, the vocal folds, or the vocal tract wall. Because signals collected from a GEMS device depend on the tissue movement in the speech production anatomy, it is relatively immune to degradation from external acoustic noise sources.

During voiced speech, GEMS records quasi-periodic electromagnetic signals due to vibration of the speech production anatomy. When placed at the larynx, quasi-periodic measurements are found during vowels, nasals, and voiced consonants including prior to the burst in voiced plosives, i.e., during voice bars. Single pulses have also been observed very sporadically from the GEMS.

2.4.2 PMIC sensor

The physiological microphone (P-Mic) sensor was the second additional sensor we examined and used. It is composed of a gel -filled chamber and a piezoelectric sensor behind the chamber [Scanlon, 1998]. The liquid-filled chamber is put in contact with the skin of a person and adapts to conform to the potentially very non-planar skin surface. This allows the PMIC to be flush against the skin, in maximum contact. Like the GEMS sensor, the P-Mic can be strapped or taped on various facial locations. The P-Mic at the throat measures primarily vocal fold vibrations with quasi-periodic measurements similar to that of GEMS. The P-Mic signal at the throat, however, contains some low-pass vocal tract formants with bandwidths wider than normal. Other facial locations can provide additional vocal tract characteristics. The P-mic located on the forehead, for example, gives significant vocal tract information but is far less noise-immune than the P-Mic at the throat in severe environments

The liquid filled chamber is designed to have poor coupling between ambient background noise and the fluid-filled pad thus attenuating vibrations of unwanted ambient background noise. Vibrations that permeate the liquid-filled chamber are measured by the piezoelectric sensor that provides an output signal in response to applied

forces that are generated by movement, converting vibrations traveling through the liquid-filled chamber into electrical signals. Since the liquid-filled chamber can be designed to greatly attenuate acoustic noise, the piezoelectric accelerometer can produce a signal that has much improved SNR over traditional acoustic based microphones such as the B&K microphone (see below). In this case and in the rest of this thesis SNR is defined as the ratio of the average energy of the speech waveform during speech divided by the average energy of the background noisy during speech and non-speech.

2.4.3 EGG sensor

In the corpus collections, two electroglottograph (EGG) [Rothenberg, 1992] sensor electrodes were taped on the left and right of the GEMS sensor. The EGG sensor measures vocal fold vibrations by providing an electrical potential (of about one volt rms and two-to-three megahertz) across the throat at the level of the larynx. With a pair of gold-plated electrodes, the sensor measures the change of impedance over time. When the vocal folds are closed, the impedance is decreased; when they are open, the impedance is increased. Thus the opening and closing of the vocal folds, as present in voiced speech is measured well by the EGG sensor. The EGG sensor, like the GEMS sensor however cannot detect or measure unvoiced speech reliably.

2.4.4 Resident Microphones

There were several resident microphones that were used in the recordings that were used by our algorithm. These microphones are the exact same ones that are used by personal in the environments we simulated (see Corpus description below). For example, one resident microphone was the stinger used by Blackhawk helicopter pilots. This was used since one of the noise environments were experimented with was a Blackhawk helicopter (see Corpus description below). All of the resident microphones were noise-canceling microphones - ie they were designed to suppress some of the noise present in the speech that they measured. Consequently, the noise level in the resident microphone recording reduced and the speech sounded more intelligible. However, to accomplish this noise reduction, several of the resident microphones destroyed some of the fine temporal features that were of interest to us for speech processing and speaker identification.

2.4.5 B&K microphone

The B&K microphone was a standard acoustic microphone that had no noise suppression attributes. Thus it was very sensitive to acoustic noises. This made it fairly worthless to use for speech processing in harsh environments. However, it was

extensively used in testing since it provided a very poor and noisy signal that could be used to test algorithm performance under extreme noise conditions.

2.5 Corpus and Environment Descriptions

There were two corpora used by tests in this thesis work. The first was one collected by Lawrence Livermore National Laboratory. The second was collected by the Arcon Corporation. This section describes both.

2.5.1 Lawrence Livermore National Laboratory Corpus

The Lawrence Livermore National Laboratory Corpus included 8 speakers speaking several sentences. Every sentence was recorded using an acoustic microphone and a GEMS sensor (both recorded simultaneously). Each sentence was spoken without noise so noise had to be artificially added later for speech enhancement tests. The noise added was made to mimic car engine noise. It was created from white noise that was passed through a filter designed after the transfer function of a car engine. The Lawrence Livermore National Laboratory corpus was the first corpus available to us. Consequently many of our initial experiments used it.

2.5.2 Advanced Speech Encoding Corpus

The DARPA Advanced Speech Encoding (ASE) corpus was much more extensive and realistic corpus than the Lawrence Livermore corpus. It was more realistic than the Lawrence Livermore corpus because recordings were made of the speaker in a sound room while noise from an real-life environment was being played. Consequently, noise and the affects of noise were present in all the recordings just like they would be when the sensors are used in a real-life situation and environment.

The ASE corpus was collected from ten male and ten female talkers. However, initially only one male speaker was available. Consequently most of the tests performed in this thesis use spoken phrases from this speaker. The rest of the speaker data became available in the final month of this thesis, but due to time constraints experiments using them were limited. Future work will focus on rerunning several experiments with these other speakers. Scripted phonetic, word and sentence material along with conversational material were generated by each talker. These materials were generated in nine different acoustic noise environments. The corpus was collected in two sessions (on two different days). Speakers were exposed to a variety of noise environments including both benign and severe cases. Six of the environments represented three acoustic environments with each presented at two intensity states. The presentation levels for these states differed by 40 dB SPL. Specific environments are quiet, office (56 dB), MCE (mobile command enclosure, 79 dB), M2 Bradley Fighting Vehicle (74 dB and 114 dB), MOUT (military operations in urban terrain, 73 dB and 113 dB), and a Blackhawk helicopter (70 dB and 110 dB). We call these environments (with L indicating low noise and H indicating high noise) quiet, office, MCE, M2L, M2H, MOUTL, MOUTH, BHL and BHH,

respectively. The M2H is the noisiest environment. Consequently, the M2H environment was used in all experiments involving the ASE corpus in this thesis.

For each talker and environment, combination time-synchronous data was collected from up to seven separate sensors. These sensors consisted of the previously introduced GEMS, P-Mic and EGG. Data was also collected from two acoustic microphones, a high quality B&K calibration microphone and an environment specific "resident" microphone. The resident microphone was typically the first-order gradient noise-cancellation microphone used for normal communications in that specific environment.

One GEMS and one EGG were located near the talker's larynx. Careful attention was given to tuning the GEMS sensor and in optimizing its placement. The GEMS was considered the primary sensor during the corpus collection. A specific talker's neck and shoulder geometry often required that tradeoffs be made in the placement of the secondary sensors in order to optimize the GEMS signal. Two P-Mics were used, one located in the vicinity of the talker's larynx and the other on the talker's forehead. Due to the acoustic presentation levels of some of the noise environments, all talkers used the acoustic protection systems typical of each specific noise environment. This normally consisted of some type of communication headset that provided noise attenuation on the order of 20 dB. Human subject procedures were followed carefully and noise exposure was monitored.

The complete corpus consists of from eight to nine channels of data from approximately twenty minutes of speech material in each of nine acoustic noise environments from each of the twenty talkers. All sensor data was sampled at 48kHz, though the non-acoustic data was downsampled to 16kHz for space considerations. The full corpus takes approximately 70 GB of storage.

2.6 Summary

In this chapter we described much of the framework that the rest of this thesis uses as a starting point. We discussed various noises that are important to consider, aspects of speech, the corpora we used to conduct experiments on, the sensors used in these corpora, and current algorithms that have been used for speech enhancement. Specifically, a baseline speech enhancement algorithm based on a Wiener filter was introduced. This baseline system will be a focus of chapters 3-5 and part of 6. Chapter 3 will focus on the detection stage of this algorithm. Chapters 4 and 5 will focus on the spectral enhancement portion of the algorithm.

Chapter 3: Speech Activity Detection

Speech detection is used to identify which segments of an audio waveform contain speech (with noise) and which contain only noise. It is useful because it allows one to process the different segments differently. Speech detection allows us to make the distinction between such regions and thus allows more intelligent processing. For example, in the baseline system (see chapter 2), it contributes to further adaptivity by making it possible to create distinct Wiener filters, one during background and one during speech.

This chapter discusses the issues of detection relevant to this thesis. It begins with discussing traditional implementations and limitations of detection. Then it examines the use of multiple sensors and how they can aid in detection. Next, it discusses the multi-sensor speech class detector used in this thesis for detection. It then concludes with applications of this detection scheme.

3.1 Traditional Implementation and Problems

Traditionally speech detection is performed using the audio waveform that is to be processed and enhanced via noise suppression. There are numerous speech detectors that have been developed [Sohn, Kim, Sung, 1999] [Haigh and Mason, 1993]. A common one is a speech activity detector that computes the signal energy over a segment, and decides that speech is present if the energy is above some threshold. Another variation of such a detector performs a frequency band-energy computation and then makes decisions whether speech is present over each band. This later detector was the one used in the baseline system described in section 2.

Traditionally such detectors perform adequately in moderate noise levels: the detection decision is fairly stable (here stability refers to how regularly the detection decision flips on and off in a contiguous speech or background region) and accurate (when compared to the actual regions were speech is present). However, when a large amount of noise is present (as with cockpit noise in an airplane or a very crowded room), these detectors do not perform well: they are unstable and inaccurate. The reason for this is that when the noise energy becomes large enough, it dominates the speech energy and thus unfavorably dominates the energy calculations that are used to make speech activity decisions. These effects are important because more accurate detection allows for more effective enhancement of segments of noise and thus less speech distortion

3.2 Multi-Sensor Approach

Our approach to circumventing the speech detection problem caused by noise in speech waveforms is to use the waveforms from other sensors that are more immune to acoustic noise (see chapter 2 on sensors). In particular, the GEMS sensor was shown to be reliable at detecting voiced speech; yet it was poor at detecting unvoiced speech (see Appendix A). This finding was further supported by the theory of how the GEMS sensor operates (see chapter 2): the EM waves reflecting off the vocal folds and/or tracheal wall

reliably record voiced sounds and only sporadically detect unvoiced sounds. Despite the poor unvoiced speech detection performance, the GEMS detector was shown to improve the MSE and SSNR of the processed speech (see Appendix A). This improved performance motivated further experiments in detection. The next sensor examined was the PMIC sensor. It was also shown to improve MSE and SSNR of processed speech, however it is less accurate at detecting voiced speech since it is not entirely immune to acoustic noise. Unlike the GEMS sensor, the PMIC sensor is capable of also detecting unvoiced speech (see chapter 2 and Appendix A). Consequently, one solution to the detection problem was essentially to use the accurate voiced speech detection available via the GEMS sensor and then, given this voiced speech detection decision, use the PMIC sensor waveform to create a decision on where the unvoiced speech regions occurred. This was done for one detection scheme described in Appendix A.

The final detection scheme used in this work however used non-acoustic sensors in conjunction with the noise canceling resident microphones from the ASE corpus to create a speech class detector. A speech class detector is more advantageous than the other detection schemes described above because it allows further adaptability by allowing the creation of multiple Wiener filters tailored to each speech class.

3.3 Speech Class Detection

3.3.1 Motivation

After conducting detection experiments with the PMIC decision fused with the GEMS decision, the resident microphone in the harsh ASE corpus environments was once again examined for use in fused detection. A closer examination of the resident microphone data was inspired by several factors. Firstly, unlike the PMIC and the other sensors that tend to roll off before 4kHz, the resident microphone contains much energy above 4kHz. Thus phones such as /s/, /ʃ/, /tʃ/, and /θ/ that contain large amounts of energy often centered at a high frequency above 4kHz might be better detected by using the resident microphone. Consequently the resident microphone data sampled at 16kHz was examined with the hope that the upper bands of the16kHz resident microphone signal would be able to help detection when fused with previous GEMS-based detectors. All the other sensors were sampled at 8kHz because the speech encoder used in other work conducted at Lincoln optimally operates on 8kHz signals.

Another reason for examining the resident microphone is that it has very poor SNR below 500 Hz and thus is not able to accurately detect very low frequency voice bars (voicing which precedes a voiced consonant) and some nasals; however it is able to detect most voiced speech and all unvoiced speech. An example is shown in figure 3.1. In the spectrogram in this figure and all following spectrograms, an analysis window of 25-ms and a frame length of 10-ms is used. In figure 3.1, the GEMS signal clearly gives the presence of the low-frequency nasal /n/ and voice bar in the voiced plosive /d/ in the word "dint". The resident-mic, while not revealing the nasal and voice bar, more clearly shows the high-frequency burst energy in the /d/ and in the unvoiced plosive /t/. Thus, it was thought that some form of the resident microphone detection decision could be used in conjunction with the GEMS and PMIC detection decisions to create some form of

speech class detection. As the experiments below demonstrate, this was indeed what was accomplished.

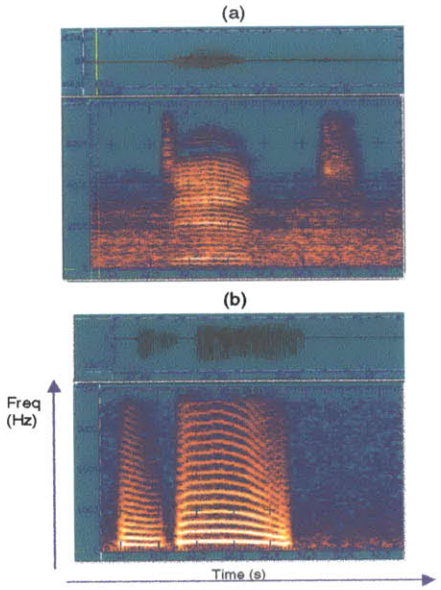


Figure 3.1: Waveforms (from the M2H ASE corpus environment) and spectrograms of the (a) resident-mic signal and (b) GEMS signal for the word "dint". The GEMS signal shows the presence of the nasal /n/ and voice bar in the initial voiced plosive /d/, while the resident-mic shows the high-frequency burst energy in the /d/ and in the unvoiced plosive /t/.

3.3.2 Classes of Speech

The class-based detection that was created and explained further below was able to separate a recording into 4 classes of speech: voiced, unvoiced, low-band, and background. The background class contains all segments of a recording that contain no spoken speech. Thus they are only noise and can be attenuated.

The unvoiced class contains all speech sounds that have no vocal fold vibrations. Such sounds include many consonants such as /f/, /t/, /p/, /k/, /s/, /tʃ/, and /ʃ/ (see chapter 2). These sounds tend to be more high frequency in nature and not harmonic. An example of an unvoiced sound is the /ʃ/ in the "sh" (centered at about 5kHz) at the beginning of the word "shag" shown in figure 3.2 below.

The low-band class contains only low frequency sounds that have little high frequency content. This includes low frequency nasals such as /n/ and /m/ and voice bars, voicing which precedes a voiced consonant such as a /b/, /g/, /d/, /z/, or /v/. These voice bars are low frequency in content and are one of several cues used for proper identification of consonants by humans and thus are very important for intelligibility. An example of a voice bar is the voicing preceding the /b/ in "boot" shown in figure 3.3. The voiced class of speech contains all segments of speech that involve use of vocal fold vibrations, except for voice bar regions. Thus the voiced class includes both voiced vowels, voiced consonants (including voiced fricatives and voiced plosives) such as a /z/ or /v/ or /b/ (but not the voice bar preceding a /b/), and liquids and glides such as /w/, /y/, /r/, and /l/. An example of a voiced segment of speech is the /z/, /e/, or,/d/ in the word "zed" which is shown in figure 3.4. As one may note from the figure, the voiced class of speech contains sounds that contain both high frequency and low frequency components.

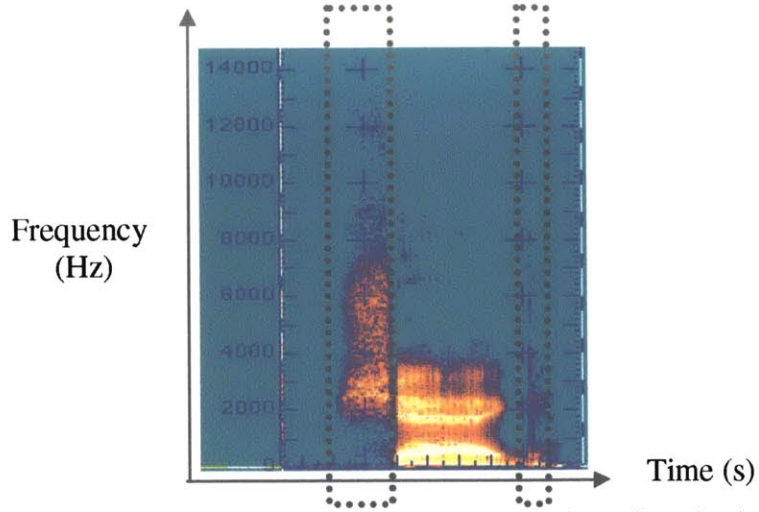


Figure 3.2: Spectrogram Example of Unvoiced Regions in the Word "Shag"

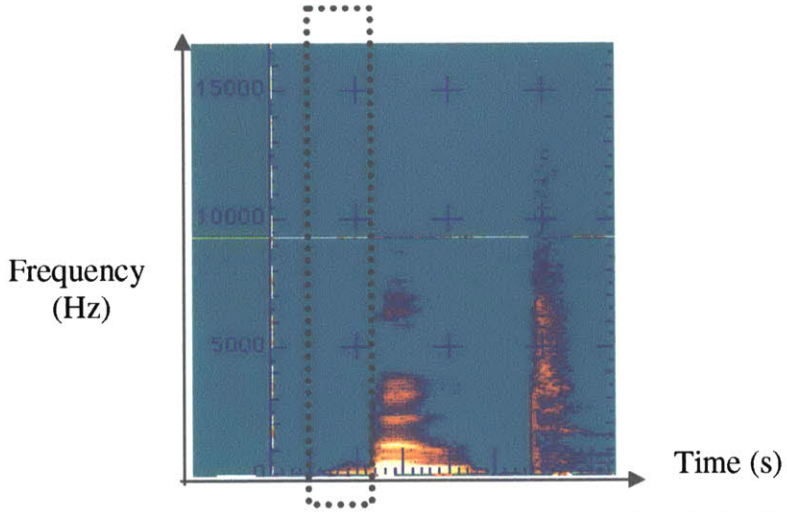


Figure 3.3: Spectrogram Example of Low-Band Region: "b" Voice Bar in "Boot"

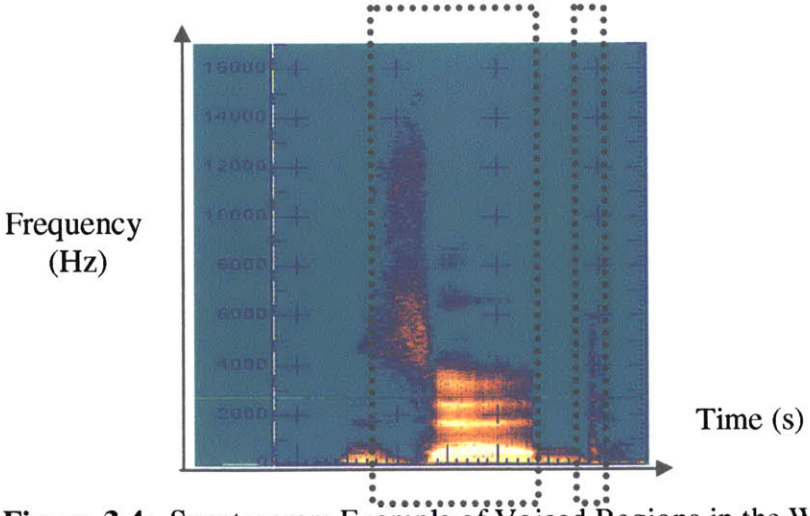


Figure 3.4: Spectrogram Example of Voiced Regions in the Word "Zed"

The differences in all four classes are summarized in figure 3.5. As one can see, voice bar regions contain only very low frequency speech content, unvoiced regions contain only high frequency speech content, voiced regions contain both high and low frequency speech content, and background regions contain no speech content.

Speech Class	Contains High Frequency Speech Content?	Contains Low Frequency Speech Content?
Low-Band	No	Yes
Unvoiced	Yes	No
Voiced	Yes	Yes
Background	No	No

Figure 3.5: Differences in 5 speech classes (high and low frequency speech content)

3.3.3 Detection of Classes

The detection and distinction of each of the 4 classes is accomplished by using two detectors, one which uses the GEMS waveform and one which uses the resident microphone waveform, and then fusing the detection decisions.

The resident microphone-based detector uses a high pass version of the 16kHz resident microphone data. It is high passed at a cutoff frequency of 3kHz because the noise present in the resident microphone waveform starts to roll off at about 3kHz while the speech does not roll off. Since several of the formants of voiced speech are above 3kHz, this high-passed resident microphone-based detector is able to detect voiced speech in addition to high frequency unvoiced speech yet is not able to detect voice bars. The GEMS based detector is able to detect voiced speech and the low-band speech such as voice bars and nasals. In order to reduce misdetections, the resident microphone decision is used only 200-ms on either side of a region detected by the GEMS based detector, much like the PMIC and GEMS fused detector decision (see Appendix A). The high SNR of the resident microphones above 3kHz allowed the 200-ms increase (from the 100-ms used in the detectors of Appendix A) because this high SNR makes the decisions more accurate. This is also beneficial because some consonants are more than 100-ms away from voicing (for example the /t/ in the word "boost"). A summary of the GEMS and high-passed resident microphone detectors is shown in figure 3.6.

	Detects Voice Bars?	Detects Voiced Speech?	Detects Unvoiced Speech?
GEMS Based Detector	Yes	Yes	No
High Pass Resident Mic Based Detector	No	Yes	Yes

Figure 3.6: GEMS and HP Resident Microphone Detection Ability

Based on the information presented in figure 3.6, class based decisions are made by using the detector decisions of the GEMS and high pass resident microphone based detectors. Regions that the GEMS detected as containing speech and the high pass resident microphone did not are labeled as a voice bar region. Regions that both detected as containing speech are labeled as a voiced speech region. Regions that the high pass resident microphone detected as containing speech but the GEMS did not are labeled as unvoiced speech regions. Regions that neither detector identified as containing speech are labeled background noise regions.

3.3.4 Data

Since Segmental SNRs and MSE measurements were once again not possible to compute with the ASE corpus (since no clean speech existed for comparison), we were forced to examine individual detection decisions of the GEMS and high-pass resident microphone detectors. A total of 61 different utterances were examined for accuracy. An example is the word "dint" which is shown below in figure 3.7. As one can see from the figure, the resulting detections are stable (not flipping on and off during contiguous segments of speech or background) and accurate (not mislabeling speech as background or background as speech) over the high band and low band of each sensor. The low band of the GEMS detects the voice bar preceding the /d/ in "dint" while the resident microphone does not at all. The resident microphone, on the other hand, detects the unvoiced /t/ while the GEMS does not. These observations allowed the creation of the class based detection described above.

3.4 Conclusion

In this chapter, we discussed the development of a speech class-based detector based on the GEMS sensor and a high-pass representation of the resident microphone sensor. This speech-class detector is able to distinguish voiced, unvoiced, low-band, and background speech regions. Classification of these different speech regions allows greater flexibility and adaptability in processing, and is essential to several of the improvements made to the estimation of the speech signal discussed in chapters 5 and 6. Detection of different speech classes may also have several other applications. It may allow improved speech encoding and speaker identification. For instance, a speaker identification system may perform better if it conducts speaker recognition over speech classes and then fuses the results. If a speech encoder has knowledge of speech classes, it may be able to allocate more bits for transmission of plosives and other consonants that are essential for intelligibility while allocating fewer bits for longer vowels or silences that are not as abrupt as the short plosives.

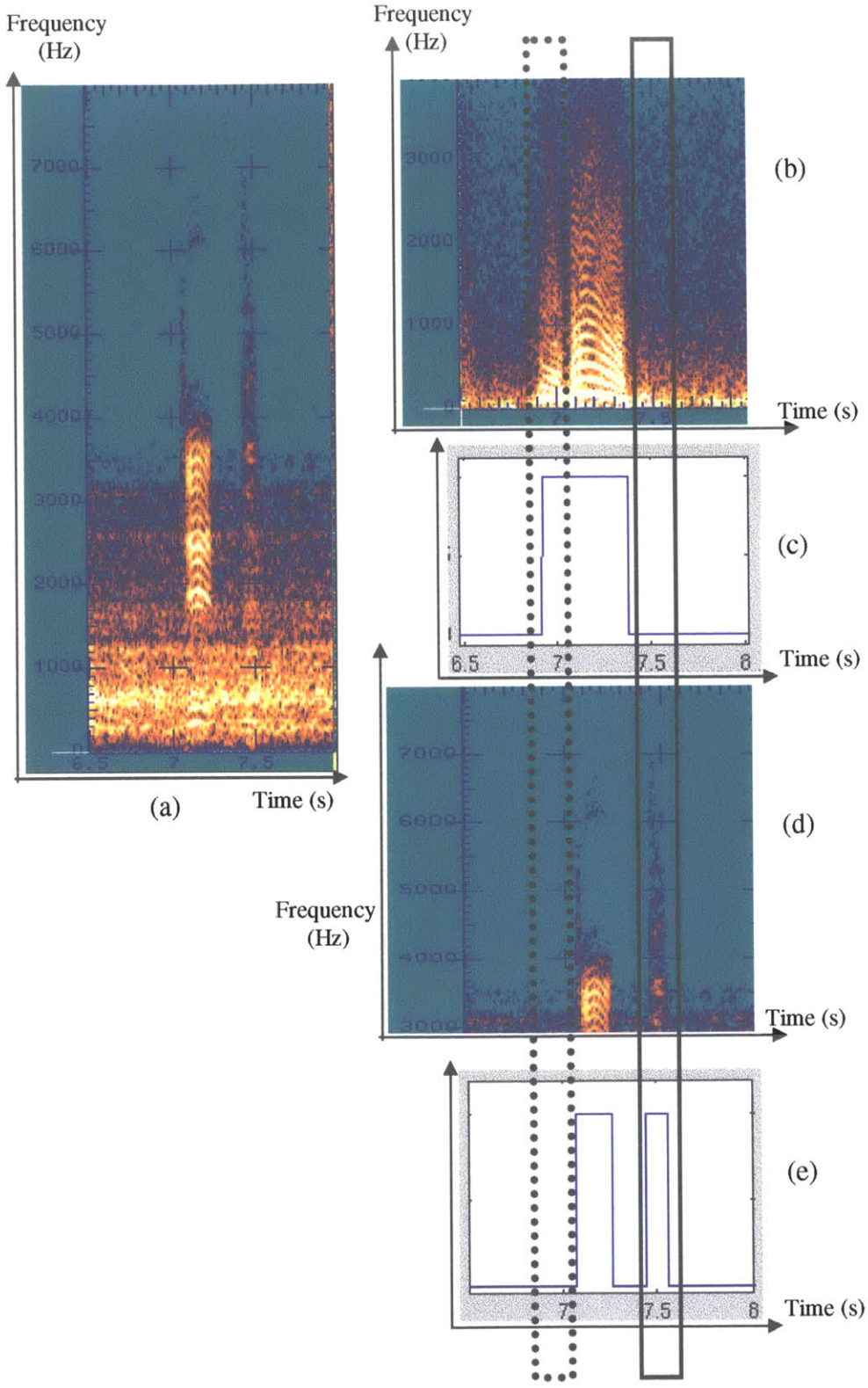


Figure 3.7: Example of two band detection (used for class detection) on the word "dint": (a) full band resident microphone signal, (b) GEMS signal, (c) GEMS-based detection, (d) upper band of resident microphone signal, and (e) resident microphone-based detection. The GEMS signal shows the presence of the nasal /n/ and voice bar in the initial voiced plosive /d/ (as in the case of figure 3.1), while the resident microphone shows the high-frequency burst energy in the /d/ and in the unvoiced plosive /t/. The GEMS detector detects the voice bar (see dotted box) while the resident microphone detector does not. The resident microphone detector detects the /t/ (see solid black box) while the GEMS detector does not. Note that the GEMS also detects the low-band energy in the /n/ while the resident microphone does not.

Chapter 4: Performance Bounds of Wiener Filtering

4.1 Background

This chapter explores the performance limitations of Wiener filtering under ideal spectral measurement conditions. The main objective is to test the limitations of Wiener filtering in order to (1) find a performance bound that can not be improved upon using any form of Wiener filtering, and (2) explore the distance of this bound from the baseline system, thus revealing how much potential there is for improvement. This chapter begins by reviewing the baseline enhancement algorithm and its performance. Then the chapter explores the performance of an "ideal" Wiener filter formed from the spectrum of clean speech. Finally the affects of phase in filtering for speech enhancement are described.

This chapter uses the Lawrence Livermore corpus because it allows control of the noise present in a signal with noise added manually, and because the Lawerence Livermore corpus has the clean speech signal available in it for comparisons. Testing involved the use of 12 sentences spoken by 2 male speakers. All the examples in this chapter are from the same male speaker speaking the sentence "Young people participate in athletic activities."

4.2 Baseline System

As described in chapter 2, the baseline system uses a Wiener filter based on the statistics of the noisy speech waveform. Specifically the Wiener filter transfer function is obtained by the equation:

$$W(k,\omega) = \frac{\hat{S}_x(k,\omega)}{\hat{S}_x(k,\omega) + \alpha S_b(\omega)}$$

Here, $S_b(\omega)$ is the average of the noise, which is estimated by using the statistics of the regions of a recording that are identified by the speech activity detector as containing only background noise and no speech (see chapter 3). $\hat{S}_x(k,\omega)$ is the estimate of the speech spectrum of the clean speech that is determined by:

$$\hat{S}_x(k,\omega) = \tau(k)\hat{S}_x(k-1,\omega) + [1-\tau(k)] \mid \hat{X}(k,\omega) \mid^2$$

$$\text{and} \quad \hat{X}(k,\omega) = W(k-1,\omega)Y(k,\omega)$$

where $W(k-1,\omega)$ is the Wiener filter from the previous segment of speech. $Y(k,\omega)$ is the STFT of the acoustic input for the current segment of speech.

The output of the baseline system is obtained by breaking the output waveform into two parts, the magnitude and phase of its frequency domain representation. The magnitude estimate of the speech was obtained from the Wiener filter as:

$$\tilde{X}(k,\omega) = W(k,\omega)Y(k,\omega)$$

where $Y(k,\omega)$ is the spectrum of the speech input. The phase of the output used is simply that of the input-waveform:

$$\angle\tilde{X}(\omega) = Y(k,\omega)$$

The phase of the input signal was originally used because of the folklore that phase is unimportant for perception, a folklore reinforced by Wang and Lim's work [Wang and Lim, 1982] which found that the spectral phase of speech is not important for quality at high SNRs, "quality" being distinctly different than "intelligibility." Consequently, traditional suppression algorithm development has not paid much attention to the difficult problem of creating a phase estimate based on acoustic noisy speech signals. Wiener filtering, for example, does not modify the phase of the original input. This thesis, however, focuses on fairly severe SNRs, and shows that the spectral phase actually plays an important role in speech enhancement. This finding is consistent with observations by Vary that phase has importance in noise reduction for signal-to-noise ratios less than 3 db [Vary, 1985].

The baseline algorithm takes severely noisy signals (for example, the negative 20 db SNR signals shown in figure 4.1) as input and produces enhanced signals shown in figure 4.2. These signals contain high noise residuals when the input has a low SNR and still do not quite resemble the clean speech shown in figure 4.3. They have a large amount of noise that completely masks the speech below 1000 Hz, much of the harmonic structure of the speech has been smeared, and when listening to the estimated speech, it is very hard to make out what is being said. However, one can see that the enhanced signals in figure 4.2 more closely resemble those of the clean speech of figure 4.3 than the original inputs in figure 4.1.

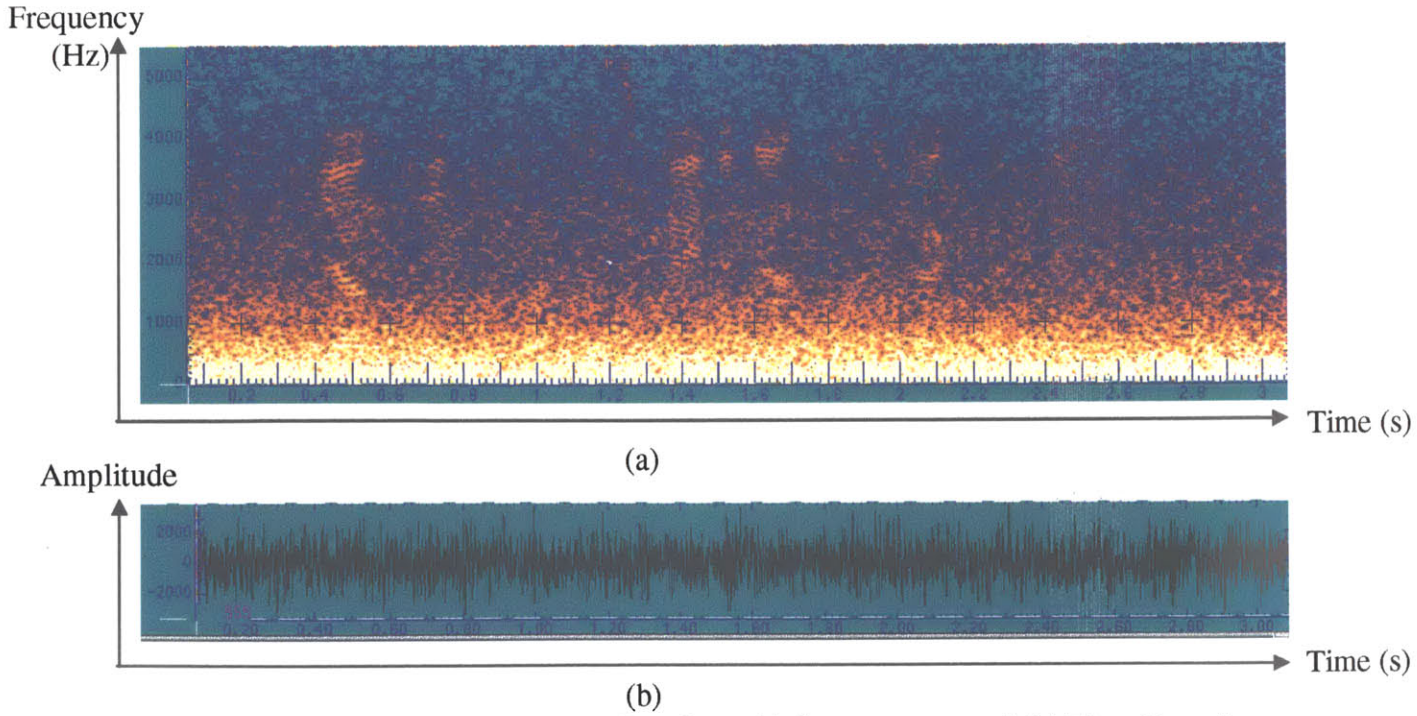


Figure 4.1: Noisy –20 db Input Waveform (a) Spectrogram, and (b) Time Domain

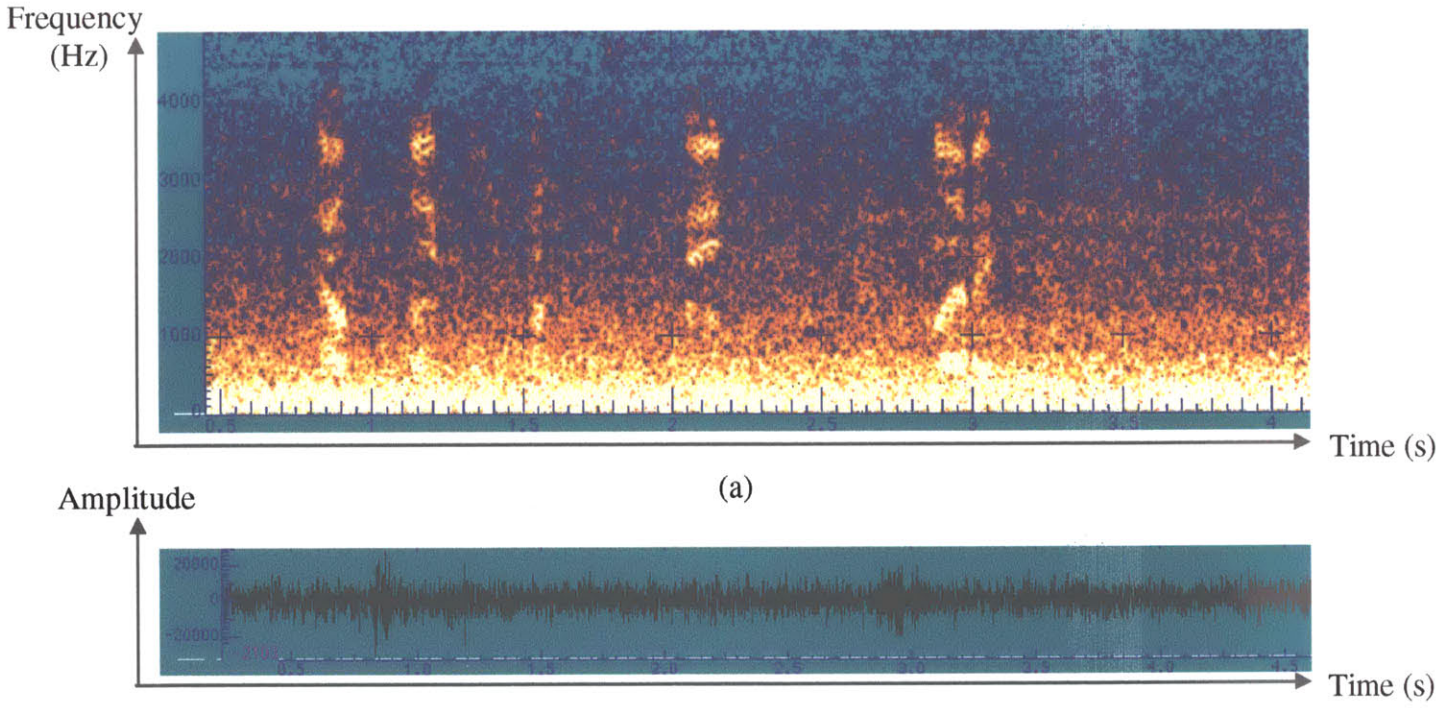


Figure 4.2: Result of Baseline Algorithm Processing (a) Spectrogram and (b) Time Domain

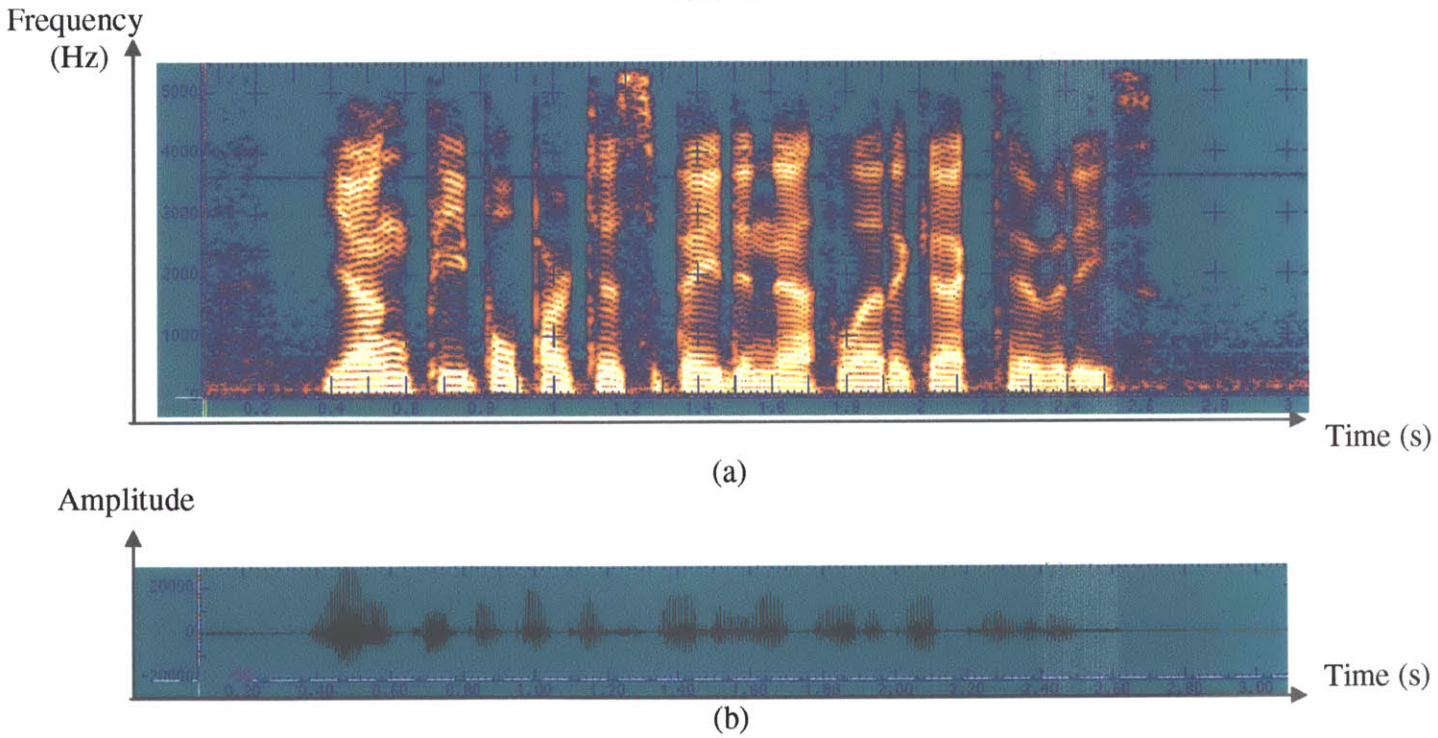


Figure 4.3: Clean Speech (a) Spectrogram and (b) Time Domain

Although the baseline system provides some enhancement to the noisy input signals, it was not known how much this enhancement could be improved upon by using

the new non-acoustic sensors at our disposal. Consequently, several tests were conducted.

4.3 Wiener Filter Magnitude Performance

In order to explore the performance bounds of Wiener filtering and gain insight into how non-acoustic sensors could be used, two tests were conducted. The first test focused on the performance of the "ideal" Wiener filter, a filter based on the actual clean speech statistics. The second test examined what the filtering should produce if it was able to perfectly reconstruct the magnitude of the signal. This was done using the "ideal" magnitude of the speech.

4.3.1 Ideal Wiener Filter

The "ideal" Wiener filter was created from the clean speech from the Lawrence Livermore corpus. Instead of using an estimate of the clean speech STFT based on the speech input and the previous frames Wiener filter, the actual clean speech was used in the transfer function of the Wiener filter:

$$W(k,\omega) = \frac{|X_{clean}(k,\omega)|}{|X_{clean}(k,\omega)| + S_b(\omega)}$$

Thus, this Wiener filter represents the "best" Wiener filter in the sense that the estimate of the speech STFT in the filter's transfer function can be no closer to the actual clean STFT. After this "ideal" Wiener filter was created, the magnitude estimate of the speech was computed in the same manner as that of the baseline algorithm:

$$|\tilde{X}(k,\omega)| = |Y(k,\omega)| \times |W(k,\omega)|$$

The phase of the output was also computed in the same way as in the baseline algorithm:

$$\angle\tilde{X}(k,\omega) = \angle Y(k,\omega)$$

When using this system, the resulting outputs (see figure 4.4) were much different than those of the baseline system that used the same Lawrence Livermore corpus sentences (see figure 4.2). In figure 4.4, the spectrogram of the output is less noisy. More of the low-band noise has been eliminated, more of the formant information is visible, and the speech stands out from the background. The time domain output waveform is also less noisy than that of the baseline algorithm output shown in figure 4.2: the speech segments clearly stand out in the waveform in figure 4.4 whereas the speech segments in figure 4.2 do not.

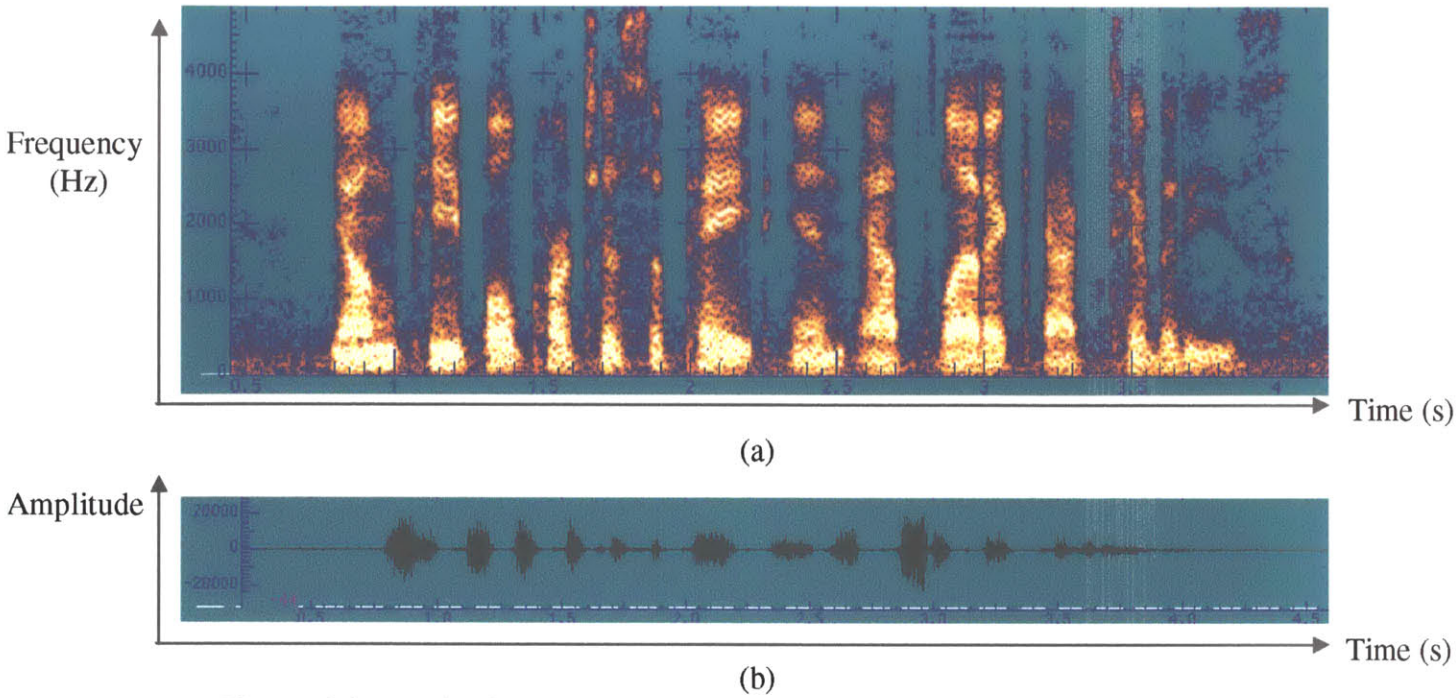


Figure 4.4: Result of Processing with "Ideal" Wiener Filter (a) Spectrogram and (b) Time Domain

All of these observations indicate that there is much room for improvement of the baseline system. However, they also indicate that a Wiener filter can only yield limited improvement. Thus the natural question becomes, what is the best magnitude enhancement. To answer this question, experiments using the actual speech magnitude were conducted.

4.3.2 Ideal Magnitude Test

For any magnitude enhancement routine, the best it can do is recover the actual clean speech magnitude as its output. Since our algorithm processes the noisy inputs in segments and then recombines these overlapping segments, this recombination may affect the system performance. To explore these affects, the clean speech STFT magnitude was used for each segment thus replacing the estimated magnitude in the baseline system:

$$| \tilde{X}(k,\omega) | = | X_{clean}(k,\omega) |$$

At the same time, the phase of each segment was kept the same as in the baseline system:

$$\measuredangle \tilde{X}(k,\omega) = \measuredangle Y(k,\omega)$$

A spectrogram showing the result of using such processing is shown in figure 4.5 (see below). As one can see from the figure, much of the noise in the speech has been eliminated when compared with the input in figure 4.1. However when compared with

the clean speech shown in figure 4.3, one notices that little of the harmonic structure has been restored. This indicates that the phase of the STFT of the speech segment is also important for proper recombination and enhancement. Since harmonics are necessary for pitch estimation, which is used for various applications such as vocoding (for example using the MELP vocoding standard), this issue of phase is very important.

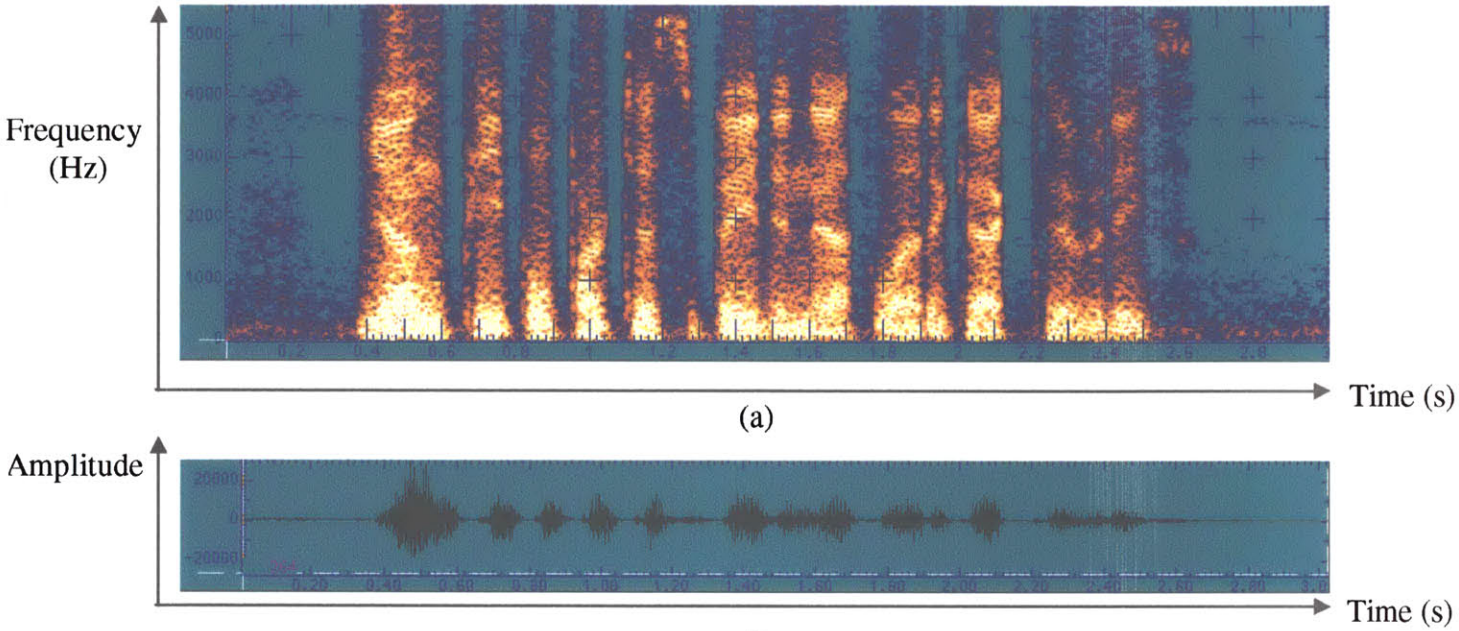


Figure 4.5: Clean Magnitude and Noisy Phase Combined (a) Spectrogram and (b) Time Domain

4.4 Phase Performance

Because the Wiener filter did not restore the harmonic nature of the speech which is important for intelligibility and vocoding, we next examined phase. The motivation for examining phase is that Wiener filtering does not change the phase of a speech segment. Thus this could be why the harmonic structure of the final reconstructed speech was destroyed.

Since no phase estimator exists, we focus on tests using the clean STFT phase. First we explore the effects of using the noisy input STFT magnitude with the clean STFT phase of a segment. Then we explore the effects of using the estimated magnitude with the clean speech STFT phase.

4.4.1 Original Magnitude and Clean Phase

To begin, the orginal noisy STFT magnitude of each segment was used in order to explore the affects of using the clean phase without any other enhancement. Specifically the magnitude and phase used was:

$$|\tilde{X}(k,\omega)| = |Y(k,\omega)| \quad \text{and} \quad \angle\tilde{X}(k,\omega) = \angle X_{clean}(k,\omega)$$

The result of this experiment is shown in figure 4.6. This result was then compared to tests using the noisy phase (thus compared to the original noisy speech input). The spectrogram of the same sentence is shown in figure 4.7 below. As one can see by examining both, the clean phase once again restores much of the harmonic structure of the speech. It also seemed to reduce a great deal of noise on its own. When conducting informal listening, it was found that the clean phase made very unintelligible speech somewhat understandable.

Observe from figures 4.3, 4.5, 4.6, and 4.7 an apparent paradox: *Modifying the short-time phase modifies the short-time magnitude.* This paradox is resolved by recalling that the enhanced signal is synthesized by an overlap-add scheme. Therefore, cleaning the phase of the noisy short-time Fourier transform may reduce noise in the magnitude of the short-time Fourier transform of the enhanced synthesized signal by virtue of overlapping and adding. In addition, the spectrogram views the synthesized waveform through a window and frame (in this example, 25 ms and 10 ms, respectively) different from that of the enhancement analysis/synthesis (12 ms and 2 ms, respectively).

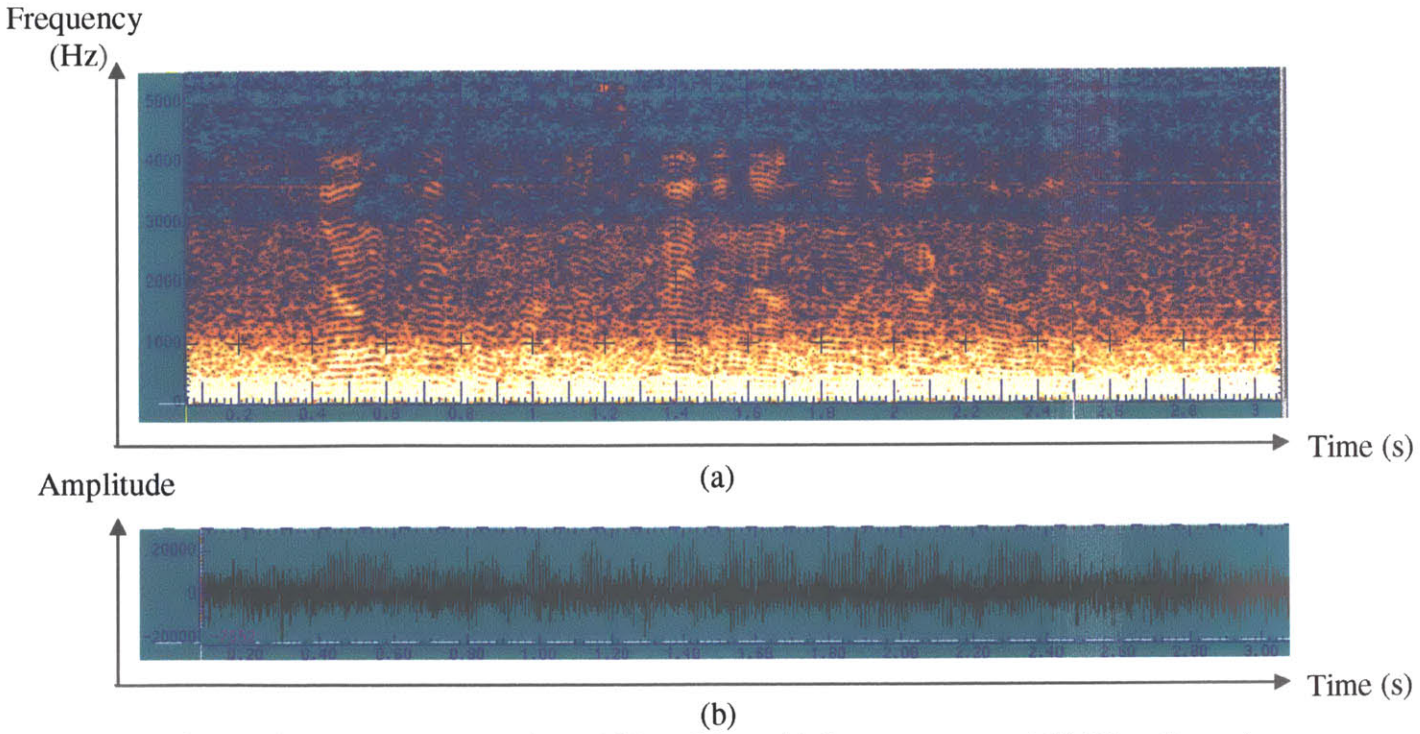


Figure 4.6: Noisy Magnitude and Clean Phase (a) Spectrogram and (b) Time Domain

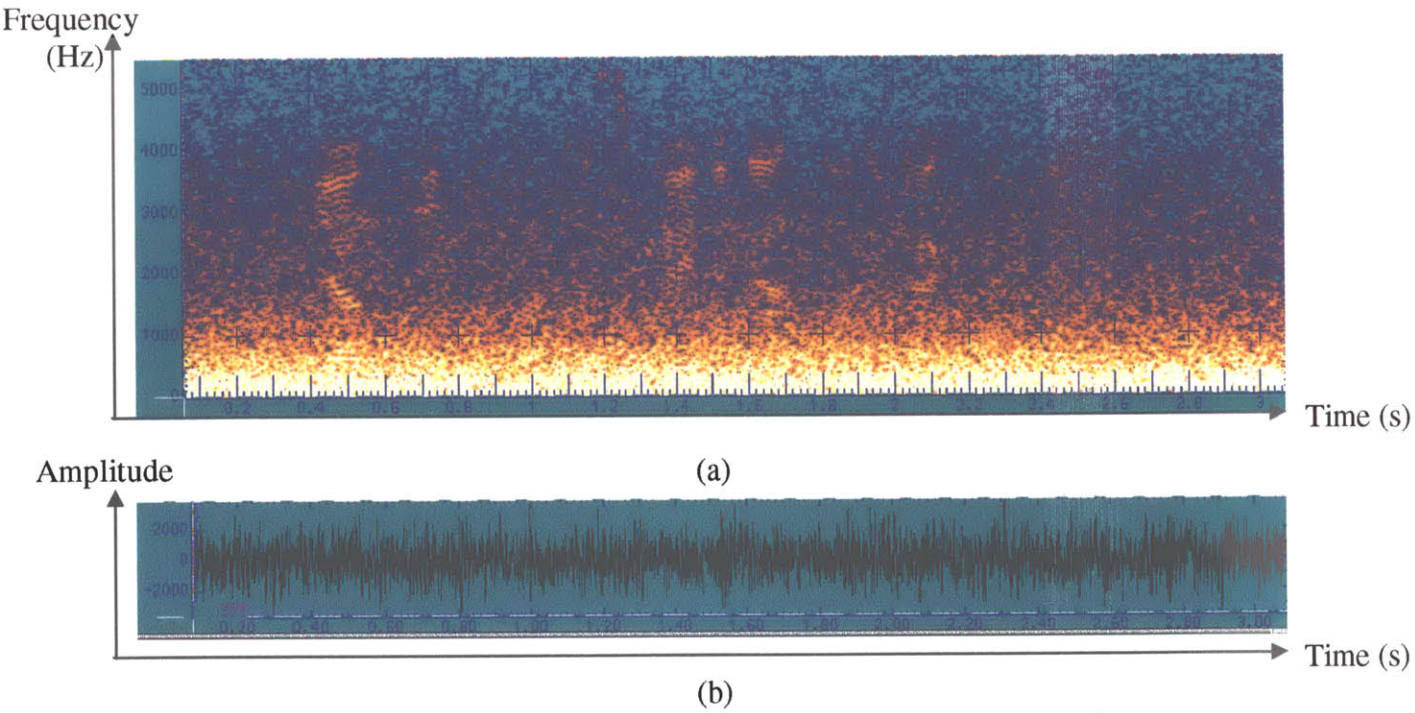


Figure 4.7: Noisy Input (a) Spectrogram and (b) Time Domain

4.4.2 Estimated Magnitude and Clean Phase

Since the clean phase alone yielded large improvements for harmonics, we next explored the affects of using the clean phase with the magnitude estimate. This is similar to the "ideal" magnitude tests in that they show what can be accomplished with the best phase achievable. It also illustrates how replacing the phase aids the baseline algorithm.

The result of this processing is shown below in figure 4.8. When compared to figure 4.7, one notes that much of the harmonics are once again restored. When compared to figure 11, one notices that much of the speech has been enhanced. Although the input was very noisy (-20 db SNR) the result yields an output whose spectrogram has speech structure that stands out more than the original input spectrogram (see figure 4.7). The result also sounds "crisper" and "clearer" than the original input.

4.5 Conclusion:

In this chapter we explored the performance limitations of Wiener filtering by examining the baseline system performance and the performance of an "ideal" Wiener filter based on the actual clean speech statistics. We showed that there was much room for improving speech magnitude estimation using Wiener filtering. We also explored several other ideal situations which are summarized in the following table 1 and compared with the original inputs and clean signals:

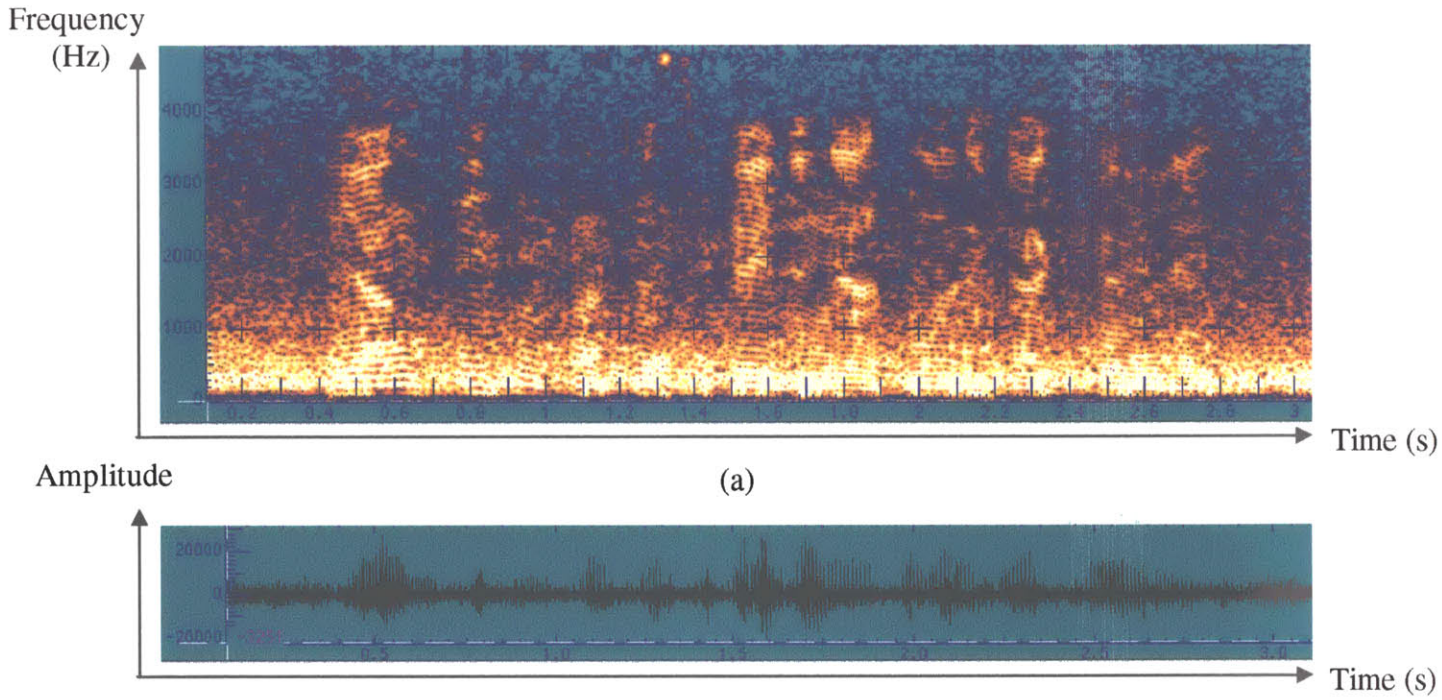


Figure 4.8: Estimated Magnitude and Clean Phase (a) Spectrogram and (b) Time Domain

Magnitude / Phase	Harmonics Degraded or Enhanced?	Background Noise High or Low?	Figure Illustration
Original / Original	Degraded	High	4.1
Clean / Clean	Enhanced	Low	4.3
Estimate / Original	Degraded	High	4.2
Clean / Original	Degraded	Low	4.5
Ideal Wiener Est / Original	Degraded	Low	4.4
Original / Clean	Enhanced	High	4.6
Estimate / Clean	Enhanced	High	4.7

Table 1: Summary of Situations and Findings: Based on 12 Sentences from 2 Male Speakers in the Lawrence Livermore Corpus

Based on these findings, it was found that there is much room for improvement using Wiener filtering if one can get at these ideal situations. Thus the focus of the rest of this thesis is on how to better reach these ideal situations using non-acoustic sensors. Chapter 5 focuses on the magnitude estimate using non-acoustic sensors. Chapter 6 focuses on the phase estimate using the non-acoustic sensors.

Chapter 5: Magnitude Estimation of Speech Spectra

This chapter discusses improvements in magnitude estimation by Wiener filtering that were made possible using several non-acoustic sensors. It begins with reviewing the baseline Wiener filtering algorithm and its performance. We then discuss Wiener filtering based on the identification of the 4 different speech classes of chapter 3. Next, we explore the uses of alternative non-acoustic sensors for use in magnitude estimation of speech. This chapter concludes by combining both ideas – filtering based on the identification of 4 different speech classes using alternative sensor data in addition to acoustic data. In this chapter the ASE corpus was used to conduct experiments since many of the alternative sensors such as the PMIC and the EGG were not present in the Lawrence Livermore corpus. Also, the ASE corpus was used in order to examine real-life noises that were not artificially added. 61 utterances spoken by a male speaker were examined in all of the following experiments.

5.1 Baseline Algorithm

5.1.1 Algorithm Details

The baseline algorithm used as a reference for comparisons was based on 2 Wiener filters. The first was used for regions of a recording that were identified as background noise and the second for regions of a recording that were identified as speech (see chapter 2). The background noise regions were greatly attenuated since the speech spectrum in the numerator of the Wiener filter transfer function was very small:

$$W(k,\omega) = \frac{\hat{S}_x(k,\omega)}{\hat{S}_x + \alpha S_b(k,\omega)} \sim \frac{0}{0 + \alpha S_b(k,\omega)} \sim 0$$

The speech estimate of the previous background frame is very small because the Wiener filter of the previous background frame is ~0 and the speech spectrum of the current background frame is estimated from the previous background frame's Wiener filter transfer function:

$$\hat{S}_x(k,\omega) = \tau(k)\hat{S}_x(k-1,\omega) + [1-\tau(k)]\,|\,\hat{X}(k,\omega)\,|^2$$
$$\sim \tau(k)0 + [1-\tau(k)]0$$
$$\text{since } \hat{X}(k,\omega) = W(k-1,\omega)Y(k,\omega) \sim 0 \times Y(k,\omega)$$

In contrast to background regions, regions identified as containing speech are not drastically attenuated since the transfer function of the Wiener filter during speech is:

$$W(k,\omega) = \frac{\hat{S}_x(k,\omega)}{\hat{S}_x(k,\omega) + \alpha S_b(\omega)} \quad > 0 \text{ for several } \omega \text{ since}$$

$$\exists \omega \text{ st } \hat{S}_x(k,\omega) = \tau(k)\hat{S}_x(k-1,\omega) + [1-\tau(k)] \mid \hat{X}(k,\omega) \mid^2 > 0$$

$$\text{and } \hat{X}(k,\omega) = W(k-1,\omega)Y(k,\omega) > 0$$

Essentially, the baseline Wiener filter algorithm divides a recording into local, short regions, each having a different desired clean speech spectrum. These different spectra are estimated by using local statistics to create a Wiener filter more tailored to that regions behavior. Thus the speech regions are processed with a distinct speech Wiener filter and the background noise regions are processed with a distinct background noise Wiener filter. This concept of dividing up a recording into regions that correspond to different speech classes and then processing each class region with different Wiener filters was further expanded upon in this thesis work (see section 5.2).

5.1.2 Baseline Algorithm Performance Examples

To determine the performance of the baseline magnitude estimation algorithm, several experiments were conducted on 61 words available in the ASE corpus using the M2H environments. In doing these experiments, speech activity detection was accomplished using the multi-detector that employed the high-pass resident microphone fused with the GEMS. Spectrograms of the outputs were made to gain insight and to compare with later experiments. Examples of 6 post-filtered utterances are shown below in figure 5.1 which are spectrograms of the words "zed", "bank', "dint', "box", "got", and "net." As one can see from the examples, the baseline magnitude estimation did a poor job of correctly estimating the magnitude on several instances. The /k/ /s/ in "box" and the /d/ in "zed" are virtually non-existent in the output waveforms. Also the voicing (from voice bars) preceding all the voiced consonants is not seen on any of the 6 spectrograms because of the very low SNR in the low band range. Also much high band noise seems to be present during these voice bars, particularly for the /z/ in "zed," /b/ in "bank," /d/ in "dint," and /g/ in "got." Elimination of this high band noise was the first objective of our research involving improvement of the magnitude estimate of speech. We sought to do this by using Wiener filtering based on the identification of different speech classes.

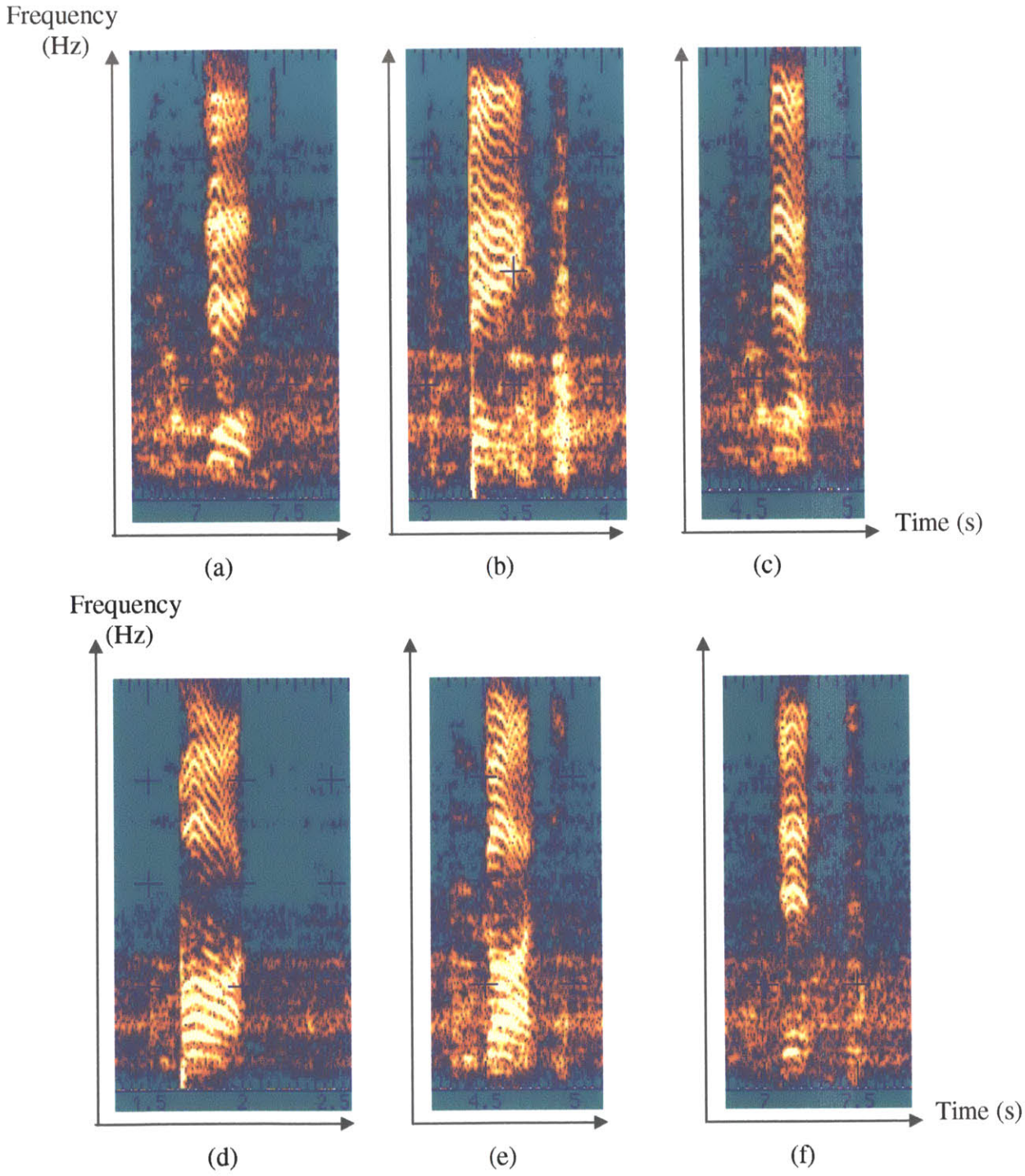


Figure 5.1: Examples of enhancement using the base-line algorithm (1 speech class and 1 input) for the words (a) "zed", (b) "bank", (c) "dint", (d) "box", (e) "got", and (f) "net"

5.2 Processing Based on Different Speech Classes Using One Sensor

It was hoped that dividing the speech waveform into different identified classes would improve magnitude estimation because the spectrums of different speech phones can be drastically different. Yet at the same time, some phones have similar spectral characteristics and can be grouped into the same phone-class. For example unvoiced phones tend to be more dominated by high-frequency energy whereas voice bar regions are dominated by low-frequency energy and contain no high band energy at all. The baseline algorithm treated all speech the same. Thus the speech segment's Wiener filter

$$W(k,\omega) = \frac{\hat{S}_x(k,\omega)}{\hat{S}_x(k,\omega) + \alpha S_b(\omega)}$$

determined by:

$$\hat{S}_x(k,\omega) = \tau(k)\hat{S}_x(k-1,\omega) + [1-\tau(k)]\,|\,\hat{X}(k,\omega)\,|^2$$

$$\hat{X}(k,\omega) = W(k-1,\omega)Y(k,\omega)$$

may be very far from the "ideal" Wiener filter described in chapter 4. For example, if a voice bar was captured in the last Wiener filter segment, much of the Wiener filter's energy will be concentrated in the low band, below 500 Hz. Thus $W(k,\omega)$ will be low pass in nature. This could negatively degrade some of the high band energy of the estimate $\hat{S}_x(k,\omega)$ as it transitions from a voice bar region into a more high energy region such as the burst in a plosive like /b/ or a fricative like /v/.

5.2.1 Division of Speech Classes

In order to improve magnitude estimation, this work separated a recording into 4 different classes of speech: (1) background noise, (2) voiced speech, (3) unvoiced speech, and (4) low-band regions. Each region was then processed with a different Wiener filter that was more attuned to the statistics of a particular class. These regions were identified using the class type multi-detector that was based on the GEMS and the high pass resident microphone detectors (see class based detection in chapter 3).

Background noise regions are regions of a recording that the detection algorithm had identified as containing no noise (see chapter 3). Consequently, the initial Wiener filter for them was designed to attenuate the region drastically.

Voiced speech regions are parts of speech that use vocal fold vibration to provide an excitation of the vocal tract. The sounds produced in such speech include all vowels, voiced fricatives and other voiced plosives and voiced fricatives such as a /z/,/v/, /b/, /d/, or /g/, and liquids and glides such as /w/, /y/, /r/, and /l/. These regions of speech contain both high frequency and low frequency sounds that can be very harmonic. An example of such voiced regions are the /z/ and /e/ in the word "zed" shown in figure 5.2.

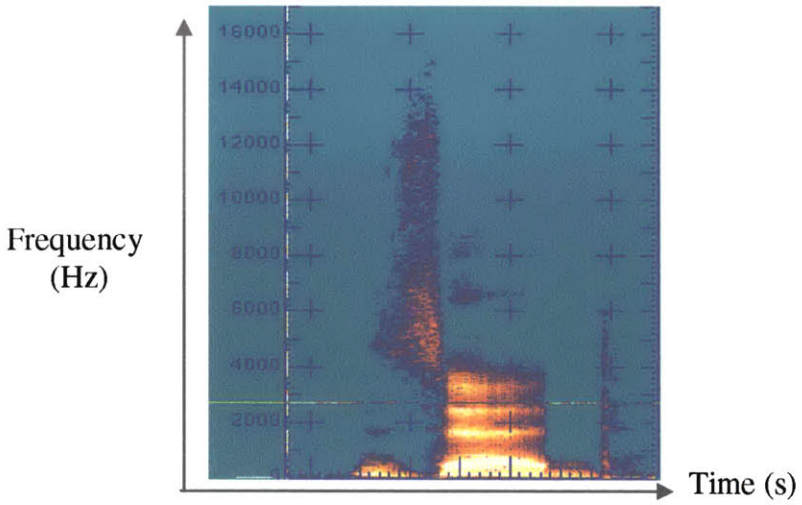


Figure 5.2: Spectrogram Example of Voiced Regions in the Word "Zed" using Quiet Environment (Not Processed)

Unvoiced speech regions are composed of unvoiced consonants such as /f/, /t/, /p/, /k/, /s/, /tʃ/ , and /ʃ /. These sounds tend to be more high frequency in nature and not harmonic. Consequently the Wiener filter employed on this class of sounds was high pass in nature (see description of Wiener filters below). An example of such an unvoiced region is the /ʃ /in the word "shag" shown in figure 5.3 (centered around 4500 Hz).

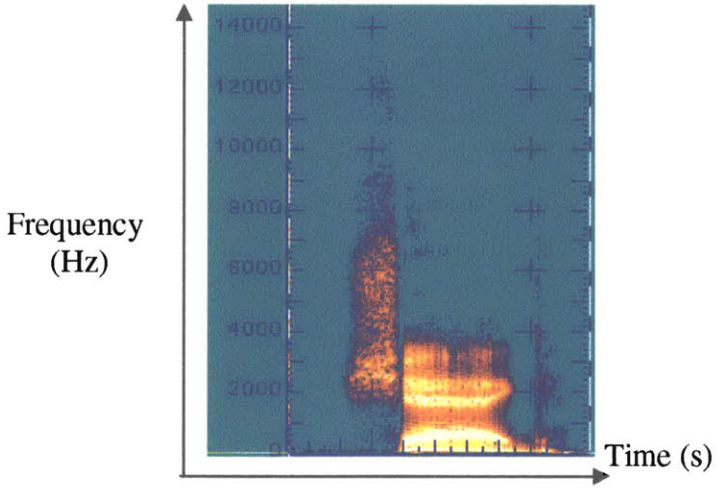


Figure 5.3: Spectrogram Example of Unvoiced Regions in the Word "Shag" using Quiet Environment (Not Processed)

Low-band regions are regions of speech that contain only very low frequency sounds such as a /m/ and /n/ or a voice bar that precedes a voiced consonant such as /b/, /g/, /d/, and some /z/s, and /v/s. Voice bars are one of numerous cues that are used by humans for proper identification of phones. Thus they are important for intelligibility and are one focus of this thesis. Because these low-band regions have only low frequency content, they are processed by a Wiener filter that is low pass in nature. An example of such a voice bar region is the voicing preceding the /b/ in the word "boot" shown in figure 5.4.

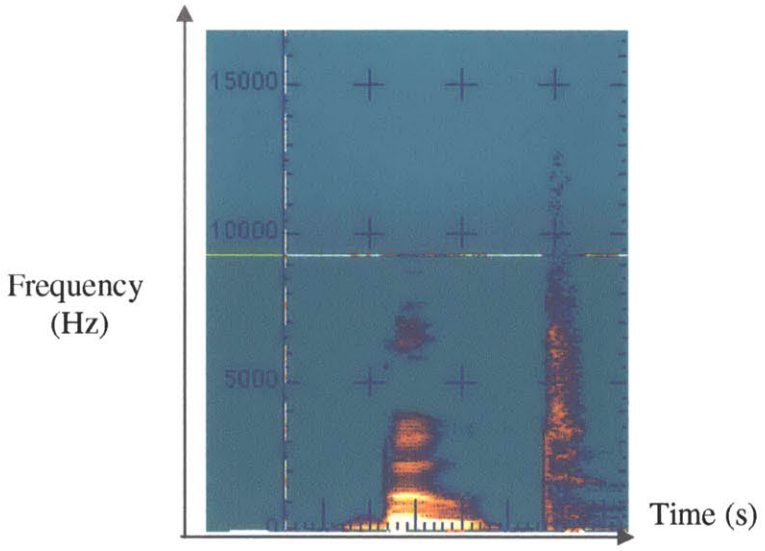


Figure 5.4: Spectrogram Example of Voice Bar Region in the Word "Boot"

5.2.2 Filtering of Each Speech Class

The Wiener filters for each of the four speech classes were divided into two parts: one part for the high-frequency bands and one for the low-frequency bands. Each band uses a separate Wiener filter as shown in figure 5.5 below. The Wiener filters are divided to process the different classes in a manner more appropriate for each class. For example, the low-band class containing voice bars and some low-band nasals would be attenuated in the high band and not in the low band.

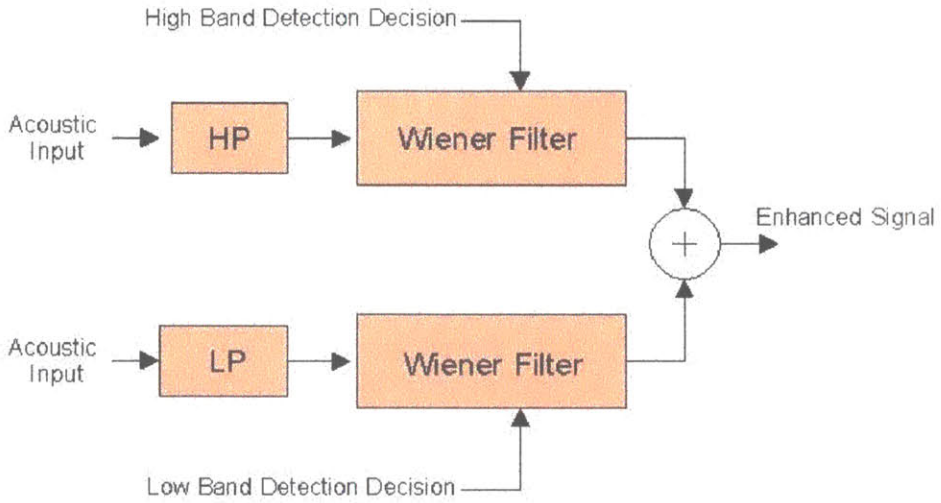


Figure 5.5: Overview of Two-Band Filtering

Both the high-frequency and low-frequency components are formed by using the statistics of the noise-canceling resident microphone to create the speech estimate used in the creation of a Wiener filter:

$$W(k,\omega) = \frac{\hat{S}_x(k,\omega)}{\hat{S}_x(k,\omega) + \alpha S_b(\omega)}$$

where $\hat{S}_x(k,\omega)$ is the speech spectrum estimate and $S_b(\omega)$ is the noise spectra. Each of the four classes of speech uses the high-pass Wiener filter and the low-pass Wiener filter in different ways. If a region of speech is identified as a low-band class sound, the low-pass Wiener filter is set to pass the speech and the high-pass Wiener filter is set to greatly attenuate the high frequencies, much like the original Wiener filter for background noise. This process still happens adaptively as in the original Wiener filter used in the baseline system. If a region is identified as a voiced speech region, the low-pass Wiener filter and the high-pass Wiener filter are set to pass the speech and keep most of the input waveform's energy. If the region is identified as an unvoiced region, the low-pass Wiener filter is set to attenuate the low frequencies and the high-pass Wiener filter is set to sharpen the high frequencies and keep most of the input waveform's high frequency energy. If a region is identified as a background region, then both Wiener filters are set to attenuate both regions since they contain only noise. A summary of this class based processing is shown in figure 5.6. In the figure, regions that are attenuated to reduce noise are marked "Attenuated" and regions that are passed and that maintain most of the energy of the band are marked "Passed."

	High Band (> 500 Hz)	Low Band (< 500 Hz)
Low-Band Regions	Attenuated	Passed
Voiced Regions	Passed	Passed
Unvoiced Regions	Passed	Attenuated
Background Noise Regions	Attenuated	Attenuated

Figure 5.6: Wiener filtering of the high and low bands for each speech class region

In order to determine the performance of this system, the same 61 utterances appearing in the ASE corpus and used for the baseline algorithm experiments were examined. Spectrograms of the same 6 words explored in figure 5.1 above are shown below in figure 5.7.

When compared to figure 5.1, one can see that much of the higher band noise has been eliminated in regions such as that preceding the /z/ in "zed," the /d/ in "dint," and the /g/ in "got." Also much of the lower-band noise has been removed, and the /k/ /s/ in box and the /d/ in "zed" are now visible in the spectrogram. Although this is a major improvement, it was noticed that much of the lower-band harmonics are missing in the magnitude estimates, most noticeably the voice bars and nasals. This is because the resident microphone of the noisy ASE environment acts like a high-pass filter and thus

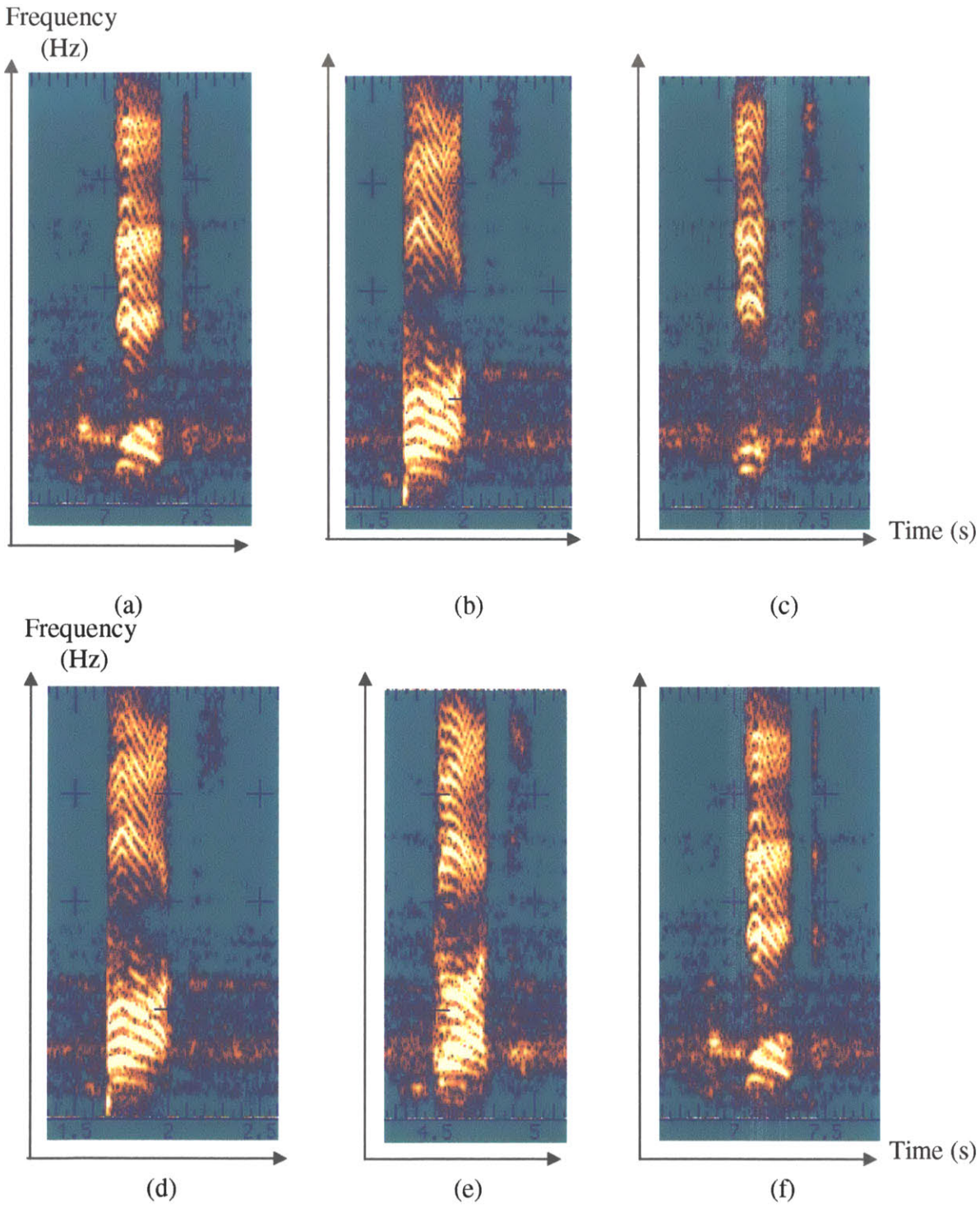


Figure 5.7: Examples of enhancement based on 4 speech classes using the resident microphone as input for the words (a) "zed", (b) "bank", (c) "dint", (d) "box", (e) "got", and (f) "net"

eliminates the energy in this region. Because of this property and because the resident microphone has greater SNR at higher frequencies (above 3000Hz the noise rolls off), it was decided to explore other sensors to use in conjunction with the high band of the resident microphone for magnitude estimation.

5.3 Magnitude Estimation with Two Sensors

5.3.1 Two-Sensor Processing with Two Speech Class

In order to make comparisons with the other early systems, processing was first conducted with two speech class, as in the baseline algorithm. The purpose of this was to test the idea of using two sensors alone, and then to later combine this with the concept of processing speech based on the identification of different speech classes. In this set of experiments, the speech was once again divided into two bands: a high band above 500 Hz and a low band below 500 Hz. The high band used the resident microphone as its input. Consequently the high band Wiener filter was created using the resident microphone in the speech object:

$$W(k,\omega) = \frac{\hat{S}_x(k,\omega)}{\hat{S}_x(k,\omega) + \alpha S_b(\omega)}$$

$$\hat{S}_x(k,\omega) = \tau(k)\hat{S}_x(k-1,\omega) + [1-\tau(k)]\,|\,\hat{X}(k,\omega)\,|^2$$

$$\tilde{X}(k,\omega) = Y(k,\omega) \times W(k,\omega)$$

In contrast, the low band Wiener filter was created using the PMIC in the speech object:

$$\tilde{X}(k,\omega) = X_{PMIC}(k,\omega) \times W(k,\omega)$$

The rationale for using the PMIC in the low band is that it has high SNR in the low band and seemed to contain a fair amount of vocal tract information. When listening to the PMIC waveform, one could often determine what was being said, even though the waveform more sounded low-passed and muffed than normal speech. Also, when examining a Spectrogram of the PMIC waveform, one notices that it has much of the same harmonic structure as that of normal speech.

The rationale for using the resident microphone in the high band is that it has high SNR in the high band even though it has low SNR in the low band. The noise in the resident microphone seemed to roll off at around 1500 Hz. Because of this, it was hoped that using it in the high-band Wiener filter (and the PMIC in the low-band Wiener filter) would improve the result of the magnitude estimation.

To make comparisons, the same 61 utterances were once again processed and then examined. Examples of the words in figure 5.1 are shown below in figure 5.8. As one can see, using the two sensors in conjunction adds much more low-frequency information. Voice bars, which are important for intelligibility of voiced consonants such

as a /b/, /z/, /d/, /v/, etc, are very visible and present. Also the first few harmonics and formants are now more present and much of the overall noise was eliminated. When conducting informal listening tests, we found that the resulting sounds were "fuller" and "clearer."

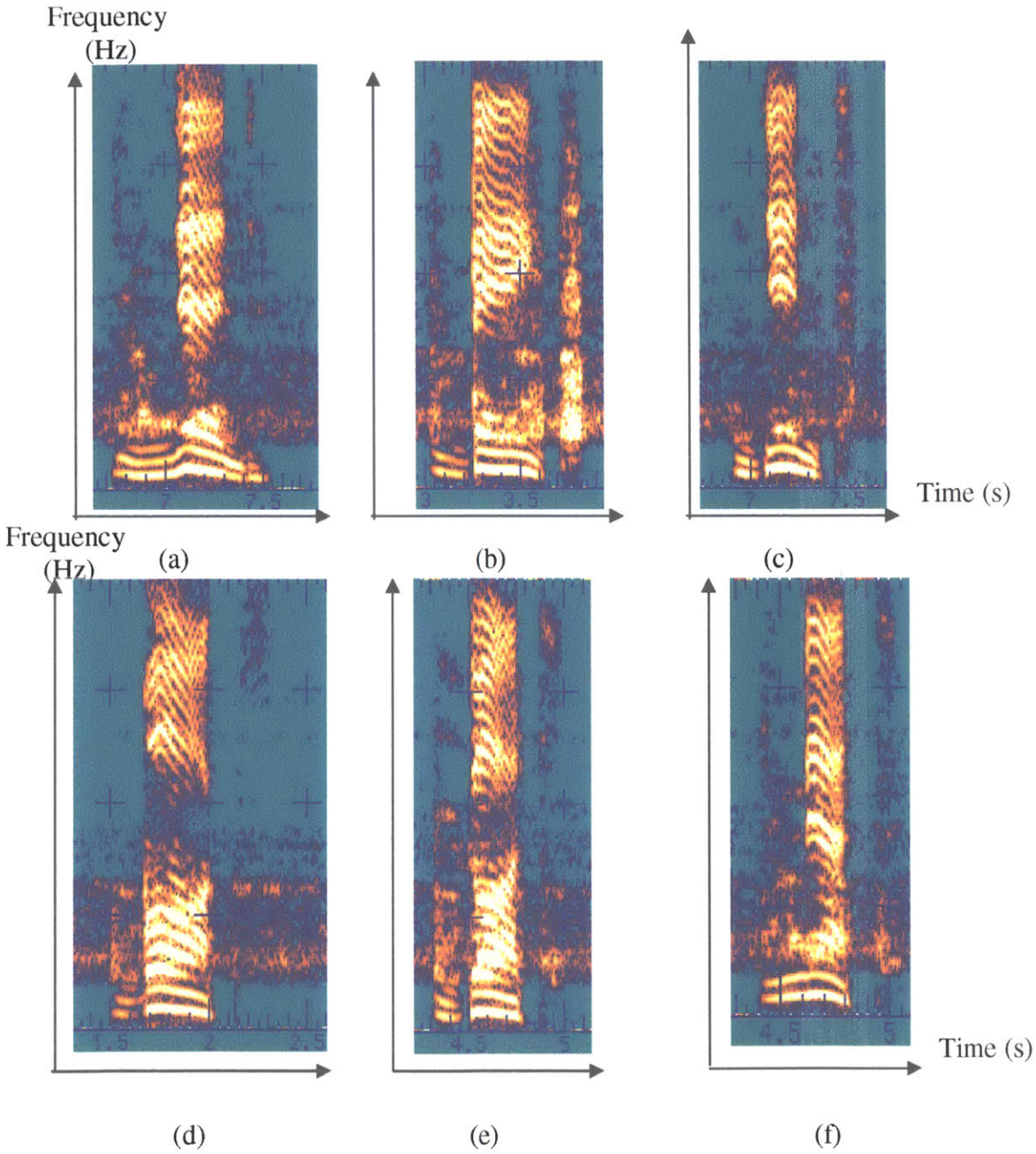


Figure 5.8: Examples of enhancement based on 1 speech classes using the resident microphone as input in the high-band and the PMIC in the low-band for the words (a) "zed", (b) "bank", (c) "dint", (d) "box", (e) "got", and (f) "net"

5.3.2 Two Sensor Processing with Multiple Speech Classes

Due to the success of the previous experiments, we sought to merge the separate ideas (class based processing and separate high-band and low-band processing based on different sensor inputs) into one system. This system thus conducted speech class identification and processed the four different speech regions in the same manner as in section 5.2. The difference now, is that for the high-band Wiener filter, the resident microphone is used as input, and for the low-band system, the PMIC located at the throat above the glottis is used as input, as in section 5.3.1. Experiments using this setup were conducted on the same 61 utterances and compared to the other tests. Examples are shown in figure 5.9 below. When compared to figure 5.8, one sees that this system helps reduce much of the noise above voice bar regions and also helped to make some of the consonants further stand out (for example the /d/ in "zed"). When compared to the baseline system (figure 5.1) one sees an improvement: the noise has been further reduced, many of the consonants are more clearly visible, the voice bars are now visible, and the low-band harmonics have been restored. Informal listening of these utterances and comparing them to those processed by the previous systems also sounded superior in terms of "fullness" and noise suppression. However, the sounds seemed to have an abnormal amount of bass in them. This abnormal bass-like sound was reduced in the final system developed by using formant sharpening in the form of pre-emphasizing the PMIC signal to reduce the low-band harmonic energy.

5.4 Conclusion

In this chapter we presented modifications to traditional Wiener filtering based on information present in non-air-acoustic sensor waveforms. Specifically, the noise canceling resident microphone (which has been shown to reduce noise in the high bands while preserving high frequency speech) was used to create a high-band Wiener filter and the PMIC (which has been shown to reduce noise in the low bands while approximately preserving low-frequency speech) was used to create a low-band Wiener filter. Each Wiener filter then used speech class-based detection, which was only made possible by the use of the GEMS and resident microphone used in conjunction with each other. This class-based detection allowed the high-band and low-band Wiener filters to more selectively attenuate and reduce noise in their band during background noise and during speech as well. The result of these modifications yielded Wiener filtering that produced sharper, more defined spectrograms containing more speech structure, and speech that sounded, through informal listening, much more "natural", "full", and perhaps understandable. In each experiment in this chapter, 61 words spoken by a male were examined since only one male speaker was initially available. Future work will involve more vigorous tests involving more speakers, both male and female, which have become available during the final weeks before this thesis work.

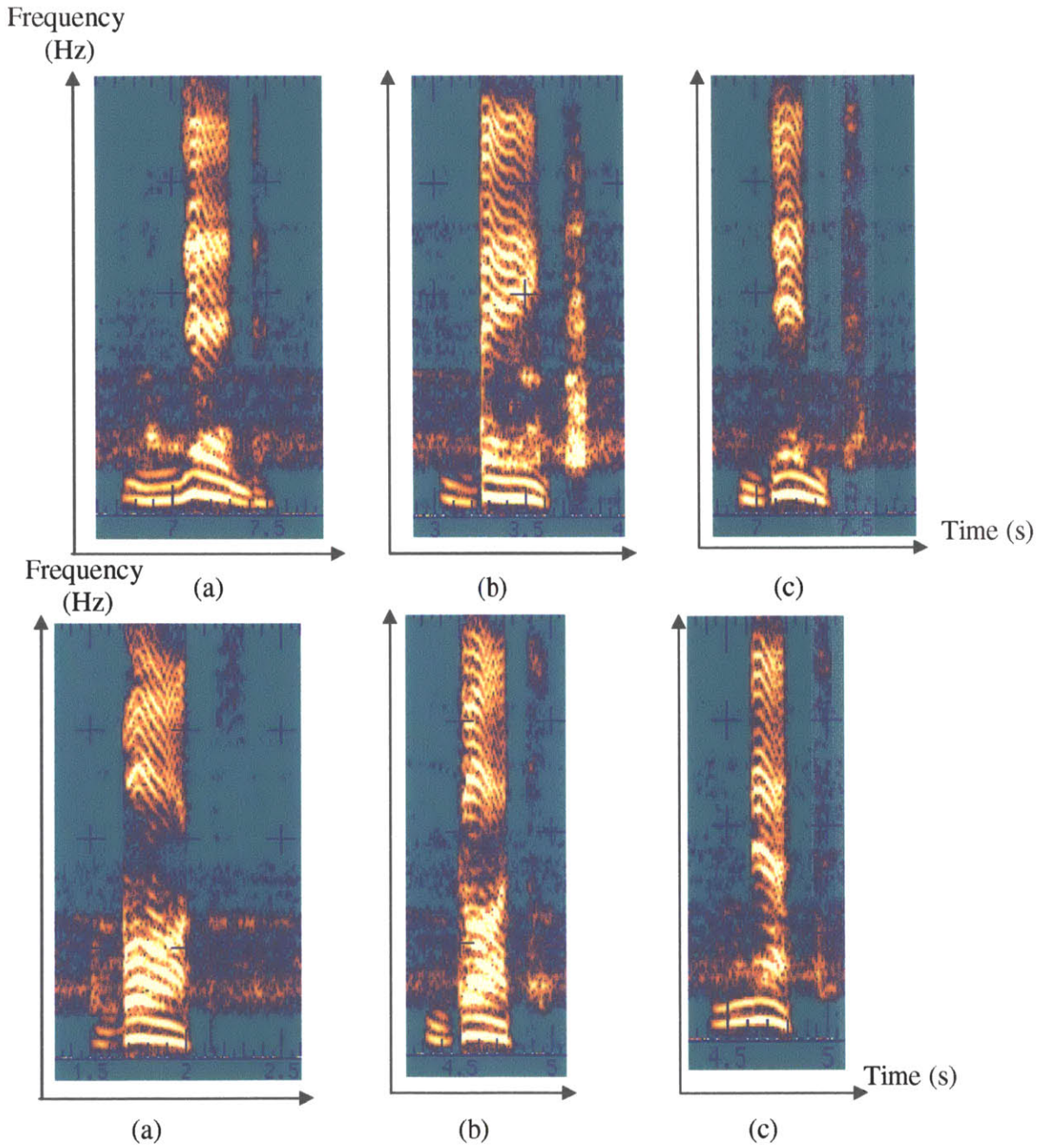


Figure 5.9: Examples of enhancement based on 4 speech classes using the resident microphone as input in the high-band and the PMIC in the low-band for the words (a) "zed", (b) "bank", (c) "dint", (d) "box", (e) "got", and (f) "net"

Chapter 6: Phase Estimation

Unlike magnitude estimation, phase estimation has not been a major focus of research in signal processing. Research has focused on creating a synthetic phase estimate based on a magnitude estimate [Oppenhiem and Schafer, 1999]. Observations by Vary have shown that phase has importance in noise reduction for signal-to-noise ratios less than 3 db [Vary, 1985]. Since some of the noise environments in this thesis are less than this threshold, phase estimation was a major focus of this thesis.

This chapter focuses on obtaining a phase estimate that produces similar affects such as that of the clean phase explored in chapter 4. Namely, a phase estimate is sought that restores harmonics and improves the quality and perhaps intelligibility of speech. This chapter begins by examining the ability of non-acoustic sensors to provide a phase estimate. It then examines the performance of a synthetic phase based on a combination of a vocal-tract phase estimate added with a glottal phase estimate. It then concludes by examining band-dependent phase replacement.

6.1 GEMS Phase

The GEMS sensor was first examined for use in phase replacement using the Lawrence Livermore corpus. For each 12-ms segment of speech, the phase of the GEMS over the windowed frame was computed. Then it was combined with the estimated magnitude. This resulted in a segment with an estimated magnitude based on the acoustic microphone and a phase from the GEMS. The segments were then combined to synthesize the speech estimate, as done in chapter 5 and the baseline system. A spectrogram of the result is shown below in figure 6.1. This result was then compared to processing using the original and clean phases of chapter 4 (which are again shown here in figures 6.2 and 6.3).

As one can see from the figures, the GEMS restored the harmonics of the speech, similar to the effect of the clean phase. However, the examples using the GEMS resulted in estimated speech that sounded " buzzy". One reason for this could be that the GEMS measures glottal vibrations of the vocal folds and the tracheal wall around the vocal folds, not the true response of the entire vocal track. This could result in a " buzzier" signal for phase estimation than desired because the resulting phase is excessively coherent across frequency. To restore the entire phase of the speech, one needs the missing phase of the vocal tract. Thus one possible way to obtain a better phase estimate may be to combine the glottal phase estimate (obtained by using the GEMS phase) with a phase estimate of the vocal tract.

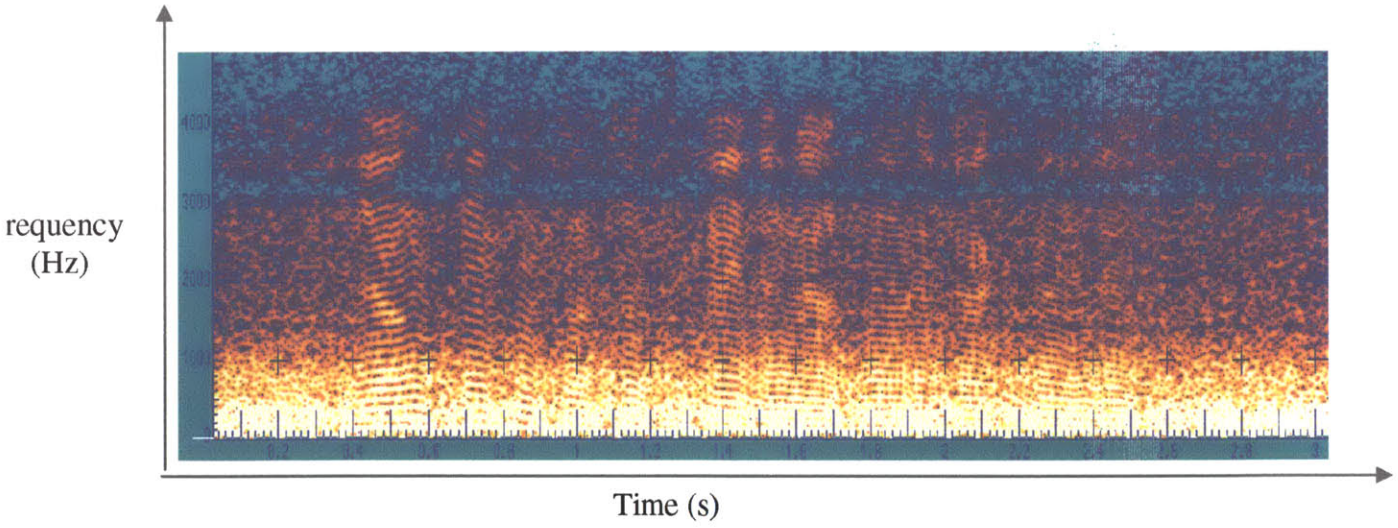


Figure 6.1: GEMS Phase with Estimated Magnitude

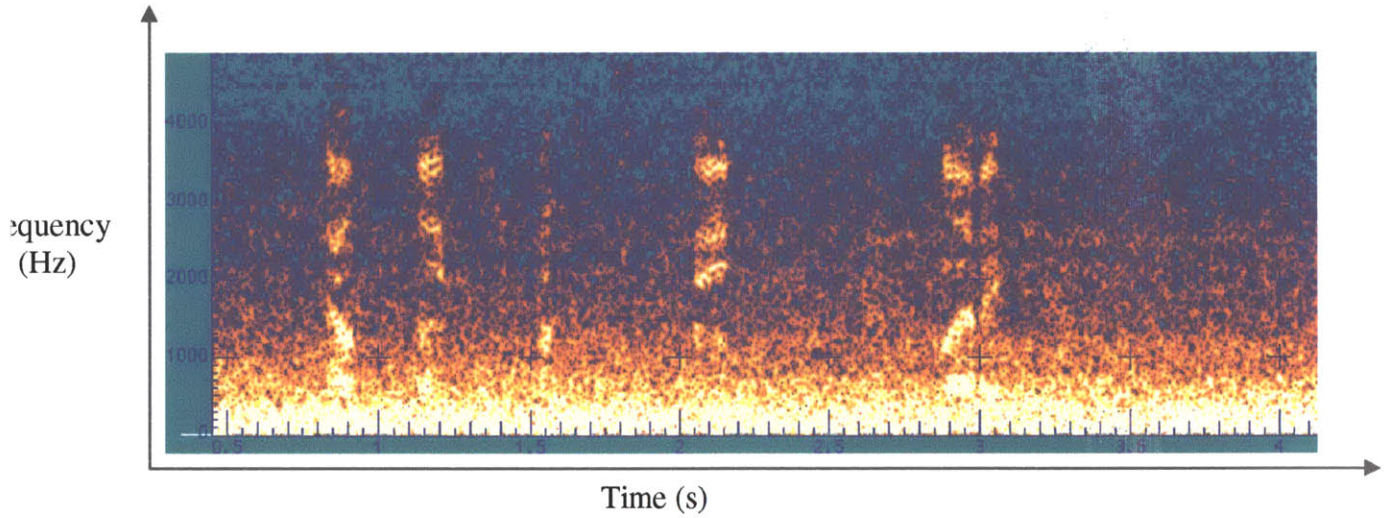


Figure 6.2: Original Phase with Estimated Magnitude

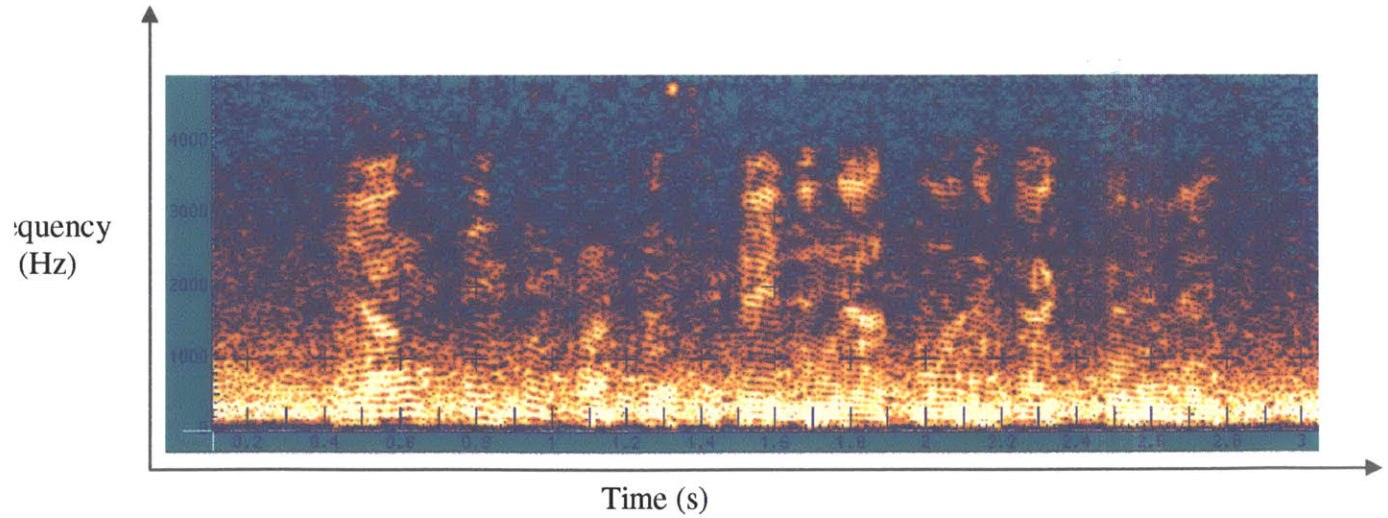


Figure 6.3: Clean Phase with Estimated Magnitude

6.2 Other Sensor Phases

In an attempt to obtain a better total phase estimate, the PMIC phase was examined and compared to the GEMS and original phase. Tests involving these experiments used 61 utterances from a male speaker in the ASE corpus since the PMIC was only available in it (future work will focus on other speakers). Consequently, the clean signal was not available for testing purposes. An example of the processing with the PMIC phase on the word "weed" is shown in figure 6.4a. This result was compared to the processing using the original phase in figure 6.4b and using the GEMS phase in figure 6.4c.

Both the GEMS and the PMIC phase produce similar results. The high-band harmonics around 3500 Hz in both examples are degraded compared to the original phase example in figure 6.4b. This is likely due to the severely low SNR of the GEMS and PMIC in this band since both of the signals roll off before 3000 Hz. In contrast the resident microphone has high SNR in this upper band and does not roll off. Consequently, the resident microphone used in this high band yields a more accurate phase function than either the GEMS or the PMIC. Below approximately 2000 Hz, both the PMIC and the GEMS phase accentuate the harmonics of the speech. This effect is most noticeable between about 1200 Hz and 500 Hz in figures 6.4a-c. This is likely due to the fact that the 500 Hz – 1200 Hz band of the estimated magnitude has the lowest SNR. Thus, as one would expect, an estimated phase used in this band has the largest effect.

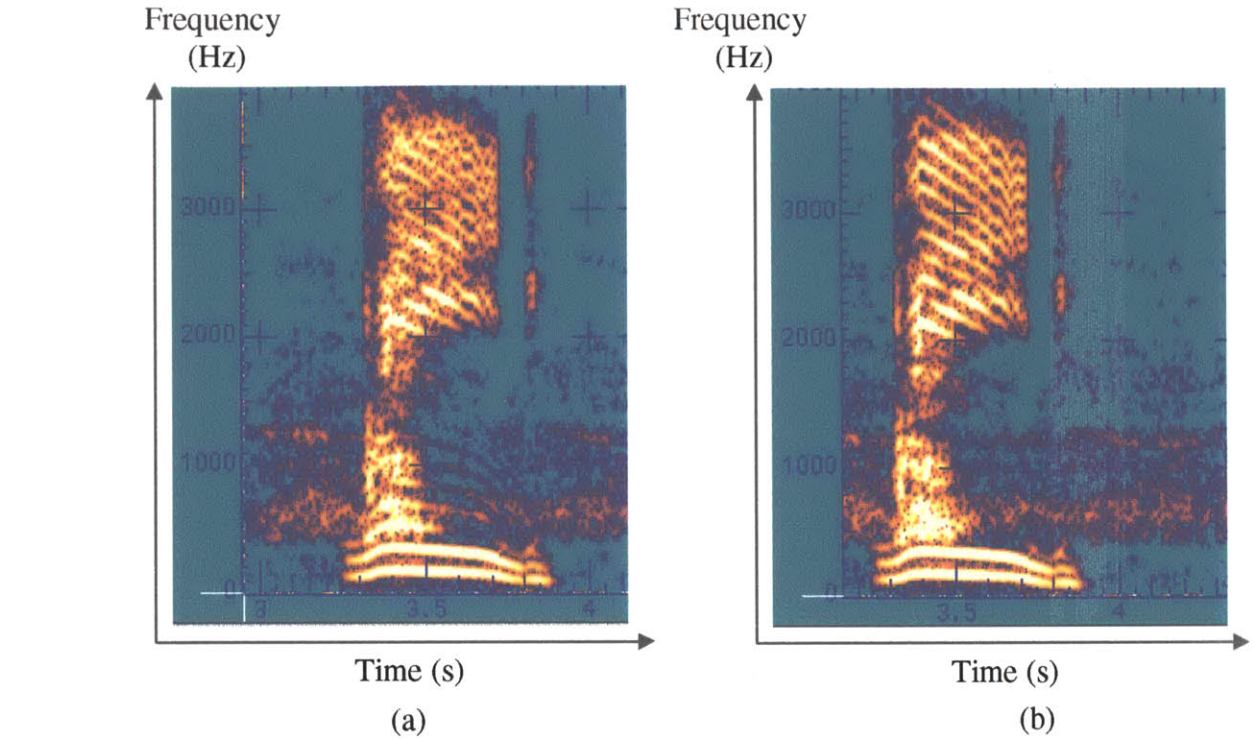
When listening to signals with the PMIC phase and comparing them to signals with the GEMS phase, both sound very similar. This seems to suggest that the PMIC phase also does not contain all the necessary vocal tract information to create an estimate of the clean speech phase. This motivates obtaining an improved speech phase estimate by combining a glottal phase estimate (obtained by using the GEMS phase) with a phase estimate of the vocal tract.

6.3 Synthetic Phase

One way to derive the phase estimate of the vocal tract is from the estimate of the magnitude of the speech. If we assume that this magnitude contains only vocal tract information (which may be approximately valid) and that the vocal tract is a minimum phase system, then we can use the Hilbert transform of the magnitude to find a phase estimate [Oppenhiem and Schafer, 1999].

Since the mid-band region between 500 Hz and 1200 Hz was shown to be the only region with an estimated magnitude with severe enough SNR to benefit from phase replacement (see section 6.2 above), we conducted phase replacement tests in the mid-band only. First, we created a synthetic phase based on the glottal phase estimate from the GEMS and the vocal tract estimate based on the Hilbert transform of an estimated magnitude. A diagram illustrating this is shown in figure 6.5. Specifically the synthetic phase was obtained by:

$$\theta_{synthetic}(k,\omega) = \theta_{GEMS}(k,\omega) + \theta_{Hilbert}(k,\omega)$$



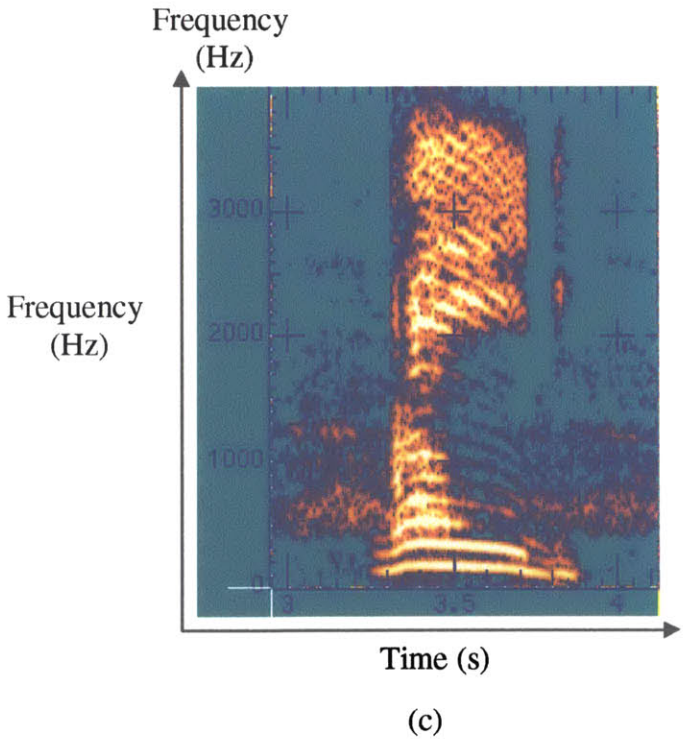


Figure 6.4: Phase construction for the word "weed" using (a) the PMIC phase, (b) the original phase from the resident microphone, and (c) the GEMS phase

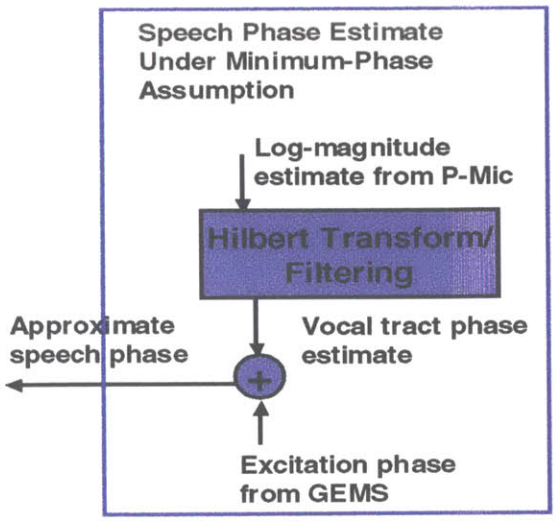


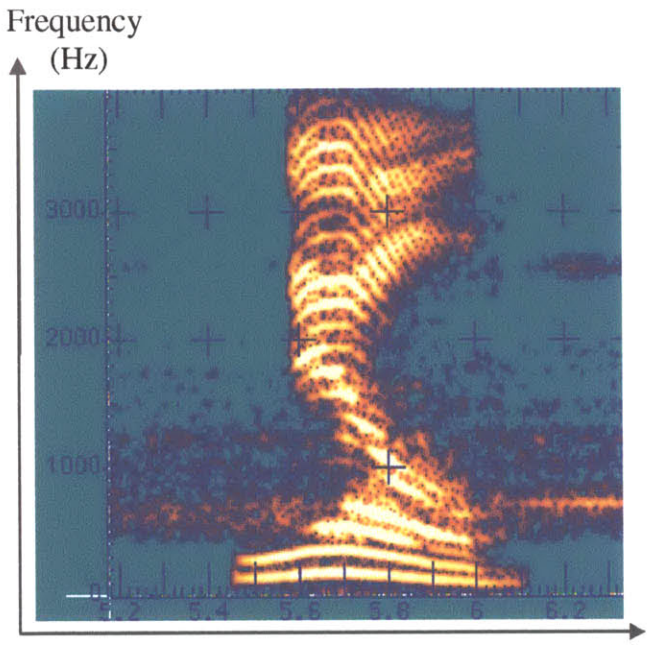
Figure 6.5: Phase construction by addition of the GEMS phase and a synthetic phase derived from the PMIC signal under a minimum-phase assumption. The Hilbert transform/filtering module implements homomorphic filtering of the enhanced PMIC signal

Where $\theta_{GEMS}(k,\omega)$ is the GEMS phase and $\theta_{Hilbert}(k,\omega)$ is the Hilbert phase of an estimated magnitude for the kth speech frame.

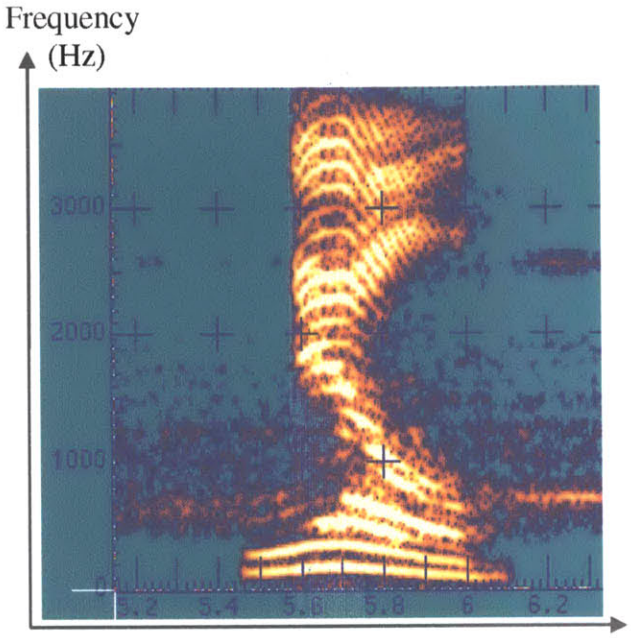
The first test (see figures 6.6b and 6.7b) used the Hilbert phase obtained from the magnitude estimate of the processed speech. The second test (see figures 6.6c and 6.7c) used the Hilbert phase obtained from the magnitude of an enhanced PMIC signal. The third test used just the GEMS phase and was used to compare with the two previous synthetic phase tests (see figures 6.6d and 6.7d). The final fourth test used the original phase and was also used as a baseline with which to compare. It is shown in figures 6a and 7a.

When examining figures 6.6 and 6.7, one can note that the cleanest, sharpest, and fullest harmonics for the 500 Hz – 1200 Hz mid-band are created by using the synthetic phase based on the enhanced PMIC magnitude. This is to be expected since the PMIC has higher SNR in this band than the resident microphone and since the GEMS signal does not contain all the vocal-tract information required to make a speech phase estimate. In informal listening, the synthetic phase based on the PMIC also sounds slightly cleaner and somewhat more natural than the other phase replacement strategies.

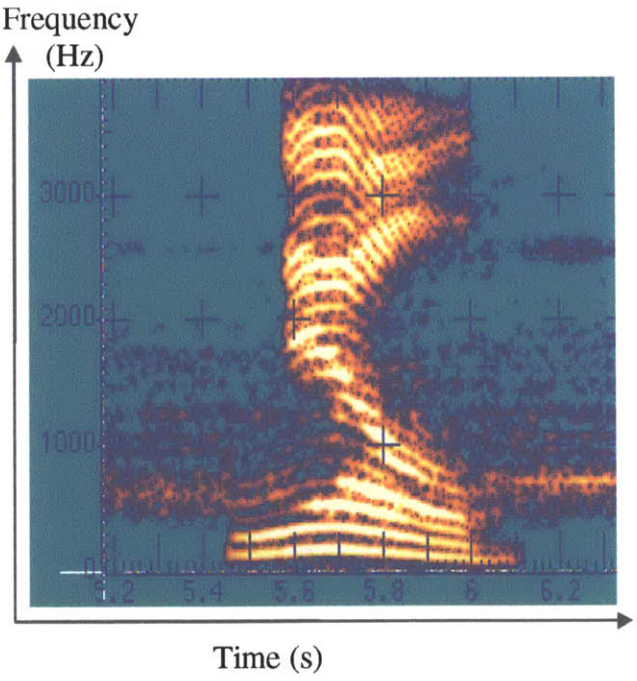
This aural difference is much more noticeable after encoding a speech signal using MELPe [Wang, Koishida, Cuperman, Gersho and Colhura, 2002].This is likely due to the dependence of the MELPe encoder on speech harmonics to create a pitch estimate which is used in encoding. Since one of the main applications of this thesis was for speech encoding using vocoding standards, such as MELP [McCree et al, 1996] and MELPe, this phase scheme is used in the final system of this thesis. Examples that use the MELPe encoding are shown in figures 6.8 and 6.9. One notices that the example with the synthetic phase estimate in the mid-band produces coded speech containing more harmonic structure. This improved harmonic structure likely causes an increased naturalness and perhaps understandability of the encoded speech that was perceived in informal listening. Future work will determine if this improves intelligibility by using Diagnostic Rhyme Tests.



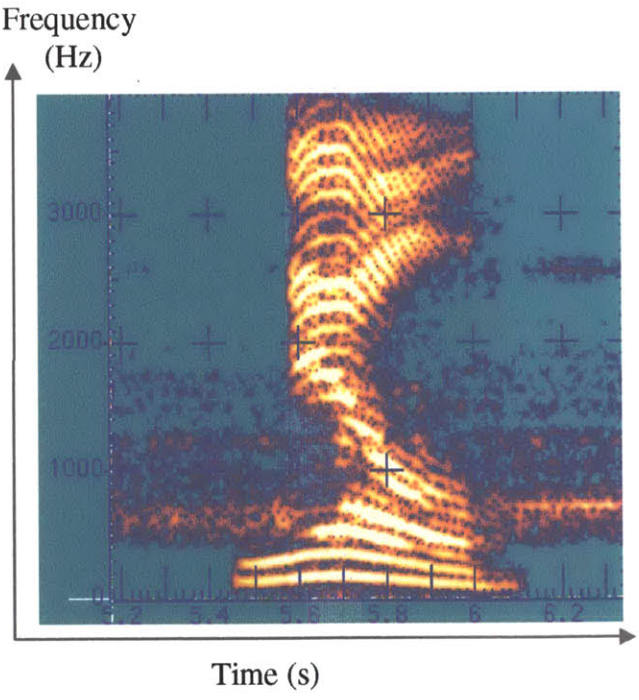
(a) Original Mid-band Phase



(b) Resident Mic Synthetic Phase

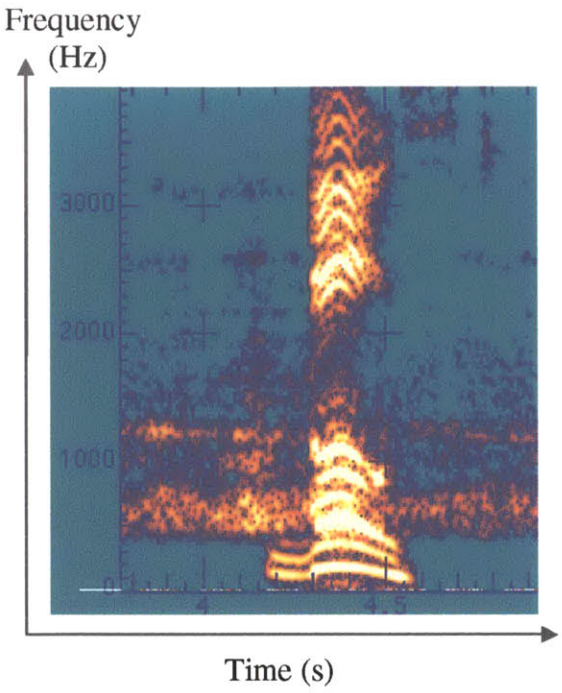


(c) PMIC Synthetic Mid-band Phase

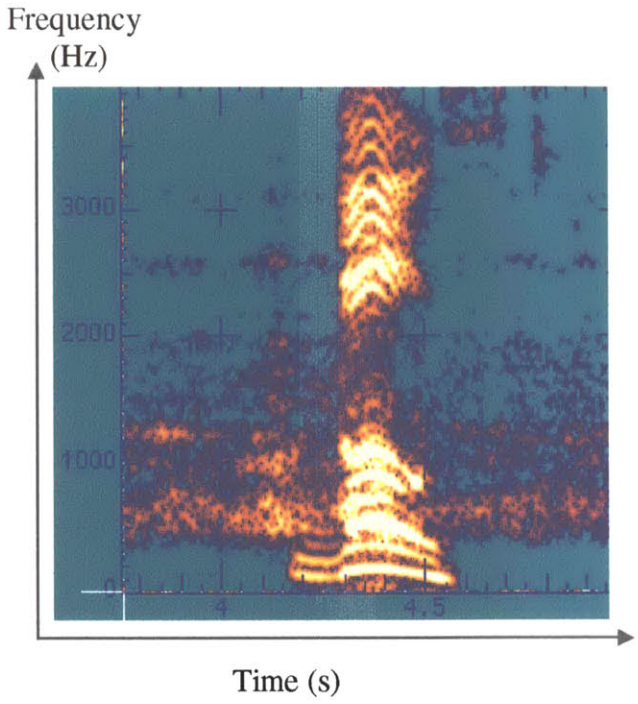


(d) GEMS Mid-band Phase

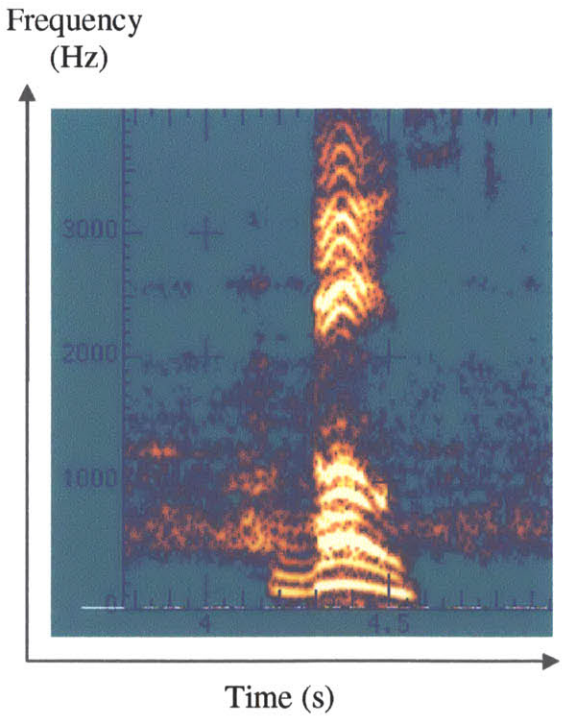
Figure 6.6: Phase construction results for the word "nil"



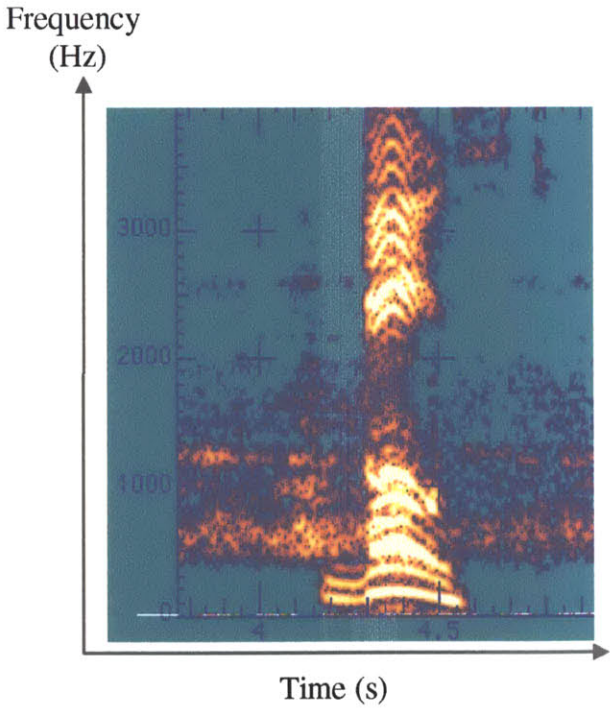
(a) Original Mid-band Phase



(b) Resident Mic Synthetic Phase



(c) PMIC Synthetic Mid-band Phase



(d) GEMS Mid-band Phase

Figure 6.7: Phase construction results for the word "boast"

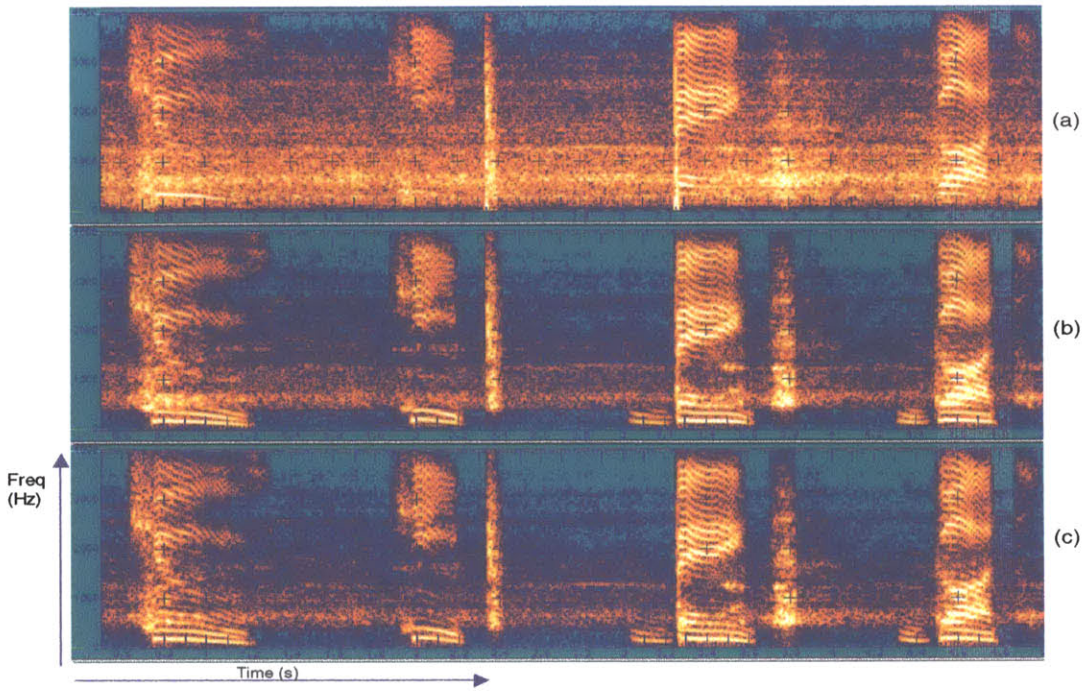


Figure 6.8: Example spectrograms of enhancement using (a) Original noisy resident-mic in baseline system; (b) Inclusion of low-passed P-Mic with 4 speech class processing; (c) Addition of of synthetic phase into system in b. The utterance consists of the word stream "choose-keep-bank-got" in the M2H environment.

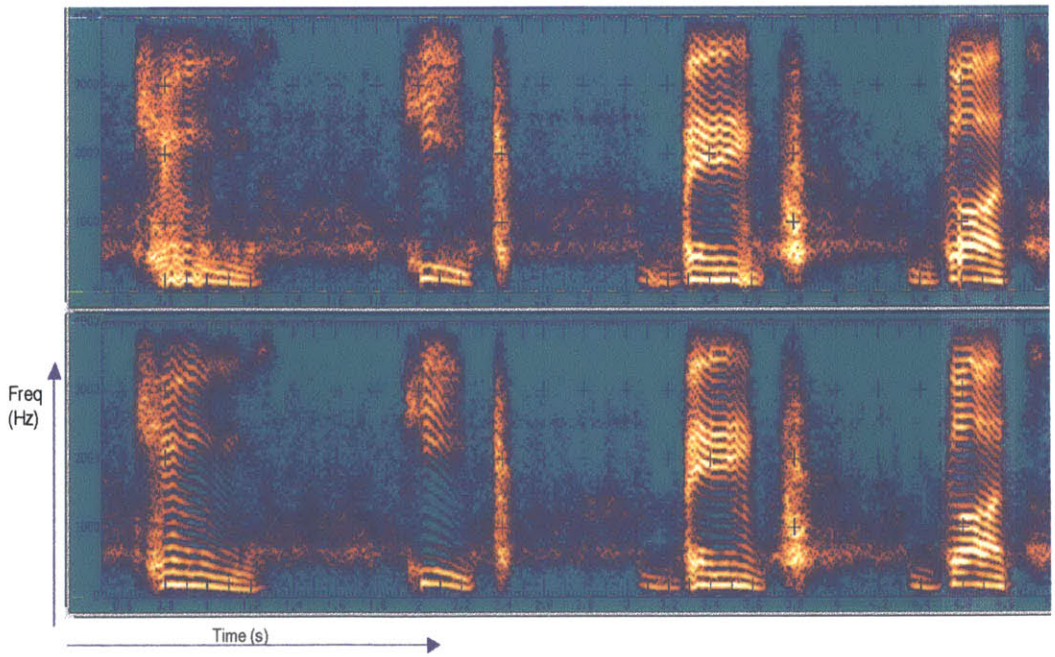


Figure 6.9: Example spectrograms of enhancement using the suppression schemes of figure 6.8 after MELP encoding of signals: (a) Original noisy phase from 6.8b; (b) Introduction of synthetic speech phase in band [500, 1200] Hz as in 6.8c.

6.4 Conclusion

In this chapter, we explored the use of several sensors for phase estimation. We began by examining the affect of using the phase of the GEMS sensor and found that the GEMS phase helped restore the harmonic structure lost in noisy speech. Then a synthetic phase was created from the estimate of the vocal tract and the glottal phase of the vocal folds. Two synthetic phases were developed – one based on an enhanced PMIC magnitude and one based on an enhanced resident microphone magnitude. For this environment (the M2H environment), each phase was used in phase replacement in the 500Hz – 1200Hz band since this band had the lowest SNR after magnitude estimation. For other environments with different conditions, the bands with very low SNR may be different. The PMIC-based synthetic phase was used in the final system developed because, for the signals under study, it seemed to best restore the harmonics in this mid-band while preserving the natural sound of the speech. Informal listening by staff at Lincoln Labs concluded that this PMIC-based synthetic phase produced enhanced speech that was slightly more natural sounding and cleaner than any of the other phase replacement methods. When encoded with a MELPe vocoder, the result using the PMIC-based synthetic phase produced encoded speech that sounded much cleaner than that of any other phase replacement method. Consequently this PMIC-based synthetic phase is used in the final system of this thesis.

The final system developed used the work from all the previous chapters and modified the baseline system in the ways discussed in this chapter and chapters 3 and 5. It was tested on and tailored to the M2H environment since this environment contained the largest amount of real-world noise in the ASE corpus. It also focused on sentences spoken by one male speaker, since this was the only speaker initially available. However, a few tests in this chapter were also run on a few female speaker utterances as well to confirm results. Future work will run tests on other male and female speakers as they become available. As a consequence, the specific frequency band-based adaptations may need to be adjusted for use with other conditions. Future work will involve the generalizations for use in other environments with a larger set of speakers. Specifically, we will focus on generalizations made according to spectral band-based quality measurements.

A diagram of the final system used with the M2H environment is shown below in figure 6.10. This system uses the acoustic resident microphone for processing the high band (above 500 Hz), and the PMIC for the low band (below 500 Hz). As stated previously, phase construction occurs only for the high band between 500 Hz and 1200 Hz since this is the band with severe enough SNR to be improved by phase estimation. During the final stages of design, spectral sharpening, i.e. reducing formant bandwidth, was added to sharpen the PMIC low band since this helped reduce perceived artificial baseness.

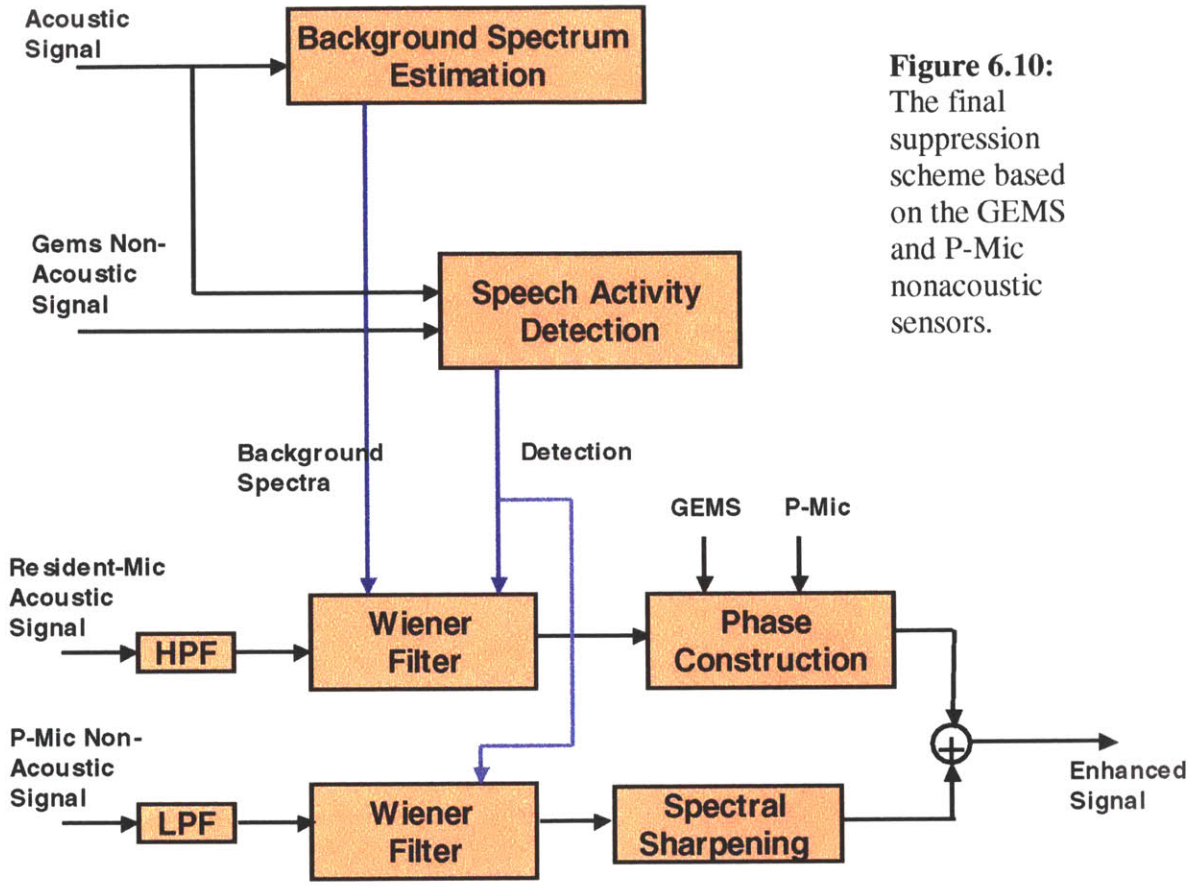


Figure 6.10: The final suppression scheme based on the GEMS and P-Mic nonacoustic sensors.

Chatper 7: Concluding Thoughts

This thesis showed how non-acoustic sensors, such as the GEMS and PMIC, could be used with acoustic noise canceling microphones, such as resident microphones, to improve speech enhancement. Specifically, we used these sensors synergistically to improve detection and spectral magnitude estimation. Also, because of these non-acoustic sensors, we were able to create a phase estimate of the speech segments. When the speech segments were recombined, this phase estimate helped restore the harmonic components of frequency bands with poor SNR. Although this change resulted in a minor perceptual difference when listened to, it yielded a large gain in waveform coding using MELP since the improved harmonic structure resulted in an improved pitch estimate which is essential in the MELP vocoder.

Now that the groundwork has been developed for processing using non-acoustic sensors, there are numerous opportunities for further research. Of immediate interest is Diagnostic Rhyme Testing (DRT) [Voiers, 1977] using several speakers. A DRT is used to measure the ability of humans to distinguish different spoken phones. Consequently it can be used as a measure of intelligibility. This thesis was unable to perform DRT tests due to time and resource constraints. Currently, DRT tests are being planned for use with the final system shown in figure 6.10 and testing should commence within the next few months.

In addition to further system performance tests, there are several ideas that could be explored relating to detection, magnitude estimation, and phase estimation. Also, the general result of the importance of phase brings up several questions. It would be interesting to focus on understanding and modeling phase and its affect on speech. For instance, when processing speech over segments, phase could be very important for maintaining and restoring short temporal events such as plosives, voice onset times between consonants and vowels, or other cues that are used by humans to understand speech.

Current phase estimation could be improved upon using the current sensors as well. For instance, the class-based detection of chapter 3 could be extended to aide in phase replacement. During unvoiced consonants that have no harmonic structure, the original phase could be used to preserve the noise-like shape of the spectrum. During vowels, the entire spectral phase could be replaced with a phase estimate to restore the harmonic structure. In voiced consonants, the phase could be replaced with an estimate in the low band where harmonic voicing is present while the high band with frication or aspiration is left alone.

Another idea involves phase replacement according to band-dependent SNR. Under this scheme, a spectral quality measure would be used to measure the relative SNR and energy of the speech signal in each band. For bands with low SNR or poor speech energy, the phase of the speech could be replaced by some phase estimate. For example, in this work, the resident microphones used in the high noise environments of the ASE corpus have very poor speech energy below 500 Hz. In the mid-bands, the resident microphones had poor SNR. In the high-bands, the resident microphones have high SNR and signal energy. Under this quality-based phase-replacement idea, the low-band and mid-band phase would be replaced with the estimate while the high band is untouched.

This would be similar to the final system of chapter 6; the main difference is that this process would be automated, thus improving robustness since different speakers in different conditions will produce speech with differing characteristics. For example a female will have harmonics that are more largely spaced than a male. Females also have formants that are higher in frequency. Thus an automated phase replacement scheme is very desirable.

Another idea that would be interesting to examine is the affect of using Kalman filtering in speech enhancement. It would be interesting to compare with the Wiener filtering since the Kalman filtering takes place in the time domain and would yield different trade-offs. It could potentially solve many issues with phase estimation, but could perhaps introduce other problems.

Speech activity detection could be improved by a few new schemes as well. For example, detection could be done using several sensors over all bands. Then a quality measure could determine the SNR over a segment of time and band of frequency. This quality measure could then be used to make a weighted spectral detection decision over each band to more robustly conduct detection of speech.

Band detection could also perhaps help spectral magnitude estimation. By dividing the magnitude into bands and creating a separate Wiener filter for each band, more selective processing could be done on the speech signal. This could help reduce noise in bands containing no speech. For example some /m/'s and /n/'s, in the ASE corpus under certain conditions, contain a large amount of low-frequency energy around 400Hz and mid-frequency energy around 2000 Hz, but little in the other bands. Under the system developed in this thesis, such a region would cause a problem for the processor and would likely result in severe loss of speech energy. A finer band-based detector and magnitude estimate would increase robustness and preserve more of the energy in the /m/ and /n/.

Appendix A: Detector Development

This appendix covers previous work that led up to the speech class detector discussed in chapter 3. For more details on the original energy detector, one should consult Dunn and Quatieri [Dunn and Quatieri, 2002].

Work on detection began with a series of experiments that were run to explore the viability of using the GEMS and PMIC sensors in detection. The first set of experiments focused on testing the GEMS sensor by using a corpus provided by Lawrence Livermore National Laboratory. This corpus was used since, at the time, it was all that was available, and since it facilitated the use of Mean-Square-Error (MSE) and Segmental Signal to Noise Ratio (SSNR) measurements. These MSE and SSNR measurements were used as a metric to determine the potential of using the GEMS sensor in detection. The second set of experiments focus on using the GEMS and PMIC sensors in tandem and use the ASE corpus collected by the Arcon Corporation (see chapter 2). This corpus was used since it contains PMIC data as well as GEMS data (and the Lawrence Livermore National Laboratory corpus did not contain PMIC data).

A.1 Metrics Used

This section discusses the quantitative metrics for comparing detectors under different SNR conditions. The two metrics used were Mean-Square-Error (MSE) and Segmental Signal to Noise Ratio (SSNR) measurements. Both of these require the use of the clean speech. The MSE measurement requires the clean speech as a reference to compute the error. The SSNR measurement requires the clean speech to compute the energy of the clean signal in order to find the SNR ratio of clean signal power-to-noise. Since clean speech was only available in the Lawrence Livermore corpus, both of these metrics were only used for tests using this corpus. Experiments involving the ASE corpus required use of comparisons of spectrograms and closer examination of detection outputs.

A.1.1 Mean-Square-Error (MSE)

MSE performance was used as a metric because it is commonly used in signal processing and estimation theory to determine a quantitative measure of the error between an estimate of a signal and the actual signal. The formula we used in computing the MSE of the speech sentences is:

$$\sum_{n}[x_c(n) - x_{est}(n)]^2$$

Here $x_c(n)$ is the normalized clean speech, $x_{est}(n)$ is the normalized estimate of the clean speech (obtained by processing the noisy speech). The disadvantage of using MSE as a metric is it may not be a good indicator of speech quality or intelligibility.

A.1.2 Segmental-Signal-to-Noise Ratio (SSNR)

Since MSE measurements may not be a good indicator of improved speech intelligibility or quality, further evaluations were conducted. These tests focused on using Segmental SNR values to compare the relative increase in signal-to-noise power after being enhanced. The motivation behind this was that this measure has been shown to be more closely related to speech quality [Quackenbush, 1988]. Segmental SNR values were computed according to the following equation:

$$1/M \sum_{m=0}^{M-1} 10\log_{10}\left\{\frac{\sum_{n=N_m}^{n=N_m+N-1} x_c^2(n)}{\sum_{n=N_m}^{n=N_m+N-1} [x_c(n)-x_d(n)]^2}\right\}$$

Here M is the number of speech segments of size N, m is the m th speech segment, N_m is the start index of the m th speech frame $x_c(n)$ is the uncorrupted speech, and $x_d(n)$ is the processed distorted speech (i.e. speech with noise after it is processed). Intuitively, the Segmental SNR is equal to the average SNR of each segment of speech.

A.2 GEMS Dection

This section focuses on the use of the GEMS sensor in detection and compares it with the acoustic detection used in the baseline algorithm. It examines the performance of the GEMS alone and in tandem with the acoustic speech by using the MSE and SSNR measurements described above in section A.1.

A.2.1 Inputs Used

To conduct initial tests, the Lawrence Livermore corpus was used. This corpus consists of two parts: speech data and GEMS data recorded simultaneously. For these tests, 6 sentences were used. All 6 sentences were noise-free. Consequently noise had to be added for testing purposes. Colored noise was added to the acoustic recordings by using an algorithm that modeled the noise after car engine noise (see chapter 2). After noise was added, experiments were conducted on the 6 spoken sentences for two speakers, yielding a total of 12 spoken sentences.

A.2.2 Experiments

MSE Measurements:

The first MSE experiments conducted involved a baseline with which to make comparisons. As a reference case, the baseline Lincoln speech enhancement was used. In that enhancement routine, a detector decision based on acoustically-recorded speech is used. We refer to this detector as LLdetectorA. Then we sought to show an increased speech enhancement performance by using the GEMS waveform as input to a speech detector that is better tailored to use the information present in the GEMS signal. This detector is referred to as LLDetectorB.

LLDetectorB processes a waveform in an auditory-like manner by sending the waveform through a filter bank of bandpass filters with logarithmic increasing bandwidth. The energy of the outputs of the lower-frequency bandpass filters is then computed and summed to create a metric that is used to determine if speech is present. Only the lower 8 bandpass filters were used in this energy computation for several reasons, one being that most of the energy of the GEMS waveform is concentrated in these lower frequencies.

Tests were run to compare LLDetectorB's ability to detect speech using the GEMS data against the LLDetectorA's ability to detect speech using acoustic input. The purpose was to show the overall improvement of using the GEMS data as input to a detector tailored to the GEMS input waveform versus using acoustic speech data and a detector that was tailored to the speech data. The resulting computed average MSE for the 12 sentences in the corpus are shown below in figure A.1. As one can see, for most SNR values, the LLDetectorB with the GEMS input performed better in the mean-squared-error sense than the LLDetectorA with the speech input - the exception being -10db. Also of importance, the output of the detector was more stable (i.e. more contiguous within speech and background regions) using the GEMS-based detector and GEMS input (LLDetectorB) than that using the acoustic based detector with speech input (LLDetectorA). The -10db sentences were atypical in that the processed versions seemed to have a bit more distortion than the other processed sentences of different SNR. This distortion is likely caused by the experimental noise reduction algorithm and likely explains the MSE result at -10db. At the very negative SNRs (-15db and -20 db), the percent MSE performance gain dwindled. This could be due to the fact that the input signal is significantly corrupted by noise at extremely negative SNRs. Consequently, the noisy input could become completely uncorrelated with the actual speech input if the noise is large enough. If this is the case, improved speech detection would yield very little to no improvement in speech enhancement. (In fact it could be worse in the MSE sense). Indeed, at -10db, one can barely hear the spoken sentence and at -15db and -20 db there is so much noise that the spoken sentence is completely masked by noise.

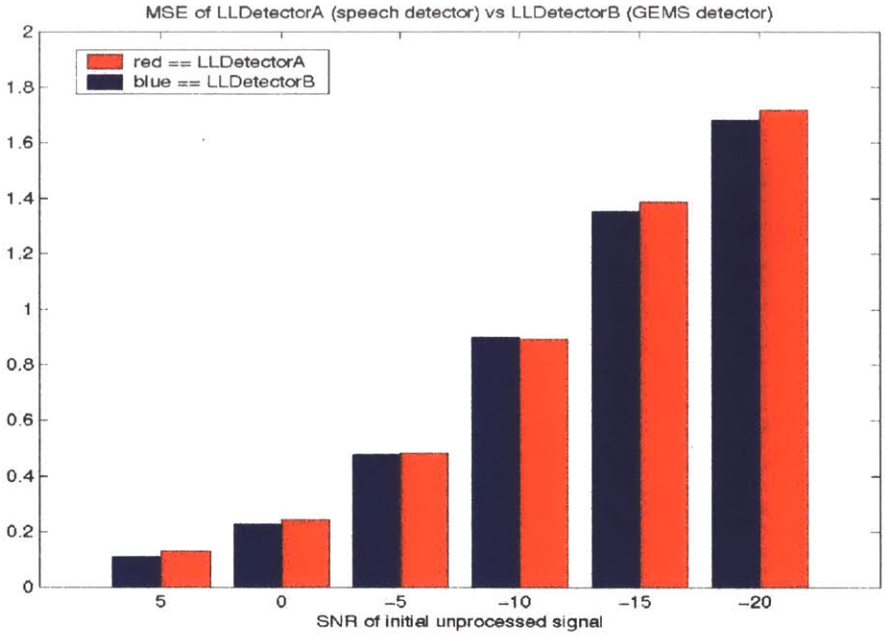


Figure A.1: MSE of LLDetectorB using GEMS data as input vs LLDetectorA using speech

Since it is unclear as to whether or not the GEMS detector can detect unvoiced speech and since a fair amount of the intelligibility of speech is due to unvoiced speech, it was decided to design a detector based on both speech and GEMS data. This detector shall be called the Multidetector. It is depicted below, in figure 2, and is composed of four parts: LLDetectorA, LLDetectorB, Part Selector of A, and an OR gate. LLDetectorA is used to make speech detection decisions based on recorded speech data (just as in previous experiments). LLDetectorB is used to make speech detection decisions based on GEMS data (just as in previous experiments). The output of LLDetectorB consequently denotes the detected voiced regions of speech. When this output is a 1, the corresponding speech segment is marked as containing speech. When it is 0, it is not. Part Selector of A is used to detect unvoiced regions of speech based on the outputs of LLDetectorA and LLDetectorB. It does this by looking 100 ms on both sides of the regions that are identified as voiced regions by LLDetectorB and using LLDetectorA to detect the unvoiced speech in these neighboring regions. More specifically, it convolves the output of LLDetectorB with a filter h(n) that is 100 seconds in length on both sides of the time origin. Then it takes the output of this convolution and "AND"s it with the inverse of the output of LLDetectorB, thus obtaining the neighborhood regions where unvoiced speech may be present. It then takes this result and "AND"s it with the output of LLDetectorA, thus obtaining the detected unvoiced regions of speech 100ms on either side of a region of detected voiced speech. The final

detected unvoiced speech regions result is then "OR"ed with the detected voiced speech regions to yield the output detected speech regions of the Multidetector (see figure A.2 below).

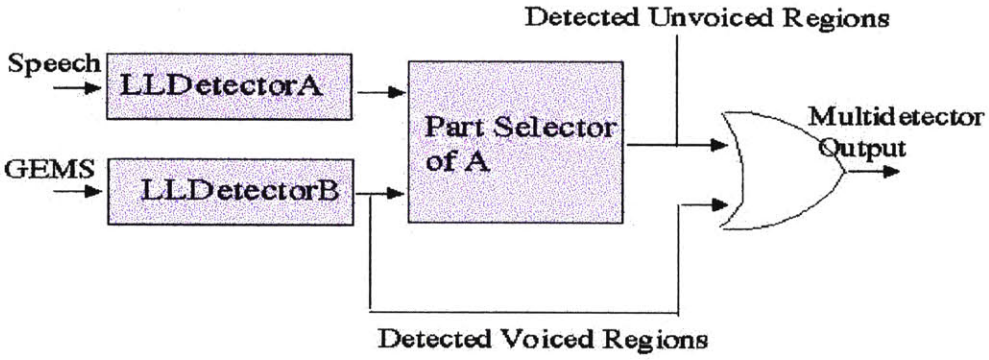


Figure A.2: Multidetector

The rationale behind this approach is due to the fact that unvoiced speech is usually contiguous to voiced speech. The window of 100 ms on both sides of the detected region is somewhat arbitrary and was chosen since it yielded the best MSE results of all windows that were explored. In reality, some regions of unvoiced speech last longer than 100 ms. This was apparent in some of the sentences that contained long unvoiced phonemes such as some /s/'s. However, this way of selecting the window was the best we could hope for since the GEMS data does not seem to be able to distinguish between unvoiced and non-speech speech regions. To explore the reliability of the Multidetector, tests using the Multidetector on all 12 corpus sentences were conducted and the resulting MSEs were computed and then compared to both the LLDetectorB using GEMS input and the LLDetectorA using acoustic speech input. The results are shown below in figures A.3 and A.4.

According to these results, the Multidetector out-performed the LLDetectorA using acoustic speech as input in all 6 different SNRs that were tested -- except the perhaps unreliable -10db case (see discussion above). At the highest 5db SNR, the Multidetector performed a little better, but at all other SNRs, the Multidetector performed a little worse in the MSE sense. It was thought that the better performance at high SNRs could be attributed to the fact that the Multidetector detects voiced speech reliably well at high SNRs (with low levels of noise) where as the LLDetectorB (GEMS decision) does not. The worse performance at the more negative SNRs (with higher levels of noise) of the Multidetector could be attributed to the fact that there is so much noise present that the Multidetector miss-detects unvoiced speech very frequently. If this is the case, the Multidetector would be expected to perform slightly worse than the LLDetectorB. However, another possibility is that the GEMS data does actually contain some of the unvoiced information of speech. If this were the case, it would not be at all surprising to have the GEMS-based detector perform better at lower SNRs. Because of this, the results of the MSE tests are a bit inconclusive. The only certain conclusion that can be made is that the GEMS-based detector yielded improved performance over the more traditional acoustic-based detector in a MSE sense.

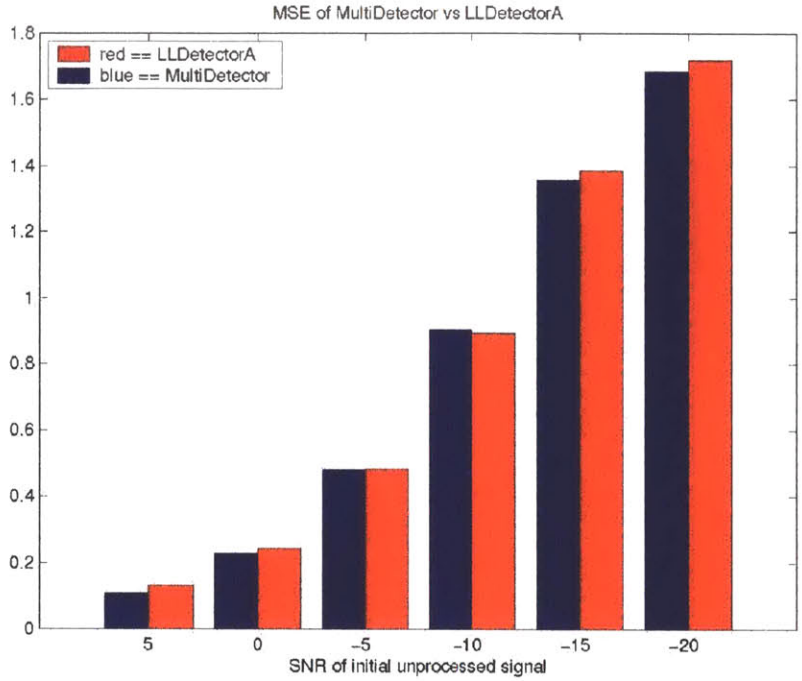


Figure A.3: Average MSE of Multidetector (red) vs LLDetectorA (blue) (5db to -20db left to right)

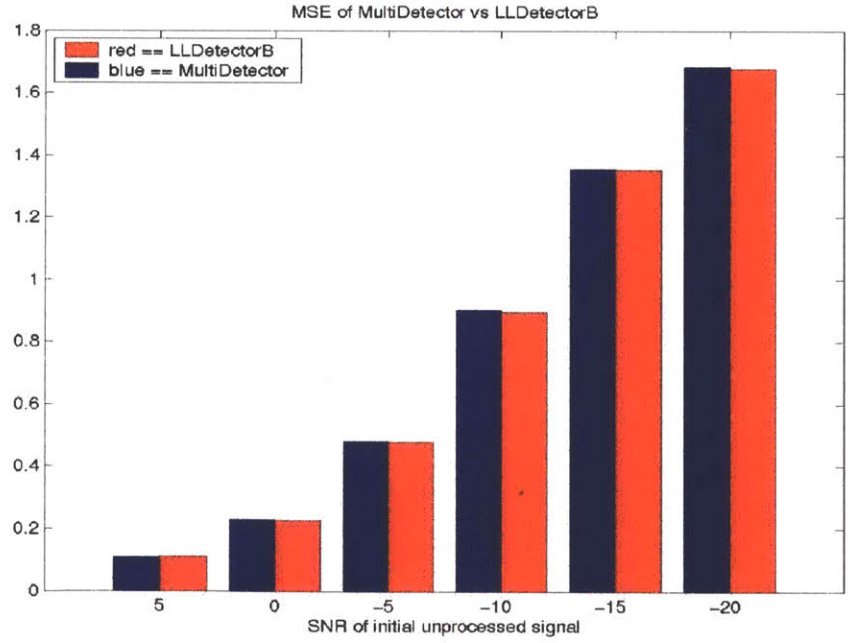


Figure A.4: Average MSE of LLDetectorB (red) vs Multidetector (blue) (5db to -20db left to right)

SSNR Measurements:

The SSNR experiments were conducted on the same waveform files as the MSE experiments. Segmental SNR values were computed for LLDetectorA using speech as input, LLDetectorB using GEMS data as input, and the Multidetector. A comparison of the results is depicted in figures A.5, A.6, and A.7 below. As was the case with the MSE tests, the Multidetector and LLDetectorB with the GEMS input showed an improvement in performance over the more traditional LLDetectorA using

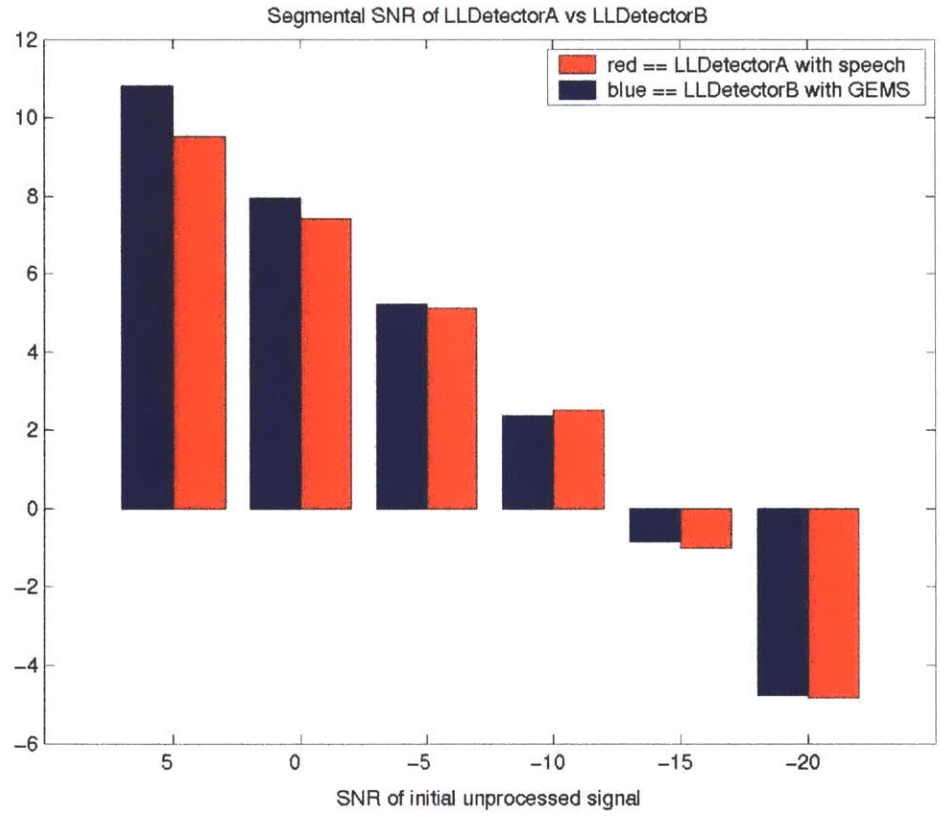


Figure A.5: Average Segmental SNR of LLDetectorA (red) vs LLDetectorB (blue) (5db to -20db left to right)

speech for detection - once again the exception was the -10db case (see discussion above). The difference between the Multidetector and LLDetectorB was small; the Multidetector performed a bit better than LLDetectorB at the high 5db SNR value but a bit worse at the other lower SNRs. This seemed to mimic the results of the MSE tests, and is likely due to the same causes (see above for discussion).

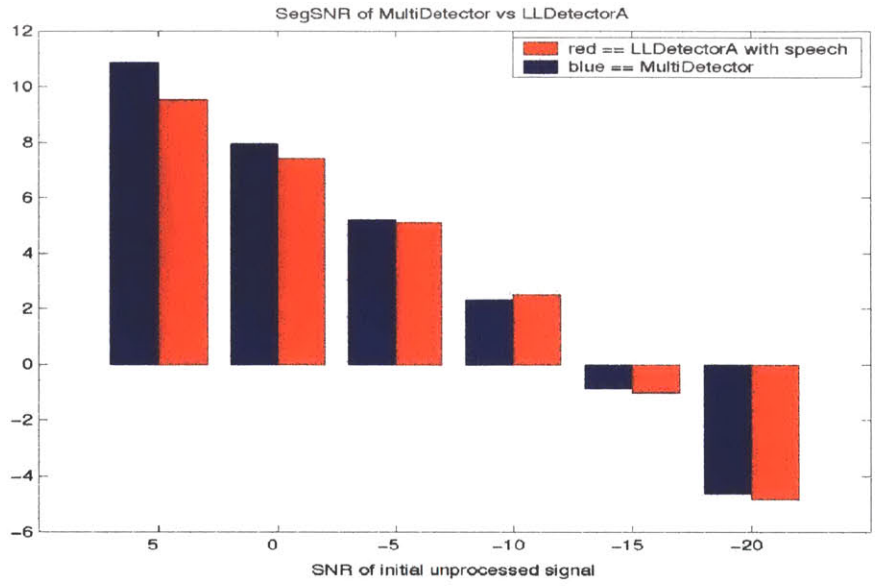


Figure A.6: Average Segmental SNR of LLDetectorA (red) vs Multidetector (blue) (5db to -20db left to right)

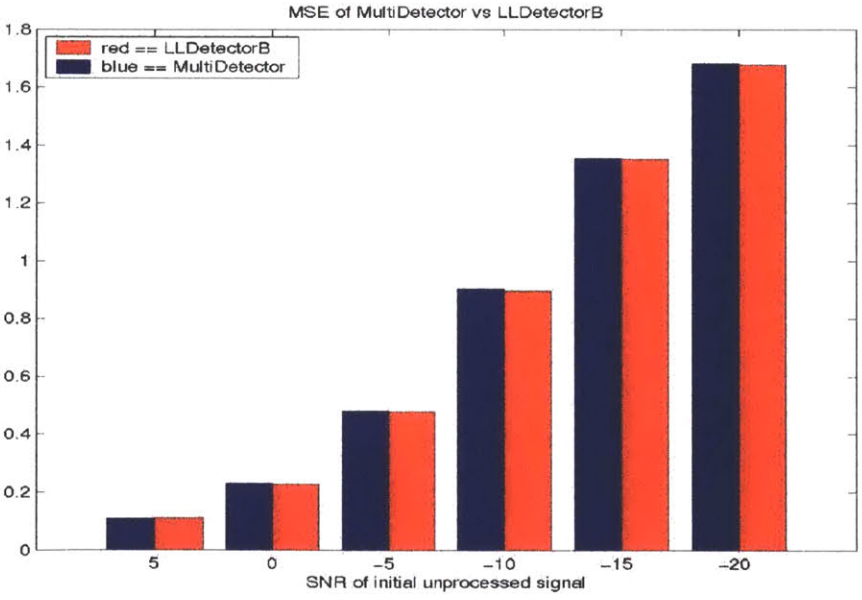


Figure A.7: Average Segmental SNR of LLDetectorB (red) vs Multidetector (blue) (5db to -20db left to right)

A.2.3 Conclusion of GEMS Detection

According to the MSE and Segmental SNR data, a GEMS based algorithm is an improvement over previous more conventional acoustical speech detection systems such as that used in the baseline system described in chapter 2. That is to say, it seems to be better identifying regions of speech than the solely acoustic-based detector used in the baseline Lincoln enhancement algorithm. According to the MSE tests and Segmental SNR tests, the Multidetector performs a little better than the LLDetectorB at 5db and a little worse at all other lower SNRs. The better performance at high SNRs could be attributed to the Multidetector detecting all speech well at high SNRs where as the LLDetectorB does not; or it could mean that the GEMS data actually contains unvoiced information. The worse performance at very low SNRs of the Multidetector could be attributed to the fact that there is so much noise present that the Multidetector miss-detects unvoiced speech very frequently. If this is the case, the Multidetector would be expected to perform slightly worse than the LLDetectorB at these severe SNRs.

In addition to quantitative MSE and Segmental SNR measurements described above, the overall improvement observed using the GEMS data is further supported by more subjective tests. These subjective tests involved Lincoln Laboratory personnel listening to the original noisy speech and then comparing it to both the enhanced speech that used the old acoustic speech detector and the enhanced speech that used the new Multidetector. In general, both agreed that the enhanced speech that used the Multidetector sounded a bit less "static-like" than the enhanced speech using the original acoustic-based detector. Also, it appeared that the enhanced speech with the greatest MSE improvement (the ones produced by the Multidetector) had the most noticeable static and popping sound reduction.

A.3 PMIC and GEMS Detection

Because the Multidetector (used above) showed some promise, experiments were conducted using other sensors to form a decision of where the voiced regions of speech are present. Specifically, data from a PMIC detector placed in position at the throat was used since it seemed to be partly immune to acoustic noise while still containing unvoiced speech information for certain speakers and locations. To conduct experiments with the PMIC and GEMS, a different corpus was used since PMIC data was not part of the Lawrence Livermore corpus. The corpus we used was the ASE corpsu.

A.3.1 Inputs Used

For tests involving the GEMS and PMIC in detection, experiments focused on full spoken sentences that were used in the collection of the TIMIT database (see chapter 2

for further details). Since this corpus contained no "clean", it was impossible to compute MSE and SSNR values. Consequently other methods of comparison were employed.

A.3.2 Experiments

Since it was not possible to calculate MSE and SSNR values to judge the performance of the PMIC, we were forced to examine the actual detection decisions more carefully and compare them to the recorded waveforms. Thus this was a much more subjective measure of performance, however it was quite insightful. To test the PMIC in severe noise environments, we conducted our experiments using the M2H tank recordings since these had the largest noise level (114 db Sound Pressure Level) and thus were the most relevant to our research.

To begin, detector decisions were obtained using the Lincoln Laboratory algorithm (as used in the Lawrence Livermore corpus experiments above). However, a different detector developed by Licoln Laboratory was used. This detector provided more efficient and accurate energy computation and speech activity decisions. Thus it was used to compute detection decisions based upon the GEMS, PMIC, and acoustic noisy speech recordings.

The first experiment compared the GEMS-based and noisy acoustic speech-based detection decisions. Figure A.8 contains an example output for the recording "Ten pins were set in order. The bill was paid every third week." As expected, at such a severe noise level as the M2H tank environment, the noisy acoustic speech recording produced a detection decision that was wildly inaccurate: it mislabeled large amounts of background noise with no speech, failed to detect many segments of speech, and was very unstable (i.e., the detection decision flipped rapidly from 1, 1 meaning speech detected, to 0, 0 meaning no speech detected, in contiguous regions). Thus it was clear the noisy acoustic recording was of little use for detection. The GEMS-based detector decision, was very stable (did not rapidly flip from 1 to 0), and typically did not mislabel background noise as speech. However, the GEMS-based detector was unable to correctly identify all the regions of speech. Specifically, in figure A.8, it was unable to identify the longer-duration unvoiced regions of speech highlighted by the red box. The first such region corresponds to the /t/ in "ten"; the second such region, to the /p/ in "paid." Indeed, if one examines figure 8, one will note that these unvoiced speech regions do not or barely appear in the time-domain GEMS waveform. Thus it was concluded that the GEMS recordings can be used to yield detection of voiced speech, however can not be used to provide accurate detection of unvoiced speech.

The second experiment compared the PMIC-based and noisy acoustic speech-based detection decisions. Figure A.8 also contains an example of these decisions on the same recording ("Ten pins were set in order. The bill was paid every third week"). From the figure, one can see that the PMIC based detector was able to detect the /t/ and /p/ unvoiced sounds that the GEMS based detector was unable to detect. Also one can easily see that the PMIC detection decision is more stable than the noisy acoustic speech-based detection decision. When listening to the actual sentence or comparing the PMIC

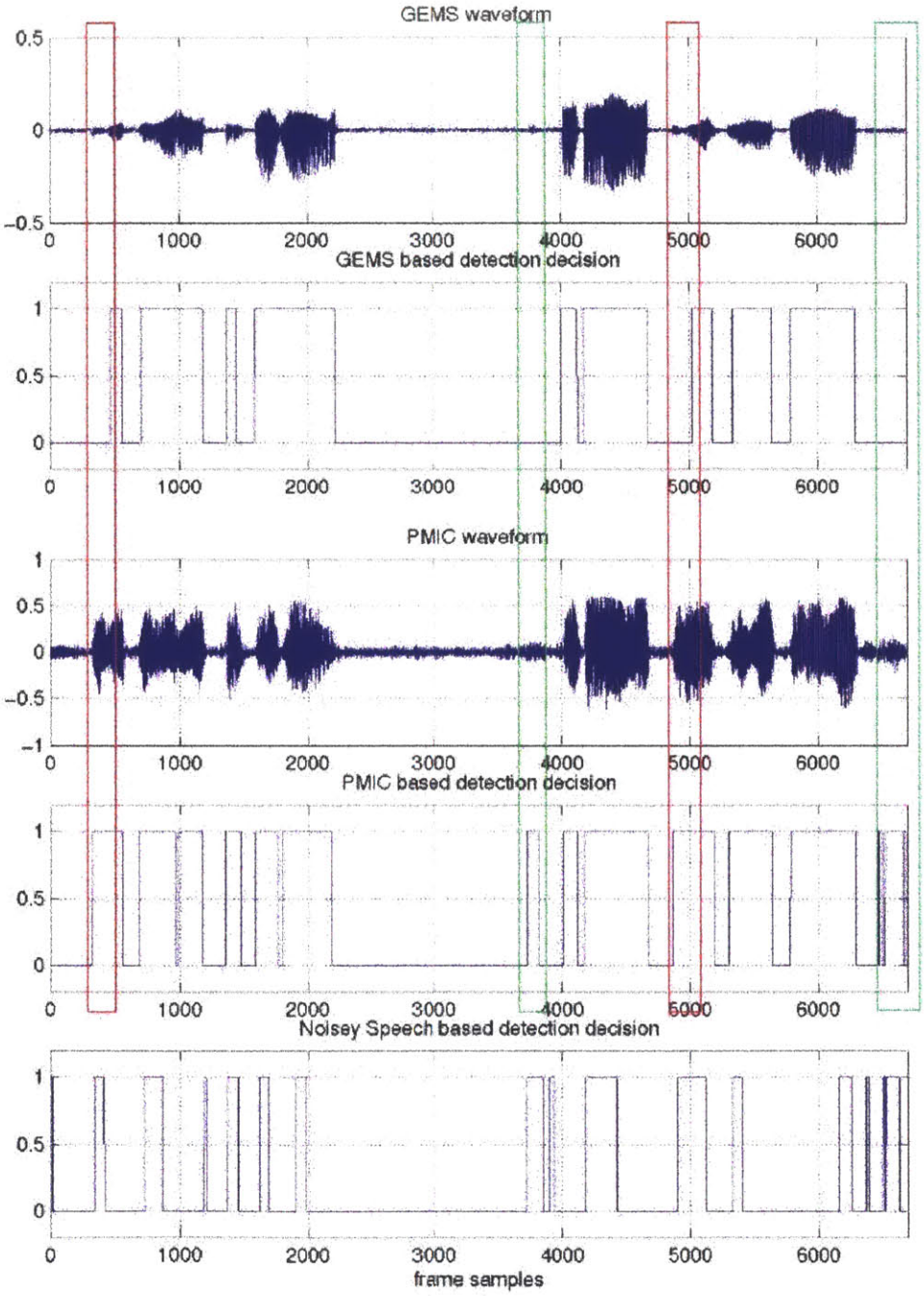


Figure A.8: GEMS and PMIC based detection
(Red box = unvoiced speech undetected by GEMS but detected by PMIC)
(Green box = tank tread noise misdetected by PMIC and B&K mic based detectors)

detection decision to the GEMS decision, one can also see that the PMIC decision is much more accurate than the noisy acoustic speech based decision. However, if one examines the PMIC decision more closely, one will find that it also misdetects some background noise regions containing no speech. The green boxes in figure 8 are good examples of this. In this figure, both green boxes coincide with higher level noise caused by a tank tread clinking sound in the M2H environment. The PMIC detector and B&K acoustic detector both are affected by these tread sounds and consequently these sounds are present in the data recordings of both sensors. The GEMS sensor however, is immune to such a noise since it operates by measuring electromagnetic wave propagation instead of measuring vibrational motions like the PMIC and B&K acoustic mic sensors. Here the B&K directly is affected by the loud tread noise since it measures vibrations through air. The PMIC is affected by the loud tread noise, because it is not completely immune to all vibrations that travel through the air.

Because the GEMS detector was shown to be an excellent detector of voiced speech and an unreliable detector of unvoiced speech and because the PMIC was shown to be fairly reliable in detecting unvoiced speech, we sought to fuse the detection decisions of the two detectors in a way that would synergistically improve the overall detection decision. To do this we used the same concepts and layout as used in the Multidetector used in the GEMS experiments. A diagram of the components of the speech detection system is shown below (see figure 9). The speech detection system consists of 3 parts: the PMIC Detector, the GEMS Detector, and the Detection Fuser. The PMIC Detector takes sampled noisy speech or a PMIC waveform as an input and produces a 1kHz sampled output that is 1 if speech is detected over a speech frame (that was 10 samples in duration) and is 0 if speech is not detected. The GEMS Detector takes in a sampled GEMS signal as input and produces a 1 kHz sampled output that is 1 if speech is detected over a speech frame and 0 otherwise. The Detection Fuser fuses the decisions of the two detectors in a way that takes advantage of the more accurate GEMS voiced decision to determine the unvoiced regions that were flagged by the PMIC speech detector (see below).

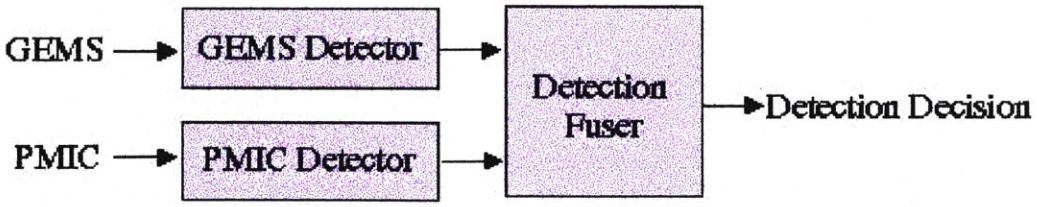


Figure A.9: Detector with GEMS and PMIC decision Fusion

The Detection Fuser creates a final speech detection decision based on the PMIC Detector and GEMS Detector decisions. Since the GEMS Detector uses the GEMS

signal as input, it is only able to accurately detect voiced speech. The PMIC Detector however is able to detect both voiced and unvoiced speech, though it detects voiced speech less reliably since the PMIC sensor is not as immune to acoustic noise as the GEMS sensor. Consequently, the Detection Fuser uses the GEMS Detector to identify voiced speech segments and then uses the PMIC Detector to identify unvoiced segments of speech 100ms on either side of the identified voiced regions (100 ms was chosen since most unvoiced speech sounds are less than 100ms in duration). Since unvoiced speech is only present before or after voiced speech, this scheme reduces detection errors in long periods of "silence" when speech is not present over many segments. The other advantage of this scheme is that it utilizes the more robust GEMS signal and the less selective PMIC signal synergistically, thus achieving what neither could alone.

The output of this detector on the recorded sentences "There are more than two factors here", "The hat brim was wide and too goofy", and "The lawyer tried to loose his case" is shown in figure A.9. Unlike the GEMS based decision, the fused decision properly detects the unvoiced /t/ in "two" (see first red box if figure A.9 below) and trailing /s/ in "factors"(see second red box in figure A.9). The GEMS-based detector found more of the tails of some of the words and also some of the residual voicing of vowels. Also of interest to note, the GEMS based detector failed on two instances to detect some voiced and unvoiced speech when the PMIC and resident microphone based detectors were able to do so (see green boxes in figure A.9 below). When examining the sections of speech, it was found that both of these regions corresponded to very loud tread noise spikes. Consequently this suggests that the GEMS sensor is not entirely immune to very loud and sudden noises. This could possibly be because the large sound level during a spike creates such a large sound pressure wave that this wave moves the GEMS device significantly enough to affect the sensor measurements. Although the GEMS sensor and detection decision was detrimentally affected by this tread spike, when combined with the PMIC based detection decision, the fused detector properly detected most of the two regions of speech following the tread spike onsets. Consequently, it was determined that the fused PMIC and GEMS detection scheme was superior to previous schemes explored.

A.3.3 PMIC and GEMS Detection Conclusion

The experiments with the ASE corpus for the M2H environment showed that the PMIC sensor could be valuable in detection when the PMIC provided unvoiced information. Furthermore, when combined with the GEMS sensor, the two sensors were shown to be capable of accurate and robust fused detection. Consequently this fused detection scheme was used until class based detection of chapter 3 was developed.

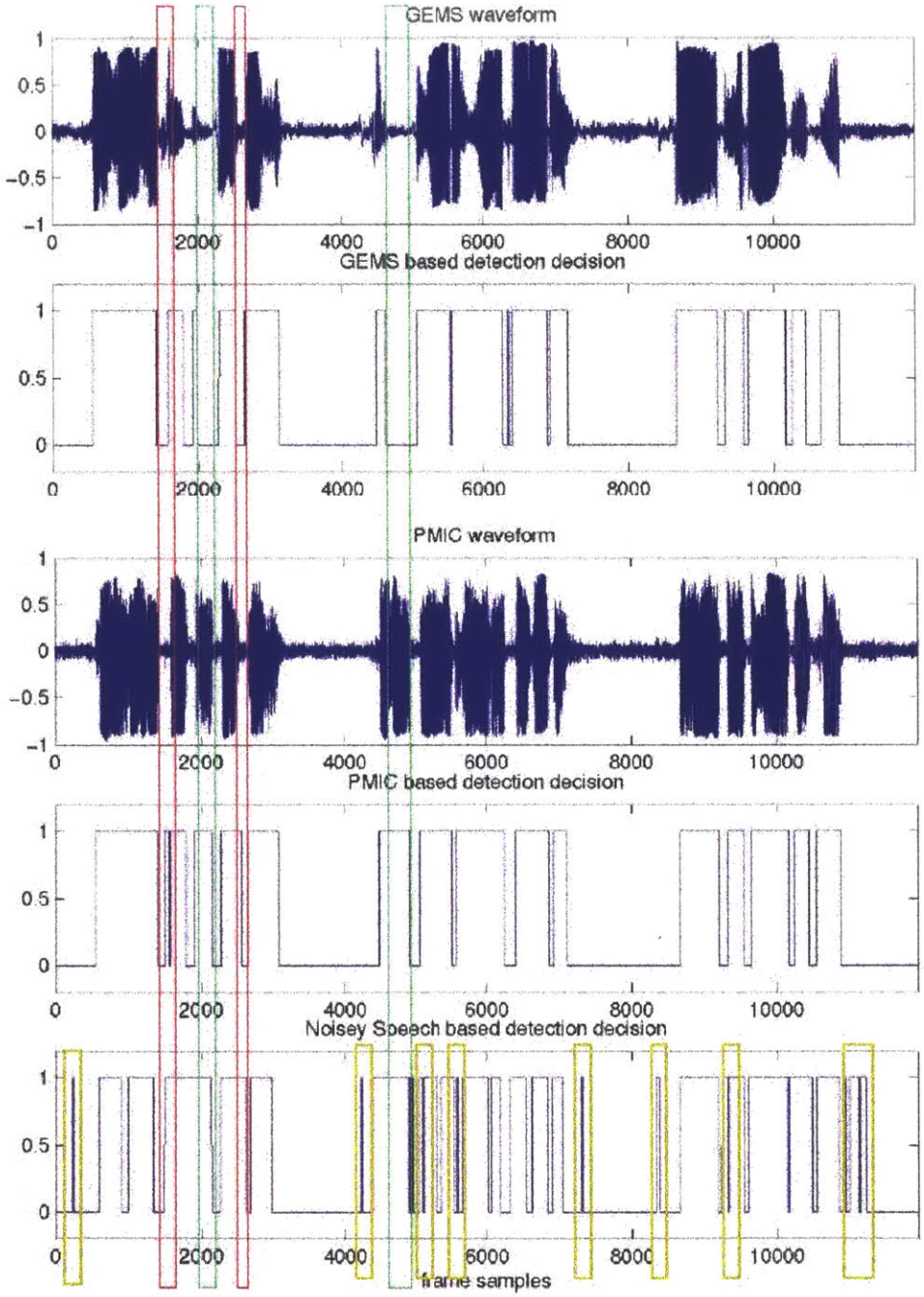


Figure A.9:GEMS and PMIC based detection
(Red box = unvoiced speech undetected by GEMS but detected by PMIC)
(Green box = tank tread noise causing GEMS based detectors to miss speech)
(Yellow boxes = detection mistakes using resident mic in detection)

A.4 Detection Conclusion

In this Appendix we discussed all of the work preceding the development of the multi-sensor speech class detector that is illustrated in chapter 3. In this work, we showed that GEMS-based detection can yield an improvement for voiced speech detection in severe noise environments. We also showed that fusing the GEMS detector decision with the acoustic detector decision yielded improved detection of unvoiced consonants. Finally, we also explored the use of the PMIC and found that when fused with the GEMS, the PMIC can create robust detection of voiced and unvoiced speech that is reliable and stable, provided the PMIC is located to give unvoiced speech information. Since not all speakers had PMICs located to give unvoiced speech information, this requirement led to the further detector development discussed in chapter 3.

[Ng et al, 2000] L.C. Ng, G.C. Burnett, J.F. Holzrichter, and T.J. Gable, "Denoising of human speech using combined acoustic and EM sensor signal processing," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Istanbul, Turkey, June 2000.

[NRC, 1989] "Removal of noise from noise-degraded speech," Panel on removal of noise from speech/noise signal, National Academy Press, Washington, D.C, 1989.

[Oppenhiem and Schafer, 1999] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing*, 2nd ed. Upper Saddle River, N.J.: Prentice Hall, 1999.

[Pinter, 1996] I. Pinter, ``Perceptual wavelet-representation of speech signals and its application to speech enhancement", {Computer Speech and Language}, vol. 10, pp. 1--22, 1996.

[Quatieri, 2002] T.F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 2002.

[Quatieri and Baxter, 1997] T.F. Quatieri and R.A. Baxter, "Noise reduction based on spectral change," *IEEE 1997 Workshop on Appl. of Signal Processing to Audio and Acoustics*, pp. 8.2.1-8.2.4, New Paltz, NY, October 1997.

[Quatieri and Dunn, 2002] T.F. Quatieri and R.B. Dunn, "Noise reduction based on auditory spectral change," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing,* Orlando, FL, May 2002.

[Quackenbush, 1988] Quackenbush, S., Barnwell III, T., and Clements, M., (1988). "Objective Measures of Speech Quality. pp.41-47

[Rhodes, 1971] I. Rhodes, "A tutorial Introduction to Estimation and Filtering," IEEE Transactions on Automatic Control, VOl 16, No.6, Dec 1971, pp. 688-706

[Rothenberg, 1992] M. Rothenberg, "A multichannel electroglottograph," *J. of Voice*, vol. 6, no. 1, pp. 36-43, 1992.

[Scanlon, 1998] M.V. Scanlon, "Acoustic sensor for health status monitoring," *Proceedings of IRIS Acoustic and Seismic Sensing*, vol. 2, pp. 205-222, 1998.

[Sohn, Kim, Sung, 1999] J. Sohn and N.S. Kim and W. Sung, "A statistical model-based voice activity deteciont," in IEEE Signal Processing Letters, vol 6, pp. 1-3, 1999

[Vary, 1985] P. Vary, "Noise suppression by spectral magnitudc estimation - mechanism and theoretical limits," *Signal Processing*, vol. 8, pp. 387-400, 1985.

[Voiers, 1977] W. D. Voiers, "Diagnostic Evaluation of Speech Intelligibility," Benchmark Papers in Acoustics, Vol. 11: Speech Intelligibility and Speaker Recognition (M. Hawlet ed.) Dowden, Hutchinson and Ross, Stoudsburg (1977).

[Wang and Lim, 1982] David L. Wang and Jae S. Lim, "The Unimportance of Phase in Speech Enhancement," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol, ASSP-30, No.4, August 1982

[Wang, Koishida, Cuperman, Gersho and Colhura, 2002] T. Wang, K. Koishida, V. Cuperman, A. Gersho and J. Colhura, "A 1200/2400 BPS Coding Suite Based on MELP", 2002 IEEE Workshop on Speech Coding, 2002

References

[Burnett et al, 1999] G.C. Burnett, J.F. Holzrichter, T.J. Gable, and L.C. Ng, "The Use of Glottal Electromagnetic Micropower Sensors (GEMS) in Determining a Voiced Excitation Function," presented at the 138th Meeting of the Acoustical Society of America, November 2, 1999, Columbus, Ohio.

[Donahue, 1994] D. Donahue and I. Johnson, ``Ideal denoising in an orthonormal basis chosen from a library of bases", {C.R. Academy of Science}, Paris, vol. 1, no. 319, pp. 1317--1322, 1994.

[Donahue and Johnson, 1994] D. Donahue and I. Johnson, "Ideal denoising in an orthonormal basis chosen from a library of bases," C.R. Academy of Science}, Paris, vol. 1, no. 319, pp. 1317-1322, 1994.

[Ephraim and Malah, 1984] Y. Ephraim and D. Malah, ``Speech enhancement using a minimum mean-square error short-time amplitude estimator", { IEEE Trans. Acoustics, Speech, and Signal Processing}, vol.ASSP-32, no. 6, pp. 1109--1121, December 1984.

[Fang and Atlas, 1995] J. Fang and L.E. Atlas, ``Quadratic detectors for energy estimation", {IEEE Trans. Signal Processing}, vol. 43, no. 11, pp. 2582--2594, November 1995.

[Haigh and Mason, 1993] J.A. Haigh and J.S. Mason, "A voice activity detector based on cepstral analysis", Proc. European Conf. Speech Communication and Technology, 1993, Vol 2, pp. 1103-1106

[Hansen and Nandkumar, 1995] J.H. Hansen and S. Nandkumar, "Robust Estimation of Speech in Noisy Backgrounds Based on Aspects of the Auditory Process," JASA, Vol. 97, No. 6, June 1995.

[Jabloun and Cetin, 1999] F. Jabloun and A.E. Cetin, ``The Teager energy based feature parameters for robust speech recognition in noise",{Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing}, vol. 1, pp. 273--276, Pheonix, AZ, March 1999.

[Knagenhjelm and Klein, 1995] H.P. Knagenhjelm and W. B. Klein, ``Spectral dynamics is more important than spectral distortion", {Proc. IEEE Int. Conf, Acoust., Speech, and Signal Proc.}, May 1995

[Lim and Oppenheim, 1979] J.S. Lim and A.V. Oppenheim, ``Enhancement and bandwidth compression of noisy speech," { Proc. of the IEEE}, vol. 67, no. 12, pp. 1588-1604, December 1979.

[McCree et al, 1996] A. McCree, K. Truong, E.B. George, T.P. Barnwell, and V, Viswanathan, "A 2.4 kbit/s MELP coder candidate for the new US Federal standard," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Atlanta, GA, vol. 1, pp. 200-203, May 1996.

[Moore, 1988] B.C.J. Moore, {An introduction to the psychology of hearing}, New York: Academic Press, pp. 191, 1988.

[Natarajan, 1995] B.K. Natarajan, ``Filtering random noise from deterministic signals via data compression", {IEEE Trans. Signal Processing}, vol. 43, no. 11, pp. 2595--2605, November 1995.