

A Microphone Array System for Speech Source Localization, Denoising, and Dereverberation

A thesis presented

by

Scott Matthew Griebel

to

The Division of Engineering and Applied Sciences
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in the subject of

Engineering Sciences

Harvard University
Cambridge, Massachusetts

April 2002



Thesis advisor
Michael S. Brandstein

Author
Scott M. Griebel

A Microphone Array System for Speech Source Localization, Denoising, and Dereverberation

Abstract

There is a great deal of potential for advancement in distant-talker speech acquisition research, and a wealth of current and future technology depends upon these advances. The goal of this work is to allow users the opportunity to roam unfettered in diverse environments while still providing a high quality speech signal and a robustness to background noise and reverberation effects. In this thesis, a microphone array speech enhancement system is presented which has three main components: source localization, background noise reduction, and dereverberation. The localization algorithm is effective in the presence of both background noise and reverberations and simultaneously produces relative time delay estimates and a source location estimate. It provides a procedure applicable to all time delay estimators which either maximize or minimize an appropriate objective function, improving the estimators' robustness to environmental degradations. The denoising algorithm is a multi-microphone extension to the Minimum Statistics denoising technique [Martin (2001)]. This algorithm also has an additional and optional SNR-dependent beamforming stage that is shown to be very useful in certain environments. The final component is a multi-channel dereverberation algorithm which models the speech source and room reverberations independently. A weighting function is estimated and applied in the Wavelet Transform domain to de-emphasize portions which are less coherent across microphone signals, an indication of reverberation effects. Results for the various components are provided as proof of the effectiveness of the proposed multi-microphone speech enhancement system.

Acknowledgements

I would like to express my gratitude for the advising and friendship of Prof. Michael Brandstein whose guidance in matters both personal and professional was always helpful. I also thank my thesis committee members Prof. Roger Brockett and Prof. Irvin Schick for their input over the years.

I would like to acknowledge my partners in crime at the Harvard Multi-Media Lab: Daniel Freedman (the smartest man I know), Ce Wang (who would die for his dream), and Daniel Kirsanov (who was always with basketball, but never played basketball). Those outside the lab that were equally appreciated include Soundouss Bouhia, Anthony Volpe, Winston Yu, and Brendan Zinn. They all sparked numerous pointless conversations that were often needed to break up a day's work.

I also thank Homer J. Simpson for his weekly words of wisdom. His knowledge regarding nuclear engineering proved helpful even in my unrelated field of speech signal processing.

I would also like to express my appreciation for my parents for their caring and support throughout my graduate career.

Finally, I thank my wonderful wife Kristen (the smartest woman I know). She waited patiently for me to begin providing some sort of real salary, but always understood that my quest for “higher knowledge” was, to me at least, well worth the trip.

If there's a way to do it better... find it.

Thomas A. Edison

Every time I learn something new, it pushes out something old.

Homer J. Simpson

Contents

1	Problem Overview and Motivation	1
1.1	Source Localization and Time Delay Estimation	2
1.2	Background Noise Reduction	4
1.3	Speech Dereverberation	5
1.4	Thesis Organization	6
2	Speech Production and Processing Background	7
2.1	The Speech Production Model	7
2.2	Speech Signal Processing Tools	12
2.2.1	The Fourier Transform	12
2.2.2	The Short-Time Fourier Transform	13
2.2.3	The Wavelet Transform	14
2.3	Reverberant Channel Modeling	29
2.4	Quantitative Speech Quality Measures	31
2.4.1	Segmental Signal-to-Noise Ratio	32
2.4.2	Bark Spectral Distortion	33
3	Robust Multi-Channel Source Localization	40
3.1	Traditional TDE-based Localization	43
3.1.1	The Mathematics of Time Delay Estimation	43
3.1.2	Generalized Cross-Correlation Time Delay Estimation	46
3.2	Localization Using Consistent Delay Vectors	50
3.3	Consistent Multi-Channel Maximum Likelihood Time Delay Estimation . .	55
3.4	Practical Time Delay Estimation Systems	59
3.4.1	Discrete Time Delay Estimation	59
3.4.2	Voice Activity Detection	60
3.4.3	Numerical Techniques	60
3.5	Experimental Results	61
3.5.1	Simulations Results	61
3.5.2	Real Room Results	69
3.6	Source Localization Summary	71

4	Background Noise Reduction	75
4.1	Single Channel Approaches	75
4.1.1	Spectral Subtraction	76
4.1.2	Wiener Filtering	77
4.1.3	Other Techniques	78
4.2	Multi-Channel Approaches	79
4.3	Background Noise Estimation and Voice Activity Detection	80
4.4	Multi-Channel Minimum Statistics Denoising (MC-MSD)	81
4.5	Experimental Results	89
4.5.1	Simulations	89
4.5.2	Real Room Array	89
4.5.3	Real Room Circular Array	91
4.6	Denoising Summary	94
5	Multi-Microphone Model-Based Dereverberation	97
5.1	Single-Channel Approaches	98
5.2	Multi-Channel Approaches	99
5.3	Multi-Channel Model-Based Dereverberation Algorithm (MC-MBD)	104
5.3.1	Joint Inverse LPC Filter	106
5.3.2	Wavelet Transform of LPC Residual Signals	107
5.3.3	Wavelet Transform Coefficient Extrema Clustering	108
5.3.4	Short-Term Coherence Scaling	109
5.3.5	Wavelet Transform Coefficient Extrema Reconstruction	112
5.3.6	Forward LPC Filter	113
5.4	Experimental Results	115
5.5	Dereverberation Summary	118
6	System Performance	121
6.1	Simulations of Denoising and Dereverberation	121
6.2	Speech Processing for Automated Videoconferencing	124
6.2.1	Audio Processing	127
6.2.2	Video Processing	130
6.3	System Summary	132
7	Conclusion	134
7.1	Discussion	134
7.2	Future Work	135
A	The Multi-Channel Maximum Likelihood Time Delay Estimator	137

List of Figures

1.1	<i>Block diagram of speech enhancement system</i>	3
2.1	<i>Physiological components of speech production</i>	8
2.2	<i>Schematic diagram of speech production system</i>	8
2.3	<i>Discrete-time components of speech production</i>	9
2.4	<i>Glottal waveform example</i>	11
2.5	<i>Two stages of the Wavelet Transform analysis filter bank for the Algorithme à Trous where $\bar{h}_k[n] = h_k[-n]$</i>	15
2.6	<i>Two stages of the Wavelet Transform synthesis filter bank for the Algorithme à Trous</i>	16
2.7	<i>The wavelet and scaling functions</i>	18
2.8	<i>A Wavelet Transform example using a derivative mother wavelet. The top plot is the original input signal and the remaining plots are the Wavelet Transform coefficients at scales 2^1 through 2^4</i>	20
2.9	<i>An example of the Wavelet Transform of a speech signal. The top plot is the original input signal and the remaining plots are the Wavelet Transform coefficients at scales 2^1 through 2^4</i>	21
2.10	<i>The Wavelet Transform of speech plus additive white noise. The top plot is the original input signal and the remaining plots are the Wavelet Transform coefficients at scales 2^1 through 2^4</i>	22
2.11	<i>Clean (red) and reverberant (blue) speech signals and corresponding Wavelet Transforms</i>	23
2.12	<i>Alternating projection algorithm [Mallat and Zhong (1992)]</i>	24
2.13	<i>Wavelet extrema reconstruction example</i>	26
2.14	<i>SNR of the first six scales and the reconstructed signal for Wavelet Transform extrema reconstruction example</i>	27
2.15	<i>Wavelet extrema reconstruction speech example</i>	27
2.16	<i>SNR of the first six scales and reconstructed signal for Wavelet Transform extrema reconstruction example using a speech signal as input</i>	28
2.17	<i>Block diagram of Bark Spectrum calculation</i>	33
2.18	<i>Hertz-to-Bark scale transformation for the approximation and experimental values</i>	36
2.19	<i>Critical band filter</i>	36
2.20	<i>Equal loudness curves. MAF = Minimum Audible Field</i>	37
2.21	<i>Conversion from phon scale to sone scale</i>	38

3.1	<i>Two-channel time delay estimation problem</i>	44
3.2	<i>Bearing angle estimates derived from independent PHAT GCC time delay estimates. The true source location is indicated by $\square$ and the 8 microphones are indicated by $\times$.</i>	52
3.3	<i>Plot of the PHAT GCC for the rightmost microphone pair in Figure 3.2. The true delay is -3.01 samples, while the estimated delay is -11.2 samples.</i>	53
3.4	<i>Bearing angle estimates for consistent PHAT GCC time delay estimation. The true source location is indicated by $\square$ and the 8 microphones are indicated by $\times$.</i>	54
3.5	<i>Qualitative plots of the $-E_{LS}$ and E_{CON} criteria typical of a source near the array for the four TDE methods considered. The source location is indicated by $\square$, the source location estimate by $+$, and the 8 microphones (two rows of 4) by $\times$.</i>	62
3.6	<i>Response pattern for a cardioid microphone</i>	63
3.7	<i>Qualitative plots of the $-E_{LS}$ and E_{CON} criteria typical of a source far from the array for the four TDE methods considered. The source location is indicated by $\square$, the source location estimate by $+$, and the 8 microphones (two rows of 4) by $\times$.</i>	65
3.8	<i>Qualitative plots of the $-E_{LS}$ and E_{CON} criteria for 2 simultaneous sources for the four TDE methods considered. The source location is indicated by $\square$ and the 8 microphones (two rows of 4) by $\times$.</i>	66
3.9	<i>10 random source locations, indicated by $\square$ used for TDE simulations. The microphones are positioned in two rows of 4, indicated by $\times$.</i>	67
3.10	<i>Analysis of the LS (XC-LS, BEN-LS, PHAT-LS, MLE-LS) and consistency-based (XC-CON, BEN-CON, PHAT-CON, MLE-CON) locators for various reverberation levels and 15 dB SNR additive white noise</i>	68
3.11	<i>Comparison of the consistency-based (XC-CON, BEN-CON, PHAT-CON, MLE-CON) locators for various reverberation levels and 15 dB SNR additive white noise</i>	69
3.12	<i>Comparison of PHAT-CON accuracy results for different angle criteria</i>	70
3.13	<i>Comparison of PHAT-CON, PHAT-CON-ALL, and MLE-CON locators for various reverberation levels and 15 dB SNR additive white noise</i>	70
3.14	<i>Real room plots of the $-E_{LS}$ and E_{CON} criteria. The source location is indicated by $\square$, the source location estimate by $\circ$, and the 8 microphones (two rows of 4) by $\times$</i>	73
3.15	<i>Real room qualitative plots of the $-E_{LS}$ and E_{CON} criteria. The source location is indicated by $\square$, the source location estimate by $\circ$, and the 8 microphones (two rows of 4) by $\times$</i>	74
4.1	<i>Outline of the proposed multi-channel denoising algorithm</i>	82
4.2	<i>Speech signal with additive noise and the time-varying DFT energy of the 586 Hz frequency bin and resulting Minimum Statistics noise estimate.</i>	84
4.3	<i>True and estimated noise variances before and after scaling to correct for downward bias of Minimum Statistics noise estimate</i>	85
4.4	<i>Oversubtraction factor used for Spectral Subtraction</i>	86

4.5	<i>Multi-channel denoising results for various SNR levels</i>	90
4.6	<i>Real room denoising results</i>	91
4.7	<i>Location of talker, fan noise source, and 16 microphones in real room denoising experiment</i>	92
4.8	<i>Circular array of 16 microphones used in real room denoising experiment</i>	93
4.9	<i>Circular array experiment room setup</i>	94
4.10	<i>Optimal SNR microphone subset selection where subset 1 is centered around the topmost microphone in Figure 4.9</i>	95
4.11	<i>Circular array denoising results</i>	96
5.1	<i>Array response for a horizontal slice of a $4\text{ m} \times 4\text{ m}$ room for $\theta = 0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ$, and 75° from vertical at a frequency of 2000 Hz where the 8 microphones are represented by $\times$ and the point of focus is indicated by $\circ$</i>	102
5.2	<i>Outline of the proposed dereverberation algorithm</i>	106
5.3	<i>Plot of 8 reverberant channels at scale 2^5 and clustered peaks</i>	109
5.4	<i>Clustered reverberant peaks and scale 2^5 of clean speech</i>	110
5.5	<i>Short-term coherence function for scale 2^3</i>	111
5.6	<i>Weighted peaks and scale 2^5 of clean speech</i>	112
5.7	<i>LPC residual of clean speech, reverberant speech, and after wavelet clustering technique</i>	114
5.8	<i>Comparison frame 1: clean, reverberant, beamformed, and MC-MBD result</i>	115
5.9	<i>Comparison frame 2: clean, reverberant, beamformed, and MC-MBD result</i>	116
5.10	<i>Comparison frame 3: clean, reverberant, beamformed, and MC-MBD result</i>	117
5.11	<i>Comparison of clean, reverberant, beamformed, and the proposed MC-MBD algorithm (Reverberation-Only Case)</i>	119
5.12	<i>Multi-channel dereverberation results for various reverberation times</i>	120
6.1	<i>Comparison of clean, reverberant, beamformed, and the proposed MC-MSD-MBD algorithm (reverberation plus noise case)</i>	123
6.2	<i>Results for varying levels of reverberation for an additive noise level of 10 dB SNR</i>	124
6.3	<i>Results for varying levels of SNR for a reverberation time of 200 ms</i>	125
6.4	<i>Outline of the overall automated videoconferencing system</i>	127
6.5	<i>Overhead view of the room showing microphone and camera positions</i>	129
6.6	<i>Speech enhancement results</i>	129
6.7	<i>Face tracking example [Wang et al. (2001)]</i>	131
6.8	<i>Examples of face orientation estimation where the detected hair region is yellow on the bottom right of each figure and the clock face indicates the estimated pose angle [Wang et al. (2001)]</i>	132

List of Tables

2.1	<i>Coefficients of analysis and synthesis filters</i>	18
2.2	<i>Mean Opinion Score scoring system</i>	31
2.3	<i>Relationship between critical band rate z, critical band limits, and critical band center frequencies</i>	35
6.1	<i>Segmental SNR scores for varying values of SNR and reverberation levels for beamforming, independent denoising and dereverberation, and joint denoising and dereverberation</i>	125
6.2	<i>BSD scores for varying values of SNR and reverberation levels for beamforming, independent denoising and dereverberation, and joint denoising and dereverberation</i>	126

Chapter 1

Problem Overview and Motivation

Speech signals have become increasingly vital components in modern human-machine interfaces. Background noise and reverberation effects produce aesthetically undesirable effects and diminish a system's ability to convey information. Signal quality also plays a key role in the performance of speech analysis systems. For instance, speech recognition and speaker identification techniques exhibit dramatic performance degradations when low quality speech signals are examined. Therefore, speech acquisition and the ensuing enhancement of that speech are both essential for the reliable performance of these interfaces.

Currently, most speech acquisition systems rely on the user to be physically close to the microphone to achieve reasonable sound quality. This close-talker condition significantly simplifies the acquisition problem by emphasizing the desired signal relative to background noise and other sources, and by reducing the effects of the physical channel on the signal of interest. However, there are many situations in which it is not possible or desirable to have a talker or talkers physically linked to the acquisition device.

The ultimate goal of this work is to allow users the opportunity to roam freely in diverse environments while still providing a high quality speech signal and a robustness to background noise and reverberation effects. Besides enhancing current human-machine

interfaces, advances in this technology are providing for a variety of new avenues for human-machine interaction. In recent years, the use of arrays of multiple microphones has received considerable attention as a means for dramatically improving the performance of traditional single-microphone systems. This multi-sensor approach has opened up a range of additional topics of study [Brandstein and Ward (2001), Flanagan and Silverman (1994), Flanagan and Silverman (1992)].

For single channel systems, recent research has focused on appropriate modeling of the speech signal, while multi-channel system research has emphasized improved spatial filtering.

In this thesis, a microphone array speech enhancement system is presented for signals acquired in challenging environments degraded by background noise and room reverberations. These novel speech source localization, denoising, and model-based dereverberation algorithms perform well in the presence of such degradations. The improved performance of the proposed system relative to classical methods is demonstrated using quantitative speech quality measures.

The three main components of the enhancement system are shown in Figure 1.1. The remainder of this chapter presents a brief introduction to each of the three components of the proposed system.

1.1 Source Localization and Time Delay Estimation

Nearly all multi-channel (i.e., multi-microphone) speech acquisition problems require the reliable estimation of active talker locations and the resulting time delays relative to the set of microphones employed. This is the goal of the first stage of the proposed speech en-

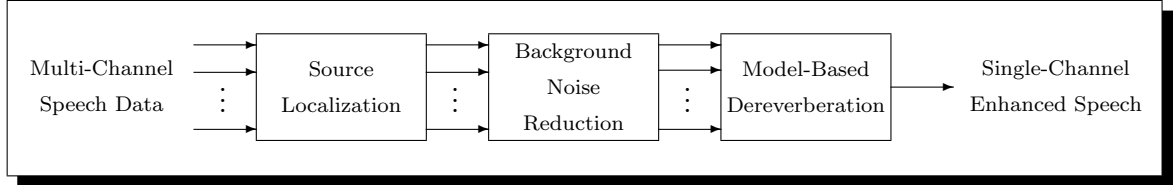


Figure 1.1: *Block diagram of speech enhancement system*

hancement system. For speech enhancement applications, accurate knowledge of locations of the desired talker and the interference sources is necessary to effectively steer the array and enhance a desired source, while simultaneously attenuating those deemed undesirable. Location data may be used as a guide for discerning individual speakers in a multi-source scenario. With this information available, it would be possible to automatically focus on and track a given source for an extended time period. Of particular interest lately is the application of the speaker location estimates for aiming a camera or series of cameras in a video-conferencing system [Wang et al. (2001), Wang and Chu (1997)]. The source location estimate and its implied vector of relative time-delay estimates are integral parts of the microphone array system. Such information is a necessary precursor to time-alignment of microphone signals, rejection of noise sources, and the solution of the general beamforming problem.

The localization algorithm presented in this thesis is effective in the presence of both background noise and reverberations, and it simultaneously produces the relative time delay estimates and a source location estimate. This approach modifies the traditional two-step localization procedure of time delay estimation followed by source localization, and simultaneously solves for each element in the estimated time delay vector. The result is a procedure that adds a level of robustness to any time delay estimation technique which attempts to maximize or minimize an appropriate objective function independently

for the different microphone pairs in the array. Simulations are performed across a range of reverberation conditions to illustrate the efficacy of the proposed method relative to standard approaches.

1.2 Background Noise Reduction

The second and third stages of the proposed speech enhancement system deal with the reduction of background noise and reverberations, respectively. Background noise (e.g., fan noise) and reverberations are typically handled with different approaches, as the former degradation is additive in nature while the latter is convolutional. Both types of degradation are common and problematic in any real environment. The enhancement of speech that has been degraded by background noise is an often-researched problem, and there is a rich history of work addressing the use of single channel methods for speech enhancement. Summaries of these techniques may be found in [Ephraim (1992), Furui and Sondhi (1992), Deller et al. (1987), Lim (1983)]. While capable of improving perceived quality in certain environments (i.e., additive noise, high signal-to-noise ratios, and no reverberations), these approaches do not perform well in the face of reverberant distortions and severe noise conditions.

Single channel techniques for background noise reduction typically involve some form of spectral modification (e.g., Spectral Subtraction [Boll (1979)]) or postfiltering (e.g., Wiener filtering [Deller et al. (1987), Lim (1983)]). However, many of these methods require an estimate of the background noise spectrum. This noise spectrum is often estimated from non-speech intervals, though the detection of non-speech intervals is a difficult problem itself. A practical method for background noise estimation is the Minimum Statistics procedure developed in [Martin (2001)]. This approach hypothesizes that the minimum of a given power spectral frequency bin over an extended period of time, including both speech

and non-speech intervals, provides a reasonable estimate of the noise floor - an intuitive argument. In this thesis, this technique is used in a multi-microphone setting as a precursor to a Delay-and-Sum Beamforming operation. This technique is both effective at removing background noise and capable of running in real time. Results will be presented for varying levels of signal-to-noise ratios to show how much of an improvement is offered relative to the use of beamforming alone.

1.3 Speech Dereverberation

The final stage of the enhancement system is multi-channel dereverberation. The proposed system incorporates nonlinear, model-based processing to reduce room reverberations. The goal is to combine the advantages of spatial filtering achieved through beamforming with knowledge of the desired time-series attributes and intuitive nonlinear processing. A multi-channel wavelet algorithm which incorporates these principles is presented for which large Wavelet Transform coefficients correspond to events - either speech production events or arrivals of multipath signals (i.e., reverberations). This allows for the discrimination of portions of the signal produced by the original speech utterance from those due to reverberation effects. Because reverberations depend largely on the geometry of the environment as well as the location of the speech source and microphones, the signal acquired at each microphone will be degraded to varying degrees, as evidenced by variations in Wavelet Transform coefficient extrema from channel to channel. By deciding which time delay compensated signal portions are coherent over all microphones, it is possible to recover the structure of the original speech from the noisy, received data.

The algorithm is shown to be capable of identifying and attenuating reverberant portions of the speech signal. This is achieved without explicitly requiring estimation of the channel characteristics, a feature which is essential for practical applications, but which is absent

from the majority of existing dereverberation approaches.

1.4 Thesis Organization

This chapter provided a brief introduction to the proposed enhancement algorithm for speech degraded by background noise and room reverberations. Chapter 2 presents some fundamentals of speech production and some speech processing tools necessary for the remainder of the thesis. Chapters 3 through 5 detail the three components of the enhancement system shown in Figure 1.1. Chapter 3 deals with robust time delay estimation. Chapter 4 presents a multi-channel extension of Martin’s Minimum Statistics approach to background noise reduction. Chapter 5 discusses the model-based dereverberation algorithm. Chapter 6 presents some results of the overall system, while Chapter 7 offers some concluding remarks and avenues for future research.

Chapter 2

Speech Production and Processing Background

This chapter discusses the speech production model as well as the mathematics of speech processing. These concepts and equations will be used throughout the remainder of the thesis and are presented here as background information for the reader new to these topics.

2.1 The Speech Production Model

The human speech production system consists primarily of the vocal cords, the vocal tract, the nasal cavity, and the lips [Rabiner and Shafer (1978), Deller et al. (1987)]. Figure 2.1, taken from [Deller et al. (1987)], shows the different physiological elements of this system. Air is forced from the lungs through the vocal cords and into the vocal tract which consists of the pharyngeal and oral cavities. Between these two cavities hangs a fold known as the velum which opens to allow for the coupling of the vocal and nasal tracts necessary to produce nasal sounds.

Figure 2.2 shows a schematic diagram encompassing the main components of the speech

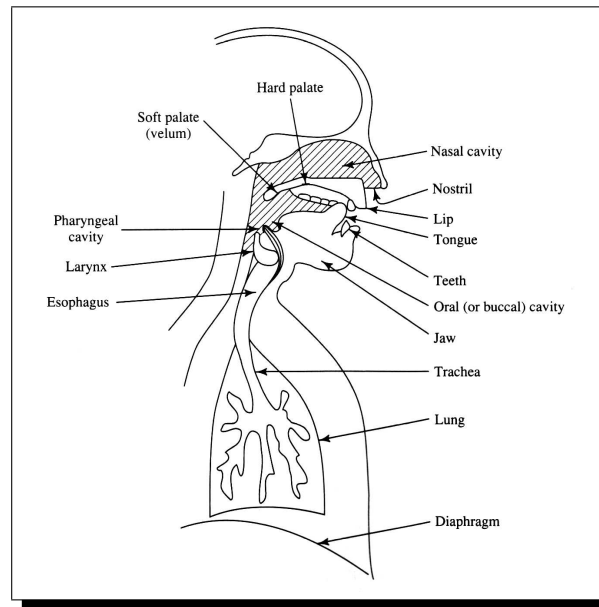


Figure 2.1: *Physiological components of speech production*

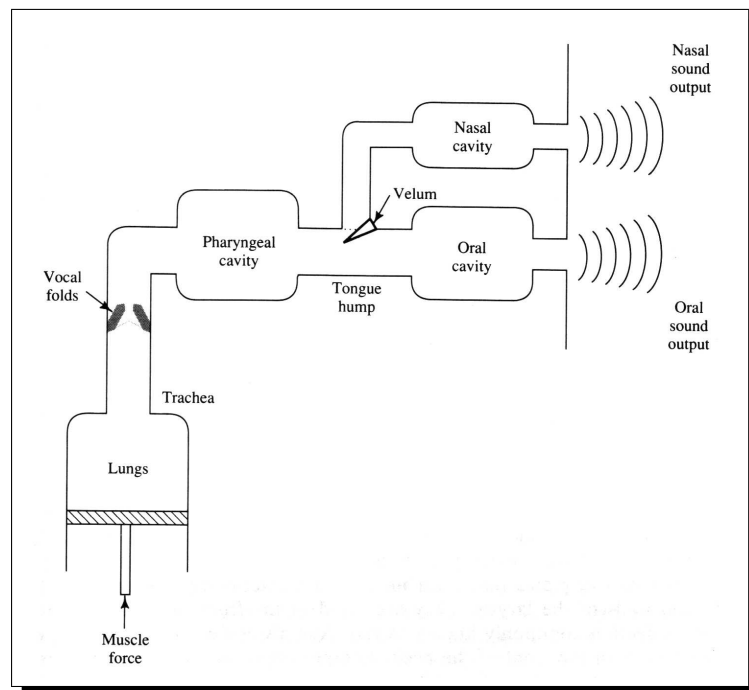
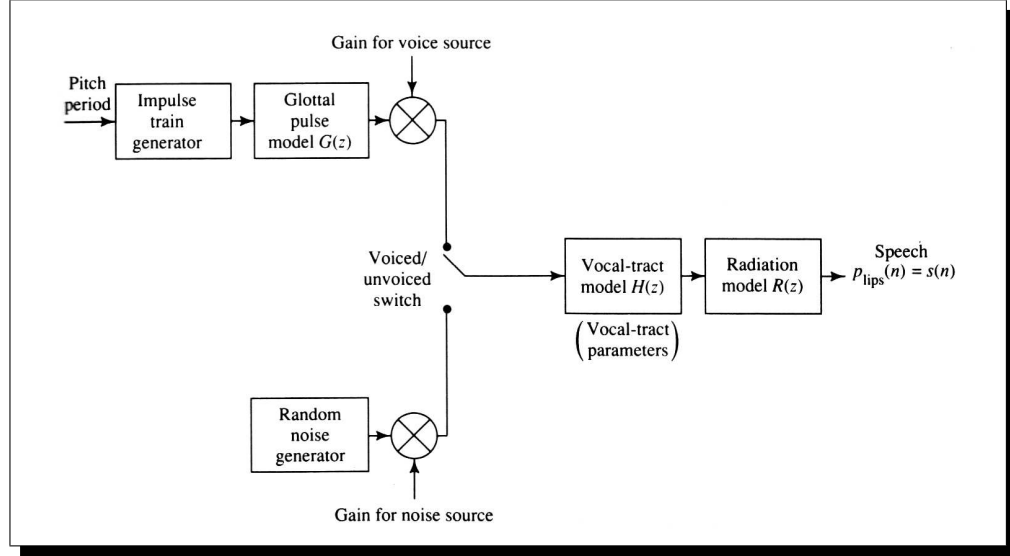


Figure 2.2: *Schematic diagram of speech production system*

Figure 2.3: *Discrete-time components of speech production*

production system [Deller et al. (1987)]. For voiced sounds such as vowels, the lungs force air through an opening in the vocal cords, known as the glottis. Adjusting the tension in the vocal cords causes them to open and close in a quasi-periodic fashion. This results in a palpable vibration of the vocal tract area. For unvoiced sounds such as /s/ and /p/, air is forced through a constriction at some point in the vocal tract and the periodicity characteristic of voiced sounds is not present. Other sounds such as /v/ and /z/ are the results of a mixed excitation signal containing both voiced and unvoiced elements. The excitation function, whether voiced, unvoiced, or mixed, drives the speech system and the resulting speech waveform is radiated from the lips (and perhaps the nostrils, should the nasal cavity be coupled to the vocal tract).

The discrete-time model of speech production based on the above discussion is shown in Figure 2.3 [Rabiner and Shafer (1978), Deller et al. (1987)]. Mathematically, voiced speech sounds are modeled by:

$$s[n] = e_p[n] * g[n] * h_v[n] * r[n] \quad (2.1)$$

where $e_p[n]$ is a quasi-periodic impulse train excitation signal, $g[n]$ is a glottal shaping pulse, $h_v[n]$ is the impulse response of the vocal tract for the particular configuration of the physiological elements, and $r[n]$ models the radiator effects at the lips. Unvoiced speech is modeled by:

$$s[n] = e_w[n] * h_v[n] * r[n] \quad (2.2)$$

where $e_w[n]$ is a white noise excitation signal, with $h_v[n]$ and $r[n]$ defined as above. In the Fourier domain, Equation 2.1 and Equation 2.2 become:

$$S(\omega) = E_p(\omega)G(\omega)H_v(\omega)R(\omega) \quad (2.3)$$

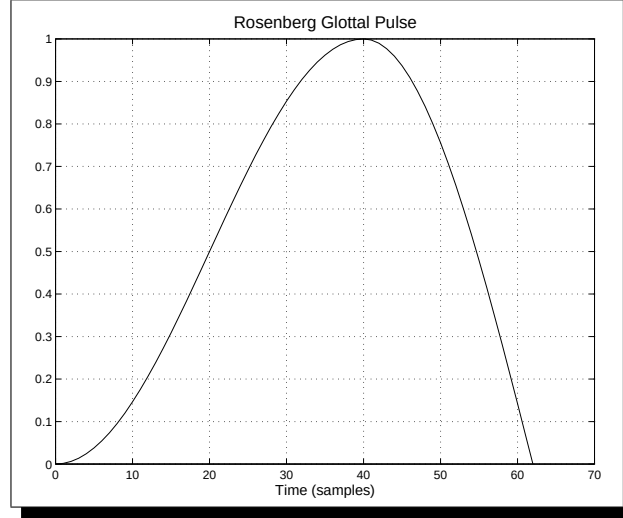
and

$$S(\omega) = E_w(\omega)H_v(\omega)R(\omega) \quad (2.4)$$

respectively.

Returning to voiced sounds, as pressure beneath the glottis increases, the vocal cords are forced open. However, as the glottis opens and the air volume velocity increases, the accompanying decrease in pressure forces the glottis closed again, a process that is repeated throughout the voiced sound. The time between successive glottal closures (or openings) can be defined as the fundamental period, or pitch, of the speech signal. The fundamental frequency is then defined as the reciprocal of the fundamental period. The fundamental frequency increases as the tension of the vocal tract is increased. The discrete-time waveform of the glottal shaping pulse is often approximated by Rosenberg's model [Rabiner and Shafer (1978)]:

$$g[n] = \begin{cases} \frac{1}{2} \left[1 - \cos \left(\frac{\pi n}{T_o} \right) \right], & 0 \leq n \leq T_o \\ \cos \left[\frac{\pi(n-T_o)}{2(T_c-T_o)} \right], & T_o \leq n \leq T_c \\ 0, & \text{otherwise} \end{cases} \quad (2.5)$$

Figure 2.4: *Glottal waveform example*

where T_o and T_c specify the times at which the glottis is wide open and completely closed respectively. A glottal waveform example is plotted in Figure 2.4.

As indicated in Equation 2.1, for voiced sounds, the glottal input to the vocal tract is the convolution of a quasi-periodic pulse train and the glottal pulse. The excitation signal for unvoiced sounds is usually modeled as white noise which is then used as the input to the vocal tract. The vocal tract transfer function, $H_v(\omega)$, is often modeled as a K^{th} order all-pole filter of the form:

$$H_v(\omega) = \frac{C}{1 - \sum_{k=1}^K a[k]e^{-j\omega k}} \quad (2.6)$$

where $j = \sqrt{-1}$, for some gain constant, C , and a set of filter coefficients, $a[k]$. Since this auto-regressive transfer function is equivalent to linear prediction in the time domain, the $a[k]$ terms are frequently referred to as Linear Prediction Coefficients (LPC) and the overall scheme is called the LPC speech model. Typically, an LPC order of $K = 10$ to $K = 16$ is sufficient to model realistic vocal tracts. The pole frequencies of the transfer function are known as the *formant frequencies* and roughly correspond to the resonant frequencies

of the vocal tract. Although the all-pole model is not accurate for all speech sounds, it is the model most often used in speech processing algorithms. It is motivated from a physical approximation of the vocal tract which treats the region from the vocal folds to the lips as a sequence of lossless tubes with varying cross-sectional areas.

The lip radiator, $r[n]$, models the relationship between the air volume leaving the vocal tract to a sound pressure waveform - what is typically measured by a microphone. A typical choice for the lip radiation function is:

$$R(z) = 1 - z_0 z^{-1}, \quad z_0 \approx 1, \quad z_0 < 1 \quad (2.7)$$

which is a simple highpass filter [Deller et al. (1987)].

This concludes the discussion of the speech production model. A series of mathematical tools is now presented which will be useful throughout the remainder of this thesis.

2.2 Speech Signal Processing Tools

2.2.1 The Fourier Transform

The Fourier transform, a very powerful tool for analyzing the frequency content of a signal, is defined as:

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt. \quad (2.8)$$

Note that the Fourier transform is the inner product of the signal $x(t)$ with an infinite-length complex exponential. That is, the Fourier transform at each frequency ω depends on the values of $x(t)$ over all time. Therefore, the Fourier transform is inherently non-local. Although it does indicate the frequencies present in the transformed signal, it gives no indication of where in time those frequencies occurred. Because of this limitation, other *time-frequency* approaches have been used to study speech signals. Time-frequency representations are often likened to musical transcription [Strang (1994)]. The Fourier transform

is equivalent to cataloging the numerous notes, or frequencies, present in a musical score with no attention being paid to their temporal location in the original signal. However, time-frequency analysis is analogous to musical notation, which shows how the different notes progress with time.

2.2.2 The Short-Time Fourier Transform

The short-time Fourier transform (STFT) is a logical extension of the Fourier transform and is given by [Cohen (1995)]:

$$\text{STFT}(t, \omega) = \int_{-\infty}^{\infty} x(s)h^*(s-t)e^{-j\omega s}ds \quad (2.9)$$

where the superscript $*$ denotes the complex conjugation operator. In this equation, a finite-length window, h , is centered at time t and multiplied by the original signal. The STFT at a given time instant t is simply the Fourier transform of a windowed portion of the signal. Although the simplicity of this interpretation is quite appealing, the STFT has two major drawbacks. First, it assumes that the signal of interest is stationary over the length of the analysis window. However, because the statistics of the speech signal can vary quite rapidly, it is not clear that such a stationarity assumption can be safely made.

The second limitation of the STFT is its fixed time-frequency resolution. In calculating the STFT, the length of the analysis window determines the time and frequency resolution of the resulting representation. If the effective duration, τ_{eff} , and effective bandwidth, ω_{eff} , are defined as [Porat (1994)]:

$$\tau_{eff} = \left[\frac{\int_{-\infty}^{\infty} s^2 |h(s)|^2 ds}{\int_{-\infty}^{\infty} |h(s)|^2 ds} \right]^{\frac{1}{2}}, \quad \omega_{eff} = \left[\frac{\int_{-\infty}^{\infty} \omega^2 |H(\omega)|^2 d\omega}{\int_{-\infty}^{\infty} |H(\omega)|^2 d\omega} \right]^{\frac{1}{2}} \quad (2.10)$$

where $H(\omega)$ is the Fourier Transform of the window $h(t)$, then the uncertainty principle is given by:

$$\tau_{eff} \cdot \omega_{eff} \geq 0.5. \quad (2.11)$$

Therefore, a shorter window will result in finer time resolution but more frequency smoothing. A longer window will result in temporal smoothing but finer frequency resolution. Unfortunately, after choosing the window function, h , the time-frequency resolution is fixed over the entire time-frequency plane.

2.2.3 The Wavelet Transform

The Wavelet Transform (WT) is a versatile tool that overcomes both limitations of the STFT, allowing a signal to be analyzed at different scales. This transform, in its traditional form, exhibits finer time resolution at high frequencies, and finer frequency resolution at low frequencies. This is well suited to speech signals which have shorter transient phenomena as well as longer slowly varying phenomena, and is also similar to how the human ear analyzes sounds [Pintér (1996)]. The Continuous Wavelet Transform of a signal $f(t)$ at scale s is given by [Mallat (1998)]:

$$Wf(u, s) = \int_{-\infty}^{\infty} f(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s} \right) dt = f * \bar{\psi}_s(u) \quad (2.12)$$

where $\psi(t)$ is the “mother wavelet” and:

$$\bar{\psi}_s(t) = \frac{1}{\sqrt{s}} \psi^* \left(\frac{-t}{s} \right). \quad (2.13)$$

Note that the $\frac{1}{\sqrt{s}}$ term is for norm preservation of $\psi(t)$ at different scales. The Dyadic Wavelet Transform constrains the scale parameter to satisfy $s = 2^j$ where j is an integer. There are certain continuous wavelet admissibility conditions that any $\psi(t)$ should satisfy, but the current discussion does not benefit from their inclusion. From a more practical filter bank perspective, Wavelet Transform calculation involves an analysis stage and a synthesis stage. Figures 2.5 and 2.6 show two analysis and two synthesis stages respectively of the wavelet filter bank used in this thesis [Mallat (1998)]. In these figures, c_j and d_j are the Wavelet Transform coefficients emitted from the lowpass and highpass filters respectively.

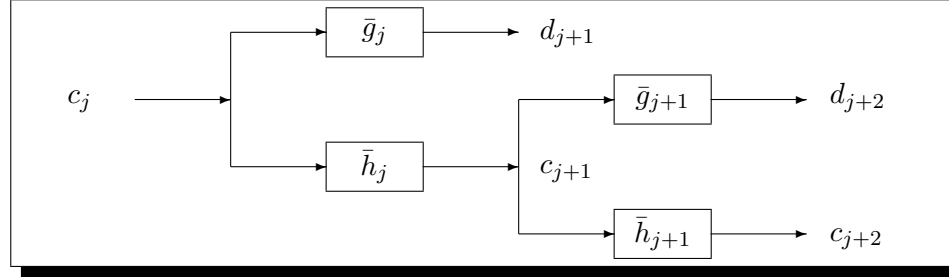


Figure 2.5: *Two stages of the Wavelet Transform analysis filter bank for the Algorithme à Trous where $\bar{h}_k[n] = h_k[-n]$*

The c_j coefficients are often called “approximation coefficients” while the d_j coefficients are referred to as “detail coefficients.” The coefficients of the analysis filters, $h[n]$ and $g[n]$, and the corresponding synthesis filters, $\tilde{h}[n]$ and $\tilde{g}[n]$, used in this thesis will be derived in the next section. Note that Figure 2.5 and Figure 2.6 show the fast dyadic Wavelet Transform known as the “Algorithme à Trous” which is French for “Algorithm with Holes.” This name arises from the fact that the filters themselves are dilated by a factor of two from one stage to the next with zeros inserted between the previous stage’s filter coefficients. This is in contrast to the traditional Wavelet Transform which leaves the filters themselves unchanged but which decimates the output of each analysis filter by a factor of 2 and upsamples the input of each synthesis filter by a factor of 2. Therefore, in Figure 2.5 and Figure 2.6, h_j comes from the filter h by inserting $2^j - 1$ zeros between consecutive h coefficients.

Quadratic Spline Wavelets

In this section, a certain wavelet system is derived where the wavelet is chosen as the derivative of a smoothing function. The usefulness of derivative wavelets will first be explored, followed by a derivation of the set of filter bank coefficients used in this thesis. This

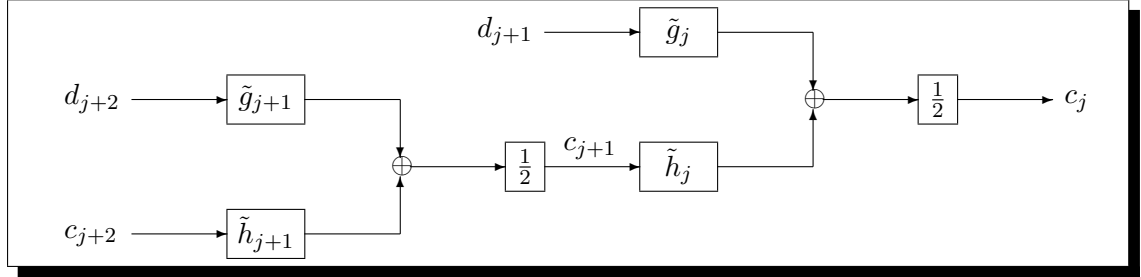


Figure 2.6: *Two stages of the Wavelet Transform synthesis filter bank for the Algorithme à Trous*

derivation is detailed in [Mallat (1998)]. Suppose the wavelet function $\psi(t)$ is defined as:

$$\psi(t) = -\frac{d\theta(t)}{dt} \quad (2.14)$$

where $\theta(t)$ is a smoothing (i.e., lowpass) function which integrates to 1 and converges to 0 at infinity [Mallat and Zhong (1992)]. Using Equation 2.14, the Wavelet Transform of Equation 2.12 can be written as:

$$Wf(u, s) = f * \left(s \frac{d\bar{\theta}_s}{dt} \right) (u) = s \frac{d}{du} (f * \bar{\theta}_s) (u). \quad (2.15)$$

Therefore, the wavelet coefficients at a given scale are proportional to the derivative of the original function smoothed to that scale.

A specific set of wavelet and scaling functions is now derived [Mallat (1998)]. When building a dyadic wavelet function, one first starts with a scaling function. The scaling function, ϕ , is the result of an infinite cascading of the lowpass filters, $h[n]$, used in the wavelet filter bank. This fact, along with the desire for a multiresolution approximation which, by definition, consists of appropriately nested subspaces [Mallat (1998)], leads to a closed-form condition of:

$$\Phi(\omega) = \frac{1}{\sqrt{2}} H\left(\frac{\omega}{2}\right) \Phi\left(\frac{\omega}{2}\right). \quad (2.16)$$

Similarly, the corresponding wavelet, ψ , satisfies:

$$\Psi(\omega) = \frac{1}{\sqrt{2}} G\left(\frac{\omega}{2}\right) \Phi\left(\frac{\omega}{2}\right). \quad (2.17)$$

Defining the scaling function to be:

$$\Phi(\omega) = e^{-j\frac{\omega}{2}} \left(\frac{\sin \frac{\omega}{2}}{\frac{\omega}{2}} \right)^3 \quad (2.18)$$

results in the following lowpass filter function:

$$H(\omega) = \sqrt{2} \frac{\Phi(2\omega)}{\Phi(\omega)} = \sqrt{2} e^{-j\frac{\omega}{2}} \left(\cos \frac{\omega}{2} \right)^3 \quad (2.19)$$

where Equation 2.16 was used along with the fact that $\sin 2a = 2 \cos a \sin a$. In order to construct a wavelet with one vanishing moment, the highpass filter is specified as:

$$G(\omega) = -j\sqrt{2} e^{-j\frac{\omega}{2}} \sin \frac{\omega}{2}. \quad (2.20)$$

Using Equation 2.17, the resulting wavelet is given by:

$$\Psi(\omega) = \frac{1}{\sqrt{2}} G\left(\frac{\omega}{2}\right) \Phi\left(\frac{\omega}{2}\right) = -\frac{j\omega}{4} e^{-j\frac{\omega}{2}} \left(\frac{\sin \frac{\omega}{4}}{\frac{\omega}{4}} \right)^4. \quad (2.21)$$

The wavelet and scaling functions are shown in Figure 2.7.

By choosing $\tilde{h}[n] = h[n]$, the dual scaling and wavelet functions will also be splines.

The reconstruction condition for this wavelet system is:

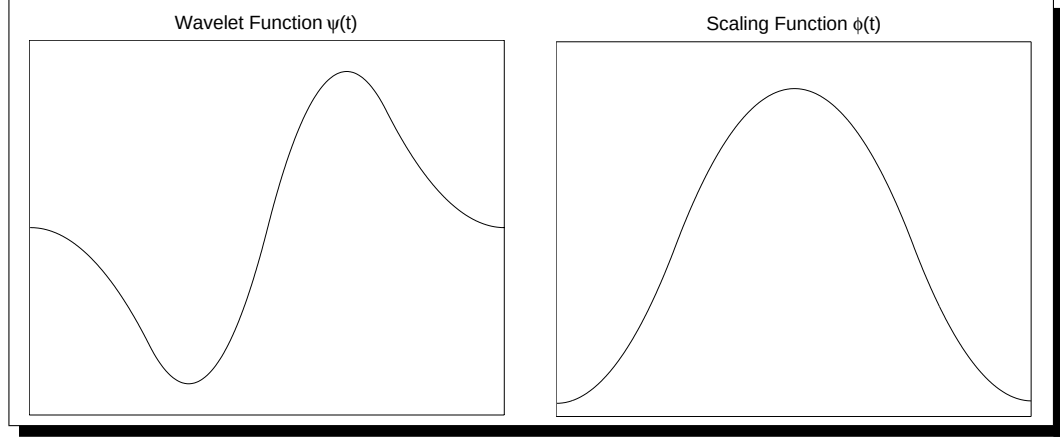
$$\tilde{H}(\omega) H^*(\omega) + \tilde{G}(\omega) G^*(\omega) = 2. \quad (2.22)$$

Therefore, the last of the filters necessary to implement the filter bank wavelet structure is:

$$\tilde{G}(\omega) = \frac{2 - |H(\omega)|^2}{G^*(\omega)} = \frac{2 - 2 \left(\cos \frac{\omega}{2} \right)^6}{j\sqrt{2} e^{j\frac{\omega}{2}} \sin \frac{\omega}{2}}. \quad (2.23)$$

Using the fact that:

$$\sum_{n=0}^2 \left(\cos \frac{\omega}{2} \right)^{2n} = \frac{1 - \left(\cos \frac{\omega}{2} \right)^6}{\left(\sin \frac{\omega}{2} \right)^2} \quad (2.24)$$

Figure 2.7: *The wavelet and scaling functions*

n	$h[n]/\sqrt{2}$	$\tilde{h}[n]/\sqrt{2}$	$g[n]/\sqrt{2}$	$\tilde{g}[n]/\sqrt{2}$
-2				-0.03125
-1	0.125	0.125		-0.21875
0	0.375	0.375	-0.5	-0.6875
1	0.375	0.375	0.5	0.6875
2	0.125	0.125		0.21875
3				0.03125

Table 2.1: *Coefficients of analysis and synthesis filters*

leads to:

$$\tilde{G}(\omega) = -j\sqrt{2}e^{-j\frac{\omega}{2}} \sin \frac{\omega}{2} \sum_{n=0}^2 \left(\cos \frac{\omega}{2} \right)^{2n}. \quad (2.25)$$

According to Fourier theory, if a signal $x(t)$ has Fourier transform $X(\omega)$, then $\frac{dx(t)}{dt}$ has Fourier transform $j\omega X(\omega)$. Therefore, the wavelet function given in Equation 2.21 is the derivative of a smoothing function $\theta(t)$, whose Fourier transform is given by:

$$\Theta(\omega) = \frac{1}{4}e^{-j\frac{\omega}{2}} \left(\frac{\sin \frac{\omega}{4}}{\frac{\omega}{4}} \right)^4. \quad (2.26)$$

The finite impulse responses of these filters are given in Table 2.1.

Figure 2.8 shows an example using this derivative wavelet. Note that discontinuities

appear as coefficients with large magnitudes in each scale. Local maxima of the coefficient magnitudes correspond to sharply changing signal features (at that scale) while local minima correspond to slowly varying phenomena. The wavelet coefficients are zero where the signal (at that scale) is constant, a fact that allows the use of such wavelets for speech coding [Srinivasan and Jamieson (1996)].

The Wavelet Transform of Speech Signals

Figure 2.9 shows an example of the Wavelet Transform for a voiced speech signal. Recall that voiced speech is modeled as a vocal tract response excited by a glottal pulse. This pulse essentially excites all of the formant frequencies simultaneously, resulting in a maximal derivative in the speech signal [Kadambe (1992), Kadambe and Boudreaux-Bartels (1992), Hess (1983)]. This results in Wavelet Transform extrema in each scale at the point of the discontinuity. Therefore, the Wavelet Transform can be used as an event detector, with the event being the instant of glottal closure. The estimation of this point of excitation has been a heavy research area [Harris and Nelson (1993), Cheng and O’Shaughnessy (1989), Ananthapadmanabha and Yegnanarayana (1975), Strube (1974)].

Figure 2.10 shows the Wavelet Transform of a speech signal corrupted by additive white noise. Note that at lower resolutions, much of the noise has been filtered out, and the local extrema still correspond to the points of glottal closure. Figure 2.11 compares the Wavelet Transform of clean speech and the same signal corrupted by reverberation. An important point to note is that over multiple channels, the extrema due to the points of glottal closure are more coherent than the extrema due to reverberations. This observation will be utilized in Chapter 5, which deals with model-based dereverberation.

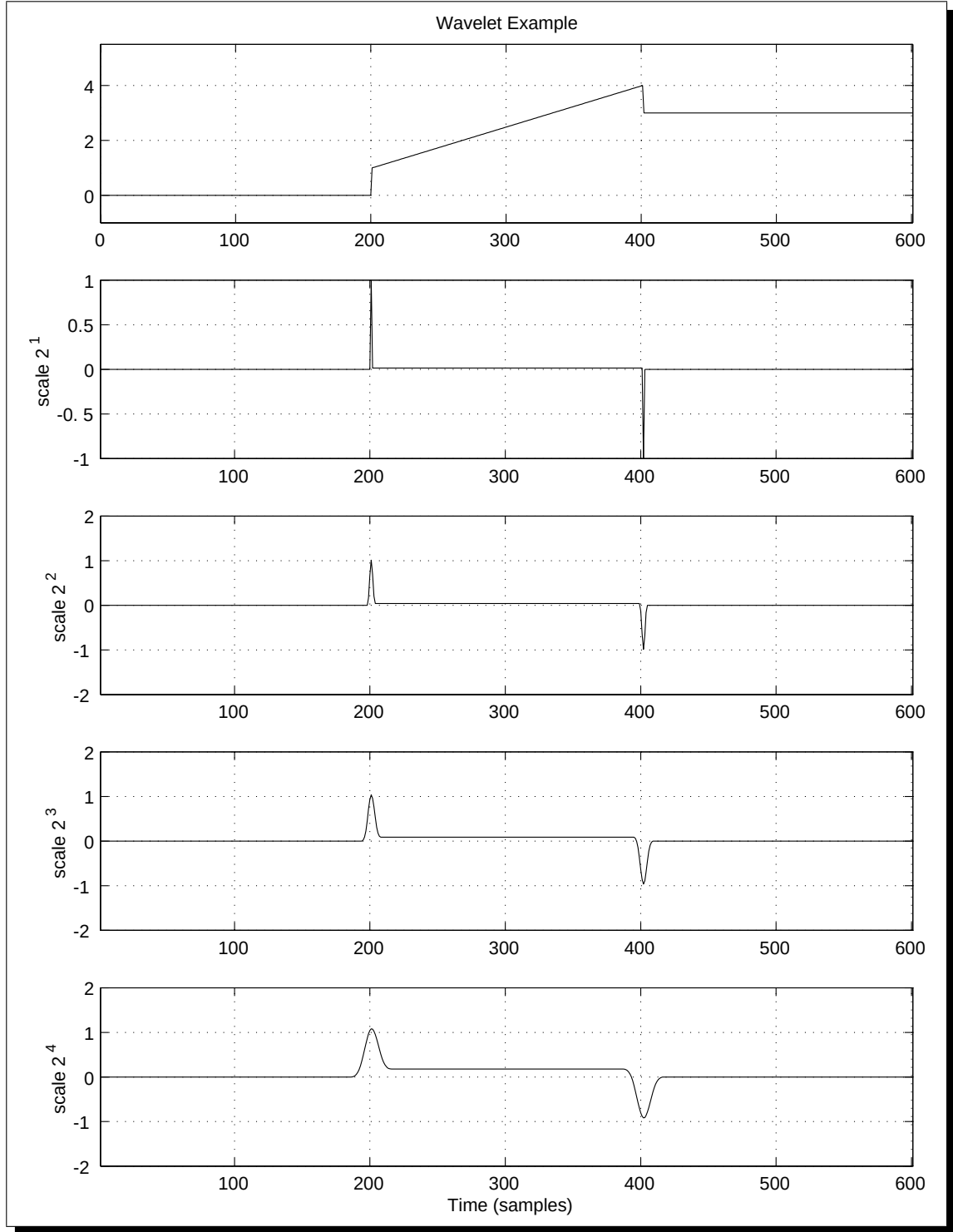


Figure 2.8: A Wavelet Transform example using a derivative mother wavelet. The top plot is the original input signal and the remaining plots are the Wavelet Transform coefficients at scales 2^1 through 2^4

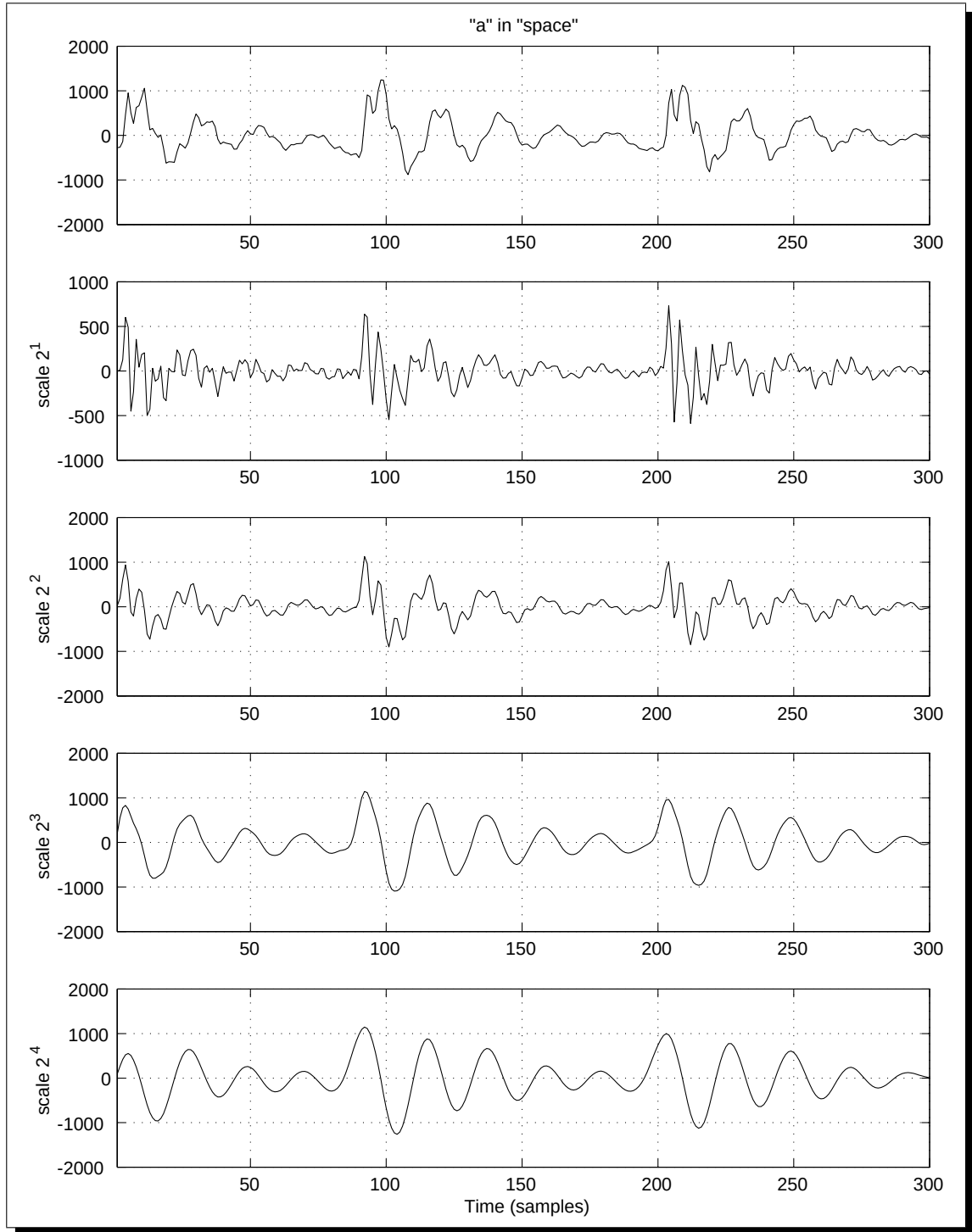


Figure 2.9: An example of the Wavelet Transform of a speech signal. The top plot is the original input signal and the remaining plots are the Wavelet Transform coefficients at scales 2^1 through 2^4

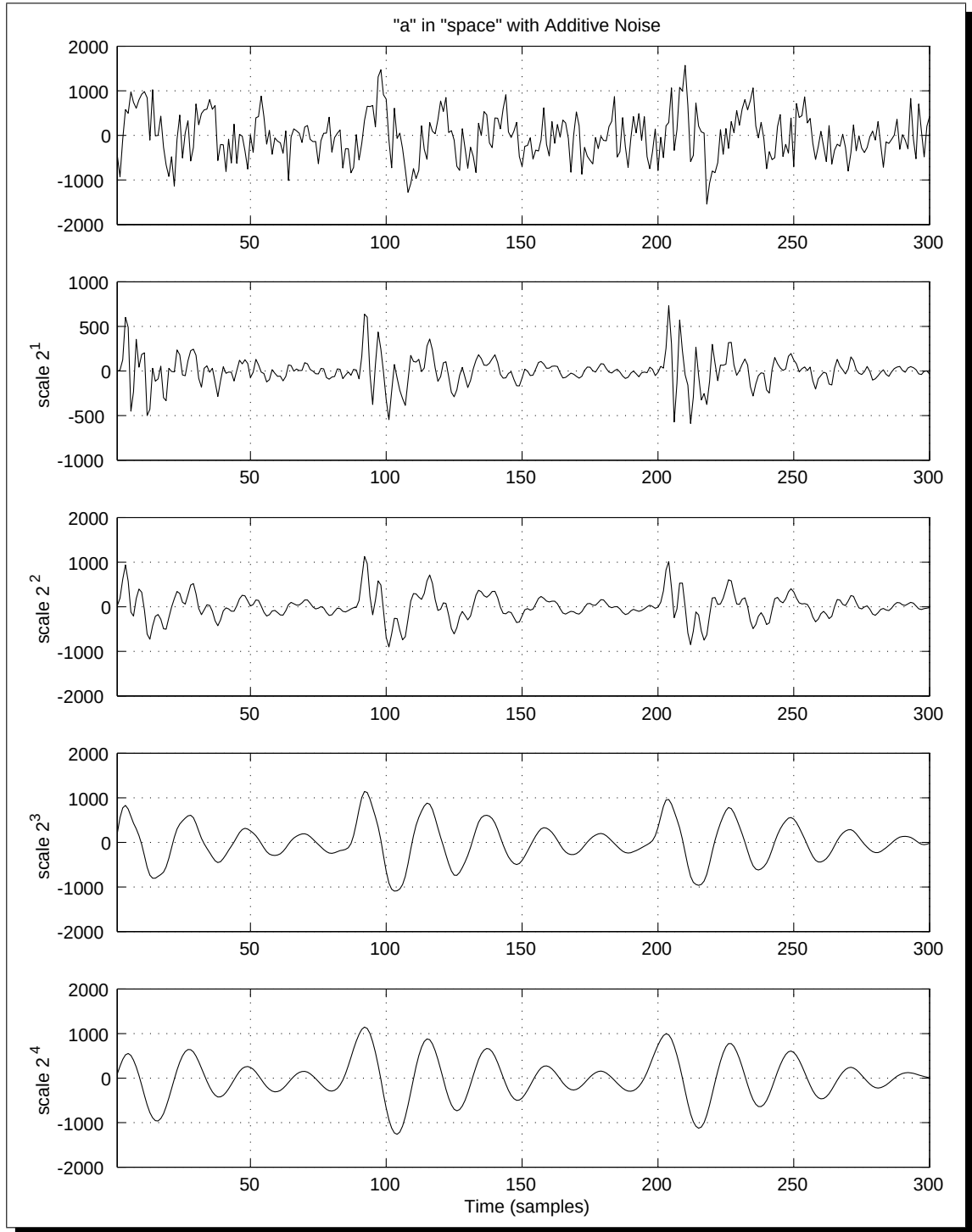


Figure 2.10: *The Wavelet Transform of speech plus additive white noise. The top plot is the original input signal and the remaining plots are the Wavelet Transform coefficients at scales 2^1 through 2^4*

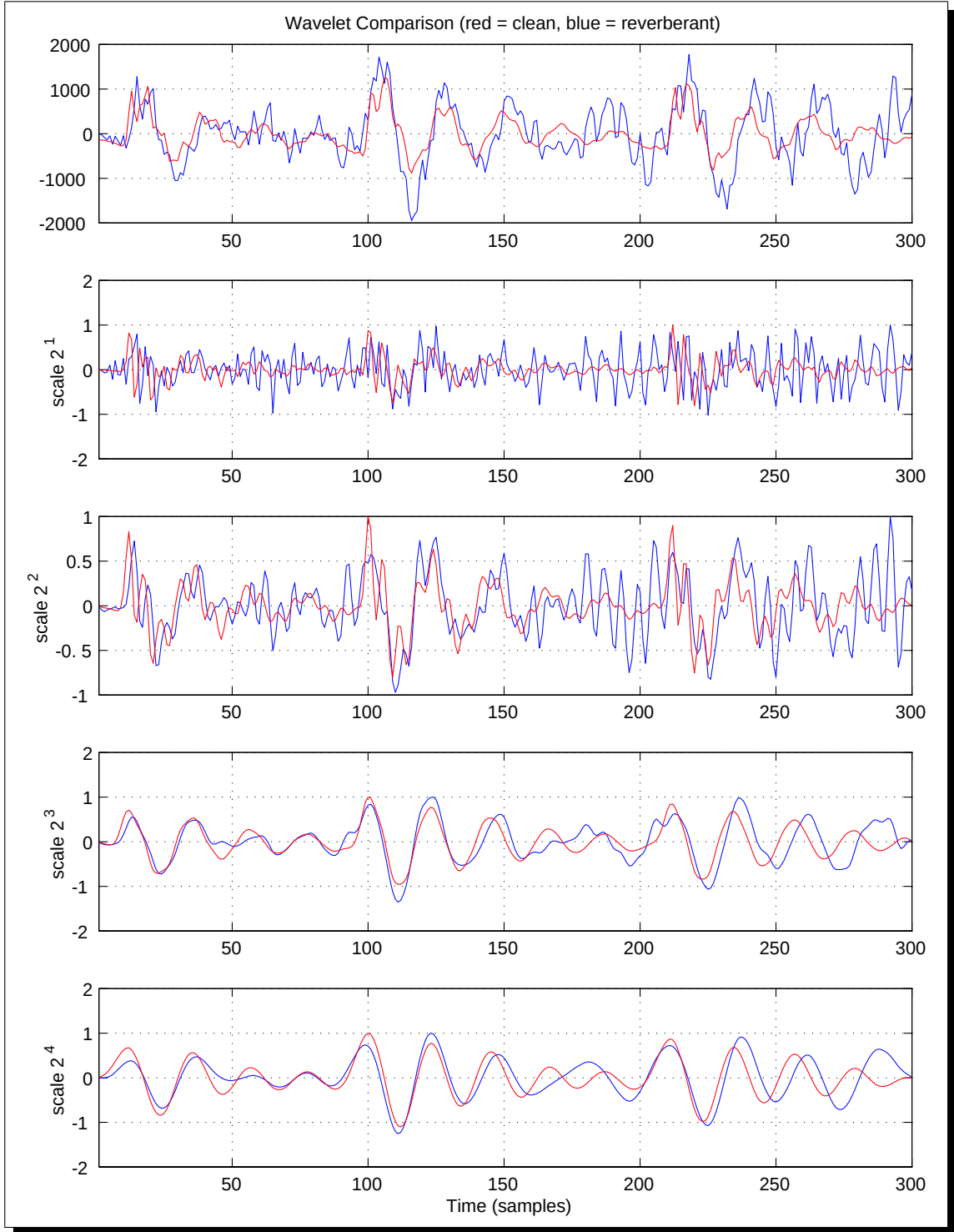


Figure 2.11: *Clean (red) and reverberant (blue) speech signals and corresponding Wavelet Transforms*

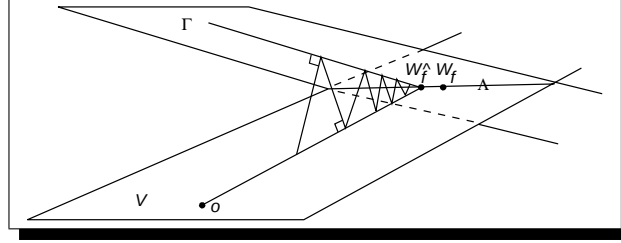


Figure 2.12: *Alternating projection algorithm [Mallat and Zhong (1992)]*

Wavelet Transform Extrema Reconstruction

The extrema of the wavelet coefficients at each scale have been used extensively for signal denoising, detection, and coding [Srinivasan and Jamieson (1996), Mallat and Zhong (1992)]. In fact, it has been shown that a signal, $f(t)$, can be closely approximated by reconstruction from only the lowest resolution approximation coefficients and the largest wavelet extrema of each detail signal. Mallat proposed an algorithm in [Mallat and Zhong (1992)] that numerically reconstructs an estimate, $\hat{f}(t)$, of the signal, $f(t)$, from the extrema of $Wf(u, 2^j)$ for integer j through the use of alternating projections. Define the V as the space of all possible sets of wavelet coefficients which can be generated by a given mother wavelet. Define Γ as the set of all sequences $g(u, 2^j)$ having the same wavelet extrema as $Wf(u, 2^j)$ at each scale. That is, if U_{e_j} is the set of time indices of all local extrema at scale j , then $Wf(u, 2^j) = g(u, 2^j)$, $\forall u \in U_{e_j}$. The approximation, $\hat{f}(t)$, is obtained through a series of alternating projections as shown in Figure 2.12. This algorithm will converge to an element of $\Lambda = V \cap \Gamma$. A function is not uniquely defined by its wavelet extrema, and exact reconstruction is not possible. However, the reconstruction error is typically confined to the highest frequency features of the original signal [Mallat (1998)]. Many papers deal with the completeness and stability of extrema representations as well as more efficient and reliable algorithms than this alternating projection algorithm [Mallat (1998), Carmona (1995),

Cvetković and Vetterli (1995), Bervman and Baras (1993)].

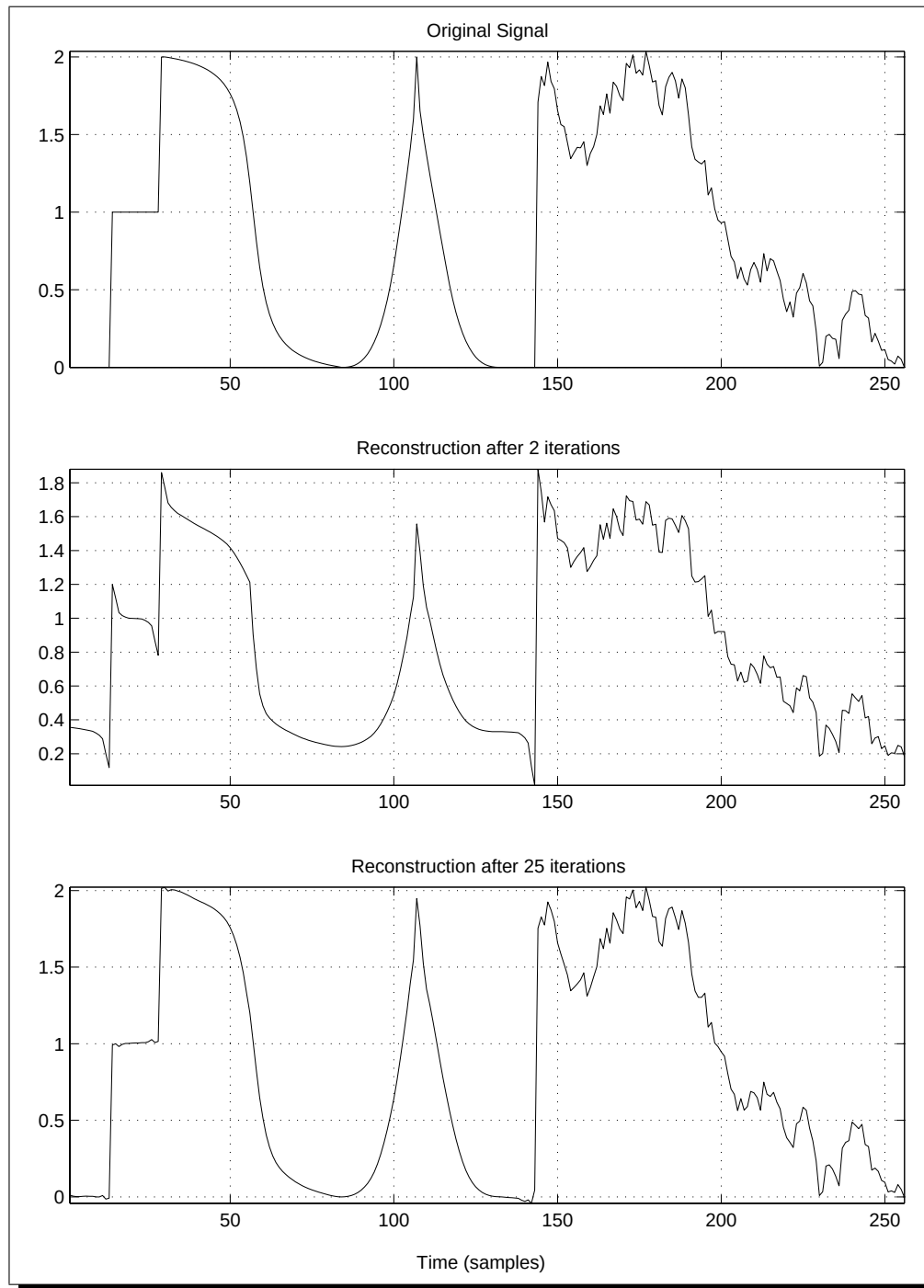
Figure 2.13 shows an example of Wavelet Transform extrema reconstruction. The first plot is the original test signal containing both smooth and sharp variations. The second plot shows the reconstruction after 2 iterations of the alternating projection algorithm, while the third plot shows the reconstruction after 25 iterations. Figure 2.14 shows the Signal-to-Noise Ratio (SNR) of the individual wavelet scales as well as the reconstructed signal. In the top plot, the highest SNR values correspond to the coarsest scales while the lowest SNR values correspond to the finest scales. Figures 2.13 and 2.14 both indicate that the reconstruction does indeed closely approximate the original signal to within a high frequency deviation. Figure 2.15 shows an example of wavelet extrema reconstruction for a speech signal. Again, the reconstruction was quite successful, proving to be both visually and audibly similar to the original speech. Figure 2.16 shows the SNR of the individual wavelet scales as well as the reconstructed signal. Again, the highest SNR values correspond to the coarsest scales while the lowest SNR values correspond to the finest scales.

Motivation for using Wavelets for Speech Enhancement

Wavelet transform extrema have been used in the past for pitch estimation of speech signals degraded by additive noise [Kadambe and Boudreaux-Bartels (1992), Kadambe (1992)]. These methods detected the points of glottal excitations by locating large Wavelet Transform coefficients (using simple thresholding procedures) in scales where the pitch information was expected to occur. The distance between consecutive excitations was then used as the pitch estimate.

The Wavelet Transform will be used for dereverberation in this thesis because:

- **Event Detection** The instances of glottal closure tend to exhibit local, if not global maxima across numerous wavelet scales. In the dereverberation procedure, a deriva-

Figure 2.13: *Wavelet extrema reconstruction example*

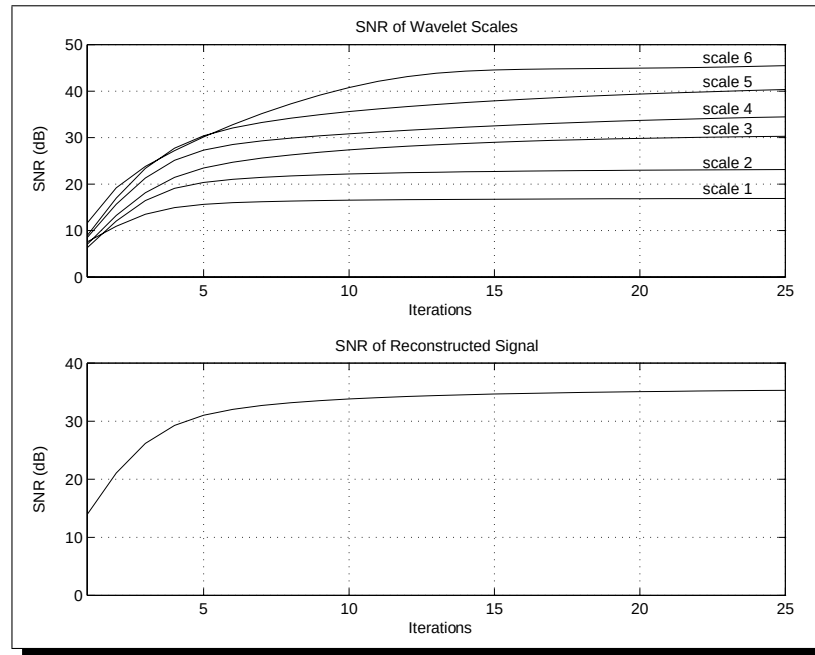


Figure 2.14: *SNR of the first six scales and the reconstructed signal for Wavelet Transform extrema reconstruction example*

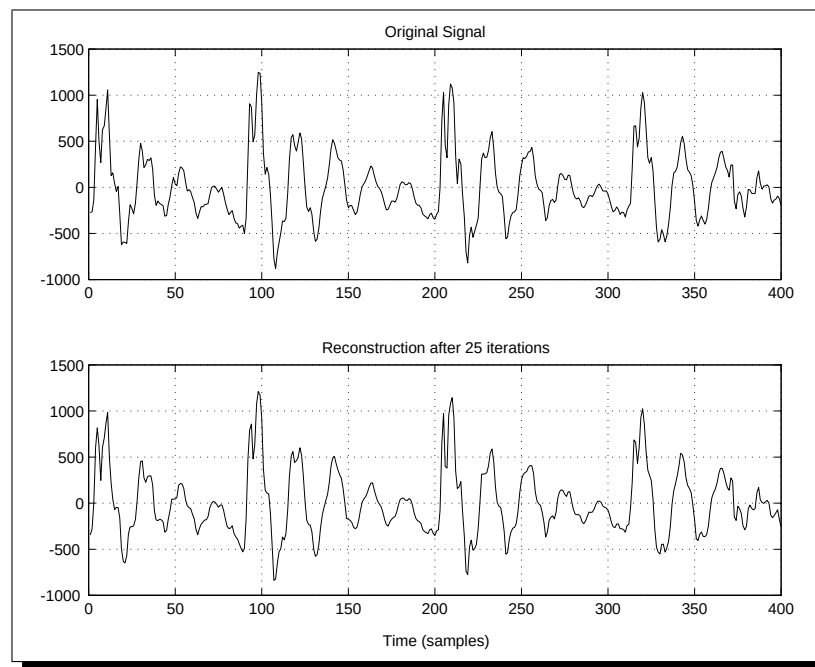


Figure 2.15: *Wavelet extrema reconstruction speech example*

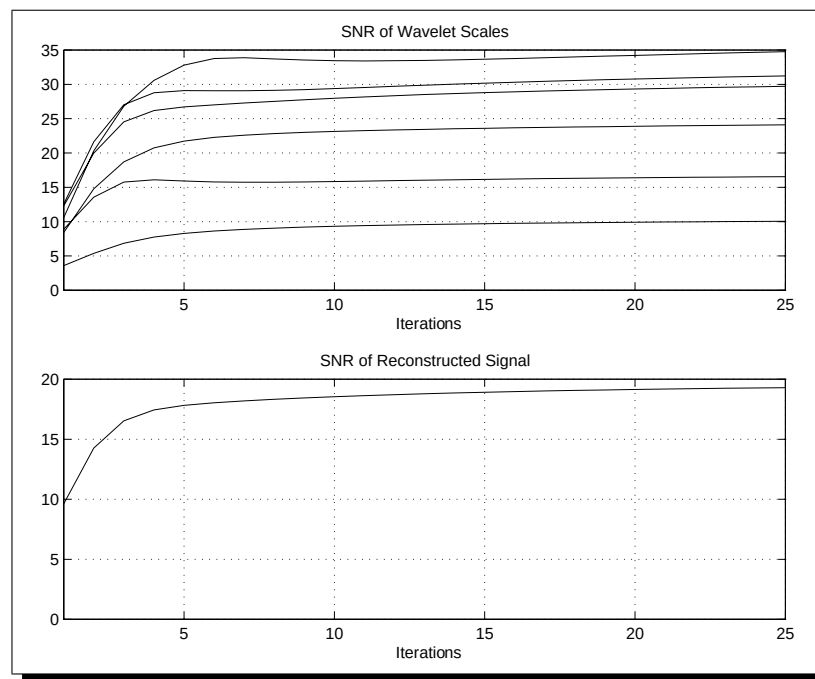


Figure 2.16: *SNR of the first six scales and reconstructed signal for Wavelet Transform extrema reconstruction example using a speech signal as input*

tive spline wavelet will be employed to emphasize and detect singularity events in the signal (i.e., instances of glottal closure and impulses due to room reverberations).

- **Extrema Reconstruction** The ability to reconstruct speech from its Wavelet Transform coefficient extrema will allow for the following procedure: Wavelet Transform calculation, followed by event detection, dampening of reverberant extrema, and finally, Wavelet Transform extrema reconstruction.

2.3 Reverberant Channel Modeling

In an ideal non-reverberant two-channel setting, a signal emanating from a distant source will typically arrive at the microphones at different times. This relative time delay is dependent on the source and microphone locations as well as the speed of sound ($c = 342$ m/s under standard conditions). The received signals at the microphones can be modeled as:

$$\begin{aligned} x_1[n] &= \alpha_1 s[n] \\ x_2[n] &= \alpha_2 s[n - \tau_{12}] \end{aligned} \tag{2.27}$$

where $s[n]$ is the source signal, α_1 and α_2 are scaling factors inversely related to the separation distance between the source and microphones, and τ_{12} is the relative time delay associated with the microphone and source geometry. In a reverberant setting, Equation 2.27 is generalized to:

$$\begin{aligned} x_1[n] &= h_1[n] * s[n] \\ x_2[n] &= h_2[n] * s[n] \end{aligned} \tag{2.28}$$

where $h_1[n]$ and $h_2[n]$ are the room impulse responses (i.e., channel responses) between the source and the two microphones (not to be confused with $h_v[n]$, the vocal tract model). Note that in the case of $h_1[n] = \alpha_1 \delta[n]$ and $h_2[n] = \alpha_2 \delta[n - \tau_{12}]$, Equation 2.28 reduces to the non-reverberant case of Equation 2.27. The extension of Equation 2.28 to $M > 2$

microphones is straightforward:

$$x_i[n] = h_i[n] * s[n] \quad (2.29)$$

at the i^{th} microphone, for $i = 1, \dots, M$.

For all simulations in this thesis, the reverberant channel responses are calculated using the Image Model method developed in [Allen and Berkley (1979)]. This is a raytracing algorithm which approximates the channel response of a rectangular enclosure by evaluating the time delay and attenuation of multi-path reflections. When calculating the channel response, $h_i[n]$, for each microphone, the reflection coefficient, $0 \leq \beta \leq 1$, of the walls, floor, and ceiling must be specified, as well as the maximum reflection order to be considered. The reflection coefficient is the fraction of the sound pressure wave that is reflected by a surface. In this thesis, varying levels of reverberation (i.e., values of β) were used, calculating the channel responses up to 10^{th} order reflections. A common parameter used to specify the amount of reverberation in a room is reverberation time. Reverberation time is the time it takes for the sound energy of an acoustic impulse to decay -40 dB from its initial energy level - a function of the room size, speed of sound in the room, and the surface reflection coefficients. Assuming all surfaces have the same reflection coefficients, the relationship between reverberation time, T_r , and the reflection coefficient is given by Eyring's formula [Kuttruff (1991)]:

$$\beta = \exp \left\{ -13.82 / [cT_r(L_x^{-1} + L_y^{-1} + L_z^{-1})] \right\} \quad (2.30)$$

where L_x , L_y , and L_z are the room dimensions. For example, for a 4 m $\times$ 4 m $\times$ 5 m room, reverberations times of 0 ms, 100 ms, 200 ms, 300 ms, 400 ms, and 500 ms correspond to β values of 0, 0.62, 0.78, 0.85, 0.89, and 0.91 respectively. To generate the different reverberant speech signals, a single-channel clean speech signal is filtered by each of the calculated channel responses.

Rating	Speech Quality	Level of Distortion
5	excellent	imperceptible
4	good	just perceptible but not annoying
3	fair	perceptible and slightly annoying
2	poor	annoying but not objectionable
1	unsatisfactory	very annoying and objectionable

Table 2.2: *Mean Opinion Score scoring system*

2.4 Quantitative Speech Quality Measures

In order to evaluate the performance of the speech enhancement algorithms presented in this thesis, some useful measures of speech quality will need to be employed. All methods for evaluating speech enhancement algorithms are either subjective or objective in nature. Subjective listening tests, though time consuming and requiring careful calibration, are widely accepted as the ideal measure of speech quality. A popular subjective measure is the Mean Opinion Score (MOS). The MOS scale is shown in Table 2.2. It consists of 5 different levels of speech quality, as perceived by the listener [Deller et al. (1987)]. Such numerical subjective scores allow for easy comparison with objective speech quality measures. However, the score does not allow for a distinction between different types of distortion.

Because human listening tests are very time consuming, the goal is to find a mathematical metric which is highly indicative of the quality of an enhanced speech signal as perceived by a wide range of listeners. In this section, the Segmental Signal-to-Noise Ratio and the Bark Spectral Distortion measures, two common quantitative speech quality assessment tools, will be discussed.

2.4.1 Segmental Signal-to-Noise Ratio

Suppose a clean speech signal, $s[n]$, is distorted and then enhanced to obtain $\hat{s}[n]$. The SNR is given by:

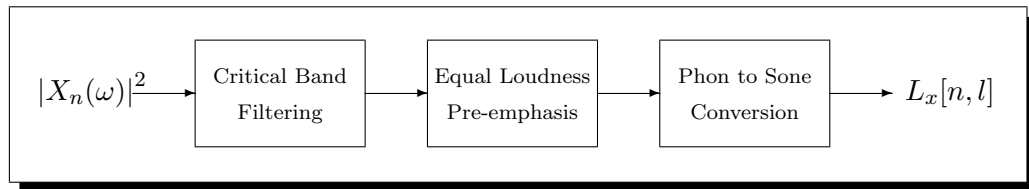
$$\text{SNR} \{s, \hat{s}\} = 10 \log_{10} \frac{\sum_n s^2[n]}{\sum_n [s[n] - \hat{s}[n]]^2} \quad (2.31)$$

which is measured in decibels. This SNR measure examines the entire speech signal and results in higher scores when the input waveforms are very similar. However, it is possible for two speech signals to sound similar even if their waveforms are very different. This explains the observation that this simple SNR measure correlates very poorly with MOS scores (correlation coefficient $\rho = 0.24$) [Wang et al. (1992), Quackenbush et al. (1988)].

The SNR metric tends to treat all time domain differences between the two signals the same, regardless of the short-term energy values which are time-varying [Deller et al. (1987)]. The Segmental SNR (SEGSNR) measure calculates the average SNR for all speech frames:

$$\text{SEGSNR} \{s, \hat{s}\} = \frac{1}{F} \sum_{i=0}^{F-1} 10 \log_{10} \left(\frac{\sum_{n=n_i-L+1}^{n_i} s^2[n]}{\sum_n [s[n] - \hat{s}[n]]^2} \right) \quad (2.32)$$

where F is the number of (typically overlapping) frames, L is the length of the each frame in samples, and n_i is the time index of the i^{th} frame endpoint. By evaluating the SNR on a frame-by-frame basis, equal weighting is given to areas of different amplitudes - a desirable result. However, in the event that a speech frame is silent, very large SEGSNR values will result, biasing the overall measure. Therefore, if silence intervals are possible, a silence detector should be implemented to exclude the corresponding frame SNR values (or simply limit them with a threshold). The SEGSNR measure has a correlation with MOS scores of $\rho = 0.77$ [Wang et al. (1992), Quackenbush et al. (1988)]. This is noticeably higher than other metrics such as the commonly used Linear Prediction Coefficient (LPC) distance ($\rho = 0.68$) and cepstral distance ($\rho = 0.63$) metrics [Wang et al. (1992), Quackenbush et al. (1988)] and will therefore be used for the eval-

Figure 2.17: *Block diagram of Bark Spectrum calculation*

uation of the proposed enhancement system.

2.4.2 Bark Spectral Distortion

Although the SEGSR metric offers a higher correlation to MOS scores than its non-segmental counterpart, it is still based on waveform matching rather than on knowledge about human perception. The Bark Spectral Distortion measure (BSD) is one such perceptually motivated metric [Wang et al. (1992)]. Studies have shown that the correlation coefficient of this measure with MOS scores is above $\rho = 0.90$ [Wang et al. (1992), Deller et al. (1987)]. The BSD metric makes use of the Bark Spectra, L_s and $L_{\hat{s}}$, of the original clean speech signal, $s[n]$, and the enhanced speech signal, $\hat{s}[n]$. The Bark Spectrum for any signal $x[n]$ is calculated by going through the various steps in Figure 2.17. The input to this process is the magnitude squared spectrum for the current analysis frame centered about the time index n , denoted $|X_n(\omega)|^2$. The output is $L_x[n, l]$ where the index l denotes frequency measured in sones. The resulting BSD score for a speech signal is the weighted average of the BSD scores for all of the analysis frames.

The first stage of the BSD calculation is Critical Band Filtering. This filtering process is based on the experimental observation that the human hearing system filters an input sound with a bank of filters with varying center frequencies and bandwidths. As frequency increases, the distances between center frequencies and therefore the resulting bandwidths

increase [Wang et al. (1992)]. Note that this is very similar to the time-frequency resolution of the standard Wavelet Transform filter bank system. Each critical bandwidth is determined experimentally by the resolution of the human hearing system. The first step of this filtering process is to transform the Fourier spectrum measured in Hertz to the Bark scale using the relation [Wang et al. (1992)]:

$$z = 13 \arctan(.76f/\text{kHz}) + 3.5 \arctan\left(\frac{f}{7.5\text{kHz}}\right)^2 \quad (2.33)$$

where f is the Fourier frequency in Hertz and z is the critical band rate measured in Barks. The Bark scale is essentially frequency divided by its critical bandwidth, resulting in the critical band rate. A Bark is also related to the distance upon the basilar membrane for which maximal stimulation is achieved for a given frequency [Zwicker and Fastl (1990)]. The experimentally measured center frequencies, critical bandwidths, and associated Bark frequency are shown in Table 2.3 [Zwicker and Fastl (1990)]. The approximation of the Hertz-to-Bark transformation of Equation 2.33 is plotted along with the experimental values of Table 2.3 in Figure 2.18 for up to 8000 Hz.

After the Bark scale transformation, the Critical Band Filtering stage is completed by convolving the transformed spectrum with the Critical Band Filter, $F(b)$, which is given by [Wang et al. (1992)]:

$$10 \log_{10} F(b) = 7 - 7.5(b - 0.215) - 17.5 [0.196 + (b - 0.215)^2]^{1/2}. \quad (2.34)$$

This filter is plotted in Figure 2.19. This spreading of the Bark scale input simulates the performance of the auditory nerve in the human hearing system. The output of this convolution is known as the “excitation pattern” [Wang et al. (1992)].

The next stage in the calculation of the Bark Spectrum is the application of equal loudness curves. The perception of the human hearing system is dependent on both the frequency and intensity of a sound waveform. Figure 2.20, taken from [Pohlmann (1995)],

z Bark	Band Limit Hertz	Center Freq. Hertz	z Bark	Band- width Hertz	z Bark	Band Limit Hertz	Center Freq. Hertz	z Bark	Band- width Hertz
0	0				12	1720			
		50	0.5	100			1850	12.5	280
1	100				13	2000			
		150	1.5	100			2150	13.5	320
2	200				14	2320			
		250	2.5	100			2500	14.5	380
3	300				15	2700			
		350	3.5	100			2900	15.5	450
4	400				16	3150			
		450	4.5	110			3400	16.5	550
5	510				17	3700			
		570	5.5	120			4000	17.5	700
6	630				18	4400			
		700	6.5	140			4800	18.5	900
7	770				19	5300			
		840	7.5	150			5800	19.5	1100
8	920				20	6400			
		1000	8.5	160			7000	20.5	1300
9	1080				21	7700			
		1170	9.5	190			8500	21.5	1800
10	1270				22	9500			
		1370	10.5	210			10500	22.5	2500
11	1480				23	12000			
		1600	11.5	240			13500	23.5	3500
12	1720				24	15500			
		1850	12.5	280					

Table 2.3: *Relationship between critical band rate z , critical band limits, and critical band center frequencies*

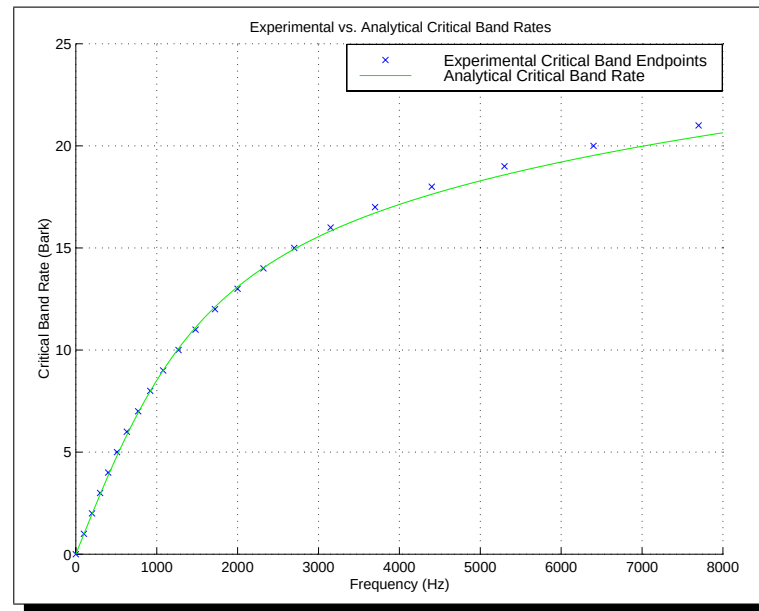


Figure 2.18: *Hertz-to-Bark scale transformation for the approximation and experimental values*

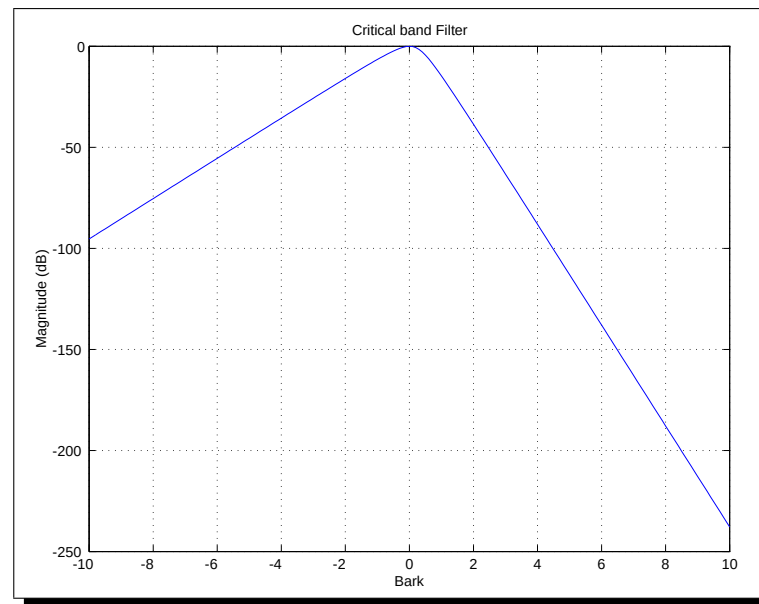


Figure 2.19: *Critical band filter*

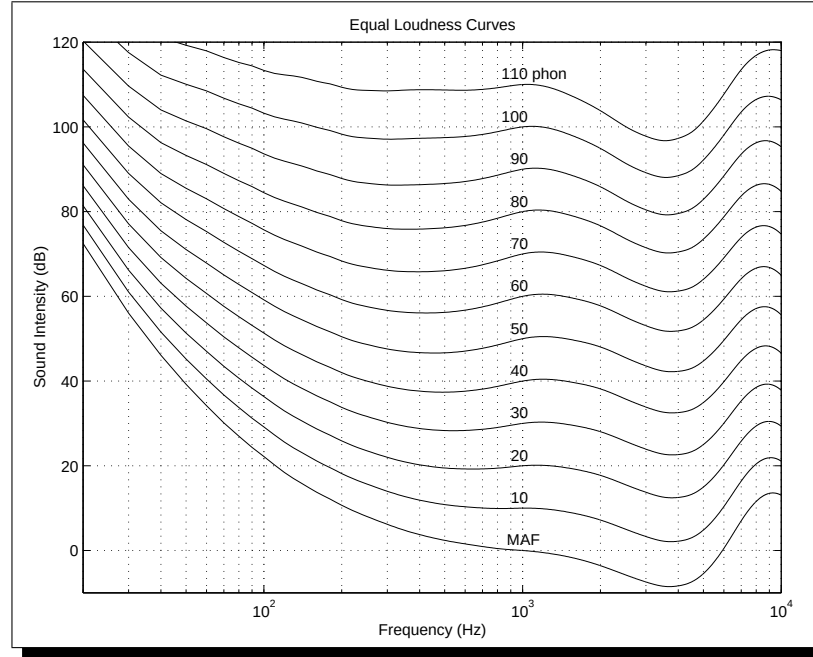
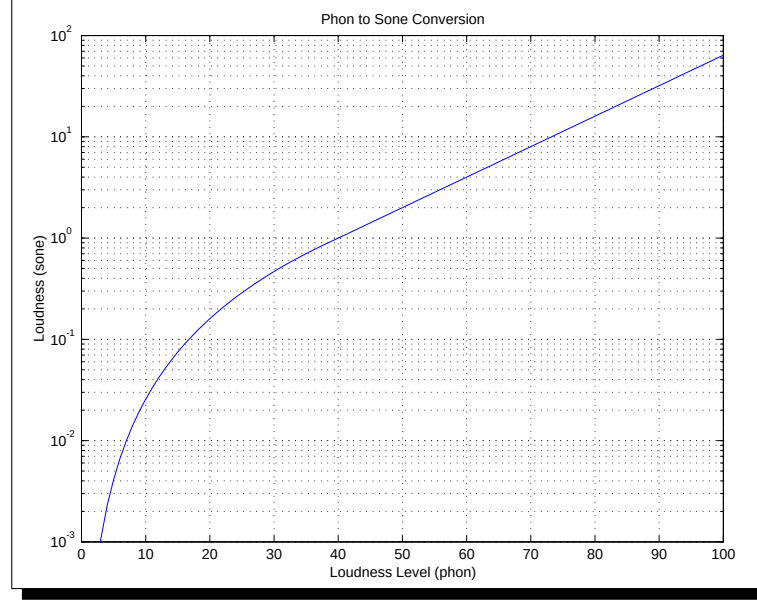


Figure 2.20: *Equal loudness curves*. *MAF* = *Minimum Audible Field*

shows the equal loudness curves on an intensity vs. frequency plot. Also shown on the plot are the values in phons corresponding to each level curve. The frequencies were pure tones played for durations longer than 500 ms [Zwicker and Fastl (1990)]. For a given tone with specified intensity and frequency, a phon is defined as the intensity in dB of a 1 kHz tone which sounds equally loud [Wang et al. (1992)]. Therefore, focusing on the plot at 1 kHz, the phon values are equal to the sound pressure. Note that a low frequency sound, in general, needs to be produced at a higher intensity than a higher frequency sound in order for the two tones to be perceived as having equal loudness. Because of these perception characteristics, the excitation pattern output of the Critical Band Filtering stage must be converted to phons according to these loudness curves.

The final stage of the Bark Spectrum calculation is the conversion from phons to sones. A sone is the increase in power which doubles the subjective loudness [Wang et al. (1992)].

Figure 2.21: *Conversion from phon scale to sone scale*

The conversion is performed because the increase in phons needed to double the perceived level varies with the loudness level. With the sones scale, the increase needed to double the perceived level is constant. The conversion factor from phons (p) to sones (l) is [Wang et al. (1992)]:

$$l = \begin{cases} (p/40)^{2.642} & p < 40 \\ 2^{(p-40)/10} & p \geq 40 \end{cases} \quad (2.35)$$

which is plotted in Figure 2.21. However, phon levels are typically greater than 40 since normal listening levels are typically at least 55 dB above the speech threshold [Wang et al. (1992)], so only the $p \geq 40$ part of Equation 2.35 will be utilized.

The output of the phon-to-sone conversion is known as the Bark Spectrum. The BSD makes use of the Bark Spectra, L_s and $L_{\hat{s}}$, of the original clean speech signal, $s[n]$, and the enhanced speech signal, $\hat{s}[n]$. The resulting BSD score for two speech signals is given

by [Adcock (2001)]:

$$BSD = \frac{\frac{1}{N} \sum_{n=1}^N \sum_l |L_s(n, l) - L_{\hat{s}}(n, l)|^2}{\frac{1}{N} \sum_{n=1}^N \sum_l |L_s(n, l)|^2} \quad (2.36)$$

where N is the number of frames and the normalization ensures that the measure is invariant to input level.

The next three chapters deal with the different stages of the proposed speech enhancement system: source localization, background noise reduction, and model-based dereverberation. Each of these chapters will make use of the concepts and equations presented in this chapter.

Chapter 3

Robust Multi-Channel Source Localization

Existing microphone array source localization procedures may be loosely divided into three categories: steered beamformers, high-resolution techniques, and time-difference of arrival (TDOA) procedures. The details of these different techniques have been explored in detail in [Brandstein and Silverman (1997)]. While beamforming has been used extensively in speech-array applications for voice capture, it has rarely been applied to the talker localization problem. The task of computing the steered output power for an appropriately dense set of candidate locations is computationally expensive and highly dependent upon the spectral content of the source signal. High-resolution approaches suffer from a lack of robustness to sensor modeling and environmental conditions, especially reverberations, and have seen little practical use for this application. With the third localization strategy, a two-step procedure is adopted. TDOA estimates of the speech signals relative to pairs of spatially separated microphones are obtained and then applied to generate a source location estimate. For examples, refer to [Huang et al. (2000), Brandstein et al. (1997), Abel and Smith (1989)].

A number of variations on this principle have been developed. They differ considerably in the method of derivation, the extent of their applicability (i.e., 2-D vs. 3-D, near source vs. distant source, etc.), and their means of solution. Primarily because of their computational practicality and reasonable performance under amicable conditions, the bulk of passive talker localization systems in use today are TDOA-based.

Accurate and robust time-delay estimation (TDE) is the key to the effectiveness of localizers within the TDOA-based genre. The two major sources of signal degradation which complicate this estimation problem are environmental background noise (i.e., additive noise) and channel multipath noise due to room reverberations (i.e., convolutional noise). The background noise case has been addressed at length, and estimating the time delay from the peak of an SNR-weighted version of the Generalized Cross-Correlation (GCC) function [Knapp and Carter (1976)] has been shown to be optimal under certain noise conditions. The problem becomes more challenging once room reverberations rise above minimal levels and the SNR-weighted GCC quickly becomes unreliable [Brandstein (1999b), Champagne et al. (1996)]. One approach to dealing with channel distortions due to room reverberations in this context has been to make the GCC function more robust by intelligently choosing the frequency-dependent weightings. The Phase Transform (PHAT) is one extreme [Knapp and Carter (1976)] which has received considerable attention as the basis of speech source localization systems [Omologo and Svaizer (1997), Wang and Chu (1997), Svaizer et al. (1997)]. By weighting each component of the cross-spectrum phase equally, the PHAT de-emphasizes signal magnitude content in favor of the desired phase content in the resulting GCC function. The PHAT time delay estimate is simply the time lag value which maximizes the PHAT GCC. A novel eigenvalue-based time delay estimation algorithm was presented in [Benesty (2000)]. This method, while not a member of the GCC family of time delay estimation techniques, explicitly models room reverberations and is

well suited to real-time implementation.

The PHAT GCC and Benesty’s time delay estimator have been shown to be fairly precise in low to moderate reverberation conditions. However, with increasing reverberation levels, they are subject to anomalous behavior in which a wall reflection is confused with the direct path delay. Although the objective functions used to estimate the time delay often do have local maxima at the true time delay, they are sometimes not global maxima, and therefore lead to erroneous TDEs. These erroneous peaks can occur due to background noise effects, a combination of multiple wall reflections, or both. This leads to one or more of the time delay estimates for a microphone array to be inaccurate, and it detrimentally affects the second step in the localization procedure. Fortunately, when estimating time delays between multiple pairs of microphones in an array, if some of the time delay objective functions have incorrect global maxima, they will rarely occur across all microphone pairs in a way that would lead to a time delay vector that is *consistent*. A consistent time delay vector is defined as a set of TDEs (one TDE for each of the microphone pairs of interest) which is theoretically achievable for a certain source location given the placement of the microphones themselves. That is, the vector of time delays is realizable for a source location based on the underlying geometry constraints of the environment.

Motivated by the observation in the preceding paragraph, this chapter explores an alternative approach to acoustic source localization which modifies the two-step localization procedure. Instead, the objective functions derived from various microphone pairs are simultaneously examined for a set of potential delay combinations consistent with candidate source locations. The result is a procedure that combines the advantages offered by the PHAT GCC or Benesty approach (or any reasonable time delay objective function) and this more robust localization procedure without dramatically increasing computational load. These results may be viewed as a more general case of the SRP-PHAT algorithm described

in [DiBiase et al. (2001)] and the CSP technique in [Nishiura et al. (2000)].

The rest of this chapter presents some typical methods for TDE-based localization, the details of the proposed robust localization method, simultaneous maximum likelihood time delay estimation, and the results of simulations and real room experiments comparing these methods.

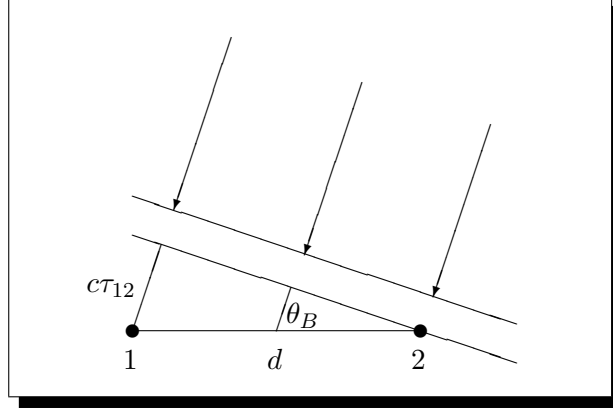
3.1 Traditional TDE-based Localization

3.1.1 The Mathematics of Time Delay Estimation

In a two-channel setting, a signal emanating from a distant source will arrive at one microphone at a slightly different time than at the second microphone (assuming that the two microphones are not equidistant from the source). This time delay is dependent on the source and microphone locations as well as the signal propagation speed in the environment. Figure 3.1 shows the setup for a typical time delay estimation problem. Often, for mathematical tractability, the source is assumed to be in the farfield so that the acoustic wave's propagation can be modeled as a plane wave. This is in contrast to the nearfield source where the propagation waves are spherical in nature. For farfield sources, the bearing angle, θ_B , which is measured as shown in Figure 3.1, can be estimated from:

$$\theta_B = \cos^{-1} \left(\frac{c\tau_{12}}{d} \right) \quad (3.1)$$

where c is the speed of sound and d is the distance between the microphones. When using multiple microphone pairs, the resulting set of bearing angles may be used to estimate the source location via triangulation methods based on the intersection points of the bearing lines [Brandstein and Silverman (1997)]. The quality of these bearing angle and location estimates is heavily dependent on the accuracy of the individual time delays.

Figure 3.1: *Two-channel time delay estimation problem*

In a non-reverberant enclosure with background noise, the signals received at two microphones, $x_1(t)$ and $x_2(t)$, may be modeled as:

$$\begin{aligned} x_1(t) &= \alpha_1 s(t - \tau_1) + u_1(t) & 0 \leq t \leq T \\ x_2(t) &= \alpha_2 s(t - \tau_2) + u_2(t) \end{aligned} \quad (3.2)$$

where $s(t)$ is the original speech utterance, τ_1 and τ_2 are the time delays of interest, α_1 and α_2 represent the scaling effects of the medium, and $u_1(t)$ and $u_2(t)$ correspond to additive noise signals uncorrelated with $s(t)$ and with each other. T is the effective time domain support of the signal. This is similar to the model presented in Chapter 2 but is expressed in continuous time and has the more realistic inclusion of background noise. Also, we define the relative time delay (i.e., relative to microphone 1) to be $\tau_{12} = \tau_1 - \tau_2$. In order to model the realistic TDE problem in a reverberant environment, the room impulse responses between the source and each microphone must be included. Thus, the signal received at two microphones, $x_1(t)$ and $x_2(t)$, may be modeled as:

$$\begin{aligned} x_1(t) &= h_1(t) * s(t) + u_1(t) & 0 \leq t \leq T \\ x_2(t) &= h_2(t) * s(t) + u_2(t) \end{aligned} \quad (3.3)$$

where $h_1(t)$ and $h_2(t)$ represent the impulse responses of the reverberant channels and $*$ denotes linear convolution. Note that the non-reverberant case results when using $h_1(t) = \alpha_1\delta(t - \tau_1)$ and $h_2(t) = \alpha_2\delta(t - \tau_2)$. Equation 3.3 is easily extended to multiple channels. In the reverberation case, the room channel response can be modeled as a sum of impulse functions [Allen and Berkley (1979)], the earliest corresponding to the time delay required for the source to reach the microphone along the direct path. With this understanding, it is clear that τ_{12} represents the time difference between these direct path energy peaks in the two impulse responses. The remaining later impulses correspond to the reflections of the source off of the ceiling, floor, various walls, and items in the room. Each of these reflections is scaled by the reflection coefficients of the surfaces encountered during the reflective path to the microphone. As is apparent from the model in Equation 3.3, a reverberant setting can and does significantly complicate the time delay estimation problem [Champagne et al. (1996)].

Given a set of P TDEs derived from the P microphone pairs of interest, the goal is to estimate the true source location, $\mathbf{q}_s$, which is a vector containing the x , y , and z coordinates of the source. Ideally, the estimate will be an element of the intersection of all the estimated bearing lines. In practice, however, this intersection is, in general, the empty set. With no ideal solution available, the source location must be estimated as the point in space which best fits the TDE data or more specifically, minimizes an error criterion that is a function of the given data and hypothesized source locations. If the TDE at each microphone pair is assumed to be independently corrupted by zero-mean additive white Gaussian noise of equal variance, then the maximum likelihood (ML) location estimate can be shown to be the position which minimizes the least squares error

criterion [Brandstein and Silverman (1997)]:

$$E_{LS}(\mathbf{q}) = \sum_{i=1}^P (\hat{\tau}_i - T(\{\mathbf{p}_{i1}, \mathbf{p}_{i2}\}, \mathbf{q}))^2 \quad (3.4)$$

where $\hat{\tau}_i$ is the TDE associated with the i^{th} pair of microphones with spatial coordinates denoted by the 3-element vectors $\mathbf{p}_{i1}$ and $\mathbf{p}_{i2}$, and $T(\{\mathbf{p}_{i1}, \mathbf{p}_{i2}\}, \mathbf{q})$ is the true TDE for a signal at source location $\mathbf{q}$ relative to the i^{th} microphone pair.

The least squares source location estimate, $\hat{\mathbf{q}}_{LS}$, is obtained from:

$$\hat{\mathbf{q}}_{LS} = \underset{\mathbf{q} \in Q}{\operatorname{argmin}} E_{LS}(\mathbf{q}) \quad (3.5)$$

where Q is the set of all candidate source locations. Note that this procedure involves the optimization of a non-linear function and necessitates the use of search methods. From a robustness and complexity standpoint this search is significantly simpler than that required by beamsteering approaches.

3.1.2 Generalized Cross-Correlation Time Delay Estimation

An intuitive method for estimating the time delay between two microphones is to find the value of τ which maximizes the cross-correlation function:

$$R_{x_1x_2}(\tau) = E[x_1(t)x_2^*(t - \tau)]. \quad (3.6)$$

The cross-power spectrum is defined as the Fourier transform of the cross-correlation function:

$$G_{x_1x_2}(\omega) = \int_{-\infty}^{\infty} R_{x_1x_2}(\tau) e^{-j\omega\tau} d\tau. \quad (3.7)$$

The Generalized Cross-Correlation (GCC) function for a given time-lag, $\hat{R}_{x_1x_2}(\tau)$, is calculated as the inverse Fourier transform of the received signal cross-spectrum, $X_1(\omega)X_2^*(\omega)$, scaled by a weighting function, $W(\omega)$:

$$\hat{R}_{x_1x_2}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} W(\omega) X_1(\omega) X_2^*(\omega) e^{j\omega\tau} d\omega. \quad (3.8)$$

The delay estimate, $\hat{\tau}_{12}$, corresponds to the lag value which maximizes $\hat{R}_{x_1x_2}(\tau)$. The standard cross-correlation (XC) function is obtained when the weighting function is uniform, i.e. $W(\omega) = W_{XC}(\omega) = 1$.

In a practical setting, the cross-power spectrum is typically estimated adaptively. That is, the cross-power spectrum for the current frame, n , is a weighted average of the estimate from the previous frame, $n - 1$, and the new cross-power information [Youn et al. (1986)]:

$$\hat{G}_{x_1x_2}^n(\omega) = \gamma \hat{G}_{x_1x_2}^{n-1}(\omega) + (1 - \gamma) X_1(\omega) X_2^*(\omega) \quad (3.9)$$

where γ is typically between 0.9 and 0.99 and G^k is the cross-power spectral estimate for the k^{th} frame. This allows for a more smoothly varying estimate than if only the current frame were used. The corresponding generalized cross-correlation function estimate for the current frame n , $\hat{R}_{x_1x_2}^n(\tau)$, is simply the inverse Fourier Transform of the cross-power spectrum estimate for that frame, $\hat{G}_{x_1x_2}^n(\omega)$. The resulting time delay estimate for the n^{th} frame, $\hat{\tau}_{12}^n$, is then:

$$\hat{\tau}_{12}^n = \underset{\tau}{\operatorname{argmax}} \hat{R}_{x_1x_2}^n(\tau) = \underset{\tau}{\operatorname{argmax}} \frac{1}{2\pi} \int_{-\infty}^{\infty} W(\omega) \hat{G}_{x_1x_2}^n(\omega) e^{j\omega\tau} d\omega. \quad (3.10)$$

The Phase Transform

For the case where the additive noise is Gaussian in nature and uncorrelated across channels, an optimal maximum likelihood (ML) weighting function derived from a frequency-dependent coherence measure is available [Knapp and Carter (1976)]. However, this ML estimator of the delay falls apart dramatically in the presence of even simple reverberations. An alternative weighting function leading in the Phase Transform [Knapp and Carter (1976), Piersol (1981)] is particularly robust to reverberations as shown in [Brandstein (1997), Omologo and Svaizer (1997), Svaizer et al. (1997), Champagne et al. (1996)]. The Phase

Transform method uses the weighting function:

$$W_{PHAT}(\omega) = \frac{1}{|G_{x_1x_2}(\omega)|} \quad (3.11)$$

to remove the effects of spectral magnitudes in the GCC function and uniformly weight the spectral phases at each frequency.

Referring to the non-reverberant model of Equation 3.2 and ignoring the scaling effects of the medium, α_i , if the noise signals are uncorrelated with the source signal, the cross power spectrum of $x_1(t)$ and $x_2(t)$ can be written as:

$$G_{x_1x_2}(\omega) = G_{ss}(\omega)e^{-j\omega(\tau_1-\tau_2)} + G_{u_1u_2}(\omega). \quad (3.12)$$

In the case of uncorrelated noise signals, $G_{u_1u_2}(\omega) = 0$. Therefore, the PHAT weighting function results in a generalized cross power spectrum of:

$$\frac{G_{x_1x_2}(\omega)}{|G_{x_1x_2}(\omega)|} = e^{-j\omega(\tau_1-\tau_2)}. \quad (3.13)$$

Taking the inverse Fourier transform results in the generalized cross-correlation signal:

$$\hat{R}_{x_1x_2}^{PHAT}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-j\omega(\tau_1-\tau_2)} e^{j\omega\tau} d\omega = \delta(\tau - (\tau_1 - \tau_2)) = \delta(\tau - \tau_{12}). \quad (3.14)$$

That is, the PHAT GCC has an impulse at the true relative time delay. In practice, the observed signals, $x_1(t)$ and $x_2(t)$, are finite length signals, requiring the above correlation functions and power spectra to be estimated using standard techniques. The resulting estimation errors as well as the addition of reverberation effects (and possibly even correlated noise signals) will cause the generalized cross-correlation function to differ from the simple form of Equation 3.14. While the PHAT is effective for moderate room reverberations, it tends to amplify errors at frequencies with a poor SNR. This limits its performance in predominantly additive noise environments, especially relative to the ML weighting.

Benesty's Technique

A number of non-GCC based time delay estimators have also been developed specifically for the acoustic room environment [Brandstein (1997), Piersol (1981)]. An eigenvalue-based estimator which models room reverberations was presented in [Benesty (2000)]. Benesty noted that in the absence of additive noise,

$$x_i(t) * h_k(t) = s(t) * h_i(t) * h_k(t) = x_k(t) * h_i(t). \quad (3.15)$$

The observation that $x_i(t) * h_k(t) = x_k(t) * h_i(t)$ allows us to write the following vectorized equation:

$$\mathbf{x}^T[n] \mathbf{u} = \mathbf{x}_1^T[n] \mathbf{h}_2 - \mathbf{x}_2^T[n] \mathbf{h}_1 = \mathbf{0} \quad (3.16)$$

where the following vectors have been defined:

$$\mathbf{x}_i[n] = [x_i[n] \ x_i[n-1] \ \dots \ x_i[n-L+1]]^T \quad (3.17)$$

$$\mathbf{h}_i = [h_i[0] \ h_i[1] \ \dots \ h_i[L-1]]^T \quad (3.18)$$

$$\mathbf{x}[n] = [\mathbf{x}_1^T[n] \ \mathbf{x}_2^T[n]]^T \quad (3.19)$$

$$\mathbf{u} = [\mathbf{h}_1^T \ -\mathbf{h}_2^T]^T. \quad (3.20)$$

By multiplying each side of Equation 3.16 by $\mathbf{x}[n]$ and taking expected values, the result is:

$$E \{ \mathbf{x}[n] \mathbf{x}^T[n] \} \mathbf{u} = \mathbf{0}. \quad (3.21)$$

Now, since $\mathbf{R}[n] = E \{ \mathbf{x}[n] \mathbf{x}^T[n] \}$ is the covariance matrix of the microphone data, the vector $\mathbf{u}$ is the eigenvector of the covariance matrix corresponding to the eigenvalue equal to 0. Since $\mathbf{u}$ captures the structure of the channel responses, the time delay estimate is chosen as the difference in peak locations between the first and second half of this vector. See [Benesty (2000)] for implementation details. An unconstrained frequency-domain least-mean-square (UFLMS) adaptive filter algorithm [Mansour and Jr. (1981)] was proposed for

the simple calculation of the relative time delay. In this method, a quasi-impulse function for one channel is searched for a peak relative to a fixed, yet arbitrarily assigned direct path impulse for the other channel. This method has been shown to be effective for low reverberation conditions.

3.2 Localization Using Consistent Delay Vectors

The traditional two-step procedure to source localization involves the estimation of time delays followed by a source localization algorithm. Often, these techniques involve a pruning or de-emphasis of possibly errant time delays, assuming they can be detected reliably, before estimating the source location. For the source locator proposed in this chapter, the two steps are combined into a single search for a vector of time delay estimates, $\hat{\boldsymbol{\tau}}$, where each estimate is *consistent* with the others in the sense that they all satisfy the physical constraints imposed by the microphone locations and hypothesized source locations. By simultaneously examining the functions used to estimate time delays (e.g., GCC functions) over all microphone pairs of interest, this localization technique is much more effective in challenging environments.

For $M > 2$ channels, the problem is to estimate a consistent delay vector $\hat{\boldsymbol{\tau}}$. That is, the bearing lines for all microphone pairs for which a time delay is estimated all intersect at one point in space. Traditional two-channel techniques are applied to the microphone pairs of interest independently, with a resulting delay vector that is not necessarily a member of the set of consistent delay vectors. To perform bearing angle estimation, one time delay estimate from the array is typically chosen and the implied bearing angle is calculated. Alternatively, one could calculate the implied delay vector for a grid of possible source locations. The resulting source location estimate is the grid position with the least squared error between its implied delay vector and the vector of estimated delays, according to

Equation 3.4. The resulting bearing angle is then the line from the microphone array center to the estimated source location.

Figure 3.2 illustrates the shortcomings of the traditional TDE-based localization scheme. It shows a set of bearing lines derived from the TDEs obtained when using the PHAT GCC independently for an array of 8 microphones which are indicated by $\times$. A time delay was estimated for each pair of neighboring microphones. For this example, the reverberation time was 400 ms and there was additive white noise with an SNR of 15 dB, a challenging situation. The source, indicated by $\square$, was located at $(x, y) = (1.85 \text{ m}, 2.75 \text{ m})$. As is apparent in the plot, two bearing lines appear to be focusing on first order wall reflections rather than the desired source. Without judicious pruning of these two delay estimates in the least squares error criterion, the resulting traditional source location estimate, $\hat{\mathbf{q}}_{LS}$, will be significantly biased away from its true location. Even the bearing angles which are not focusing on reverberations coming off of the left wall exhibit a certain amount of aiming error. This is due to the slight smearing of the global peak located at the true delay that can result when speech is degraded by room reverberations.

Figure 3.3 gives some insight into the cause of the errant bearing lines. It shows the PHAT GCC function for the rightmost pair of microphones in Figure 3.2. The true delay is -3.01 samples, but the interpolated peak in the PHAT GCC occurs at -11.2 samples (sampling rate = 20 kHz). This leads to a bearing line estimate of 163° which is well to the left of the true bearing line. A local maximum does occur in the PHAT GCC near the true time delay, but it is not the global maximum and is therefore not chosen as the TDE. As this plot suggests, reducing the information in the cross-correlation function down to a single parameter, the TDE, prior to estimating the source location is problematic. Rather than selectively culling or weighting TDEs in an error criterion [Brandstein and Silverman (1997), Di Claudio et al. (2000)], it is advantageous to work directly with the GCC functions them-

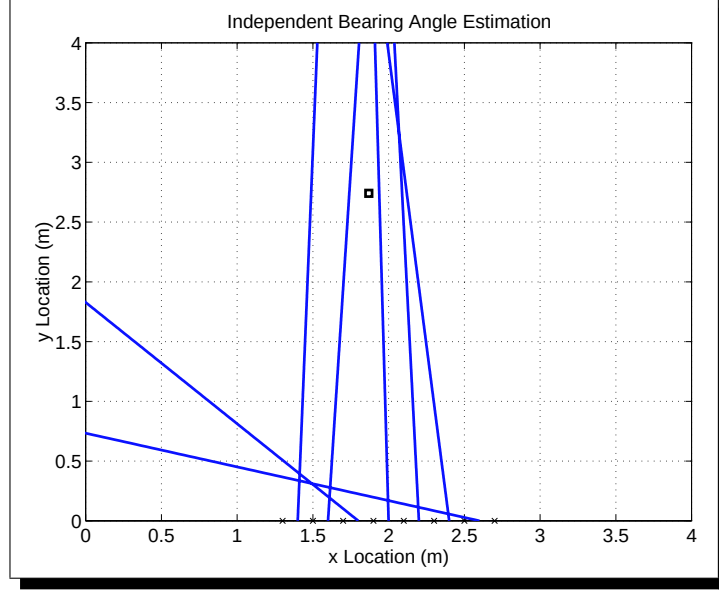


Figure 3.2: *Bearing angle estimates derived from independent PHAT GCC time delay estimates. The true source location is indicated by $\square$ and the 8 microphones are indicated by $\times$.*

selves to generate a delay vector, τ_{con} , which is consistent with a physical source location.

If we define the consistent energy metric E_{CON} as:

$$E_{\text{CON}}(\mathbf{q}) = \sum_{i=1}^P \hat{R}_{x_{\mathbf{p}_{i1}}, x_{\mathbf{p}_{i2}}} (T(\{\mathbf{p}_{i1}, \mathbf{p}_{i2}\}, \mathbf{q}))^\nu \quad (3.22)$$

where ν is an exponential weighting term (typically $\nu = 1$ or 2). The consistent source location estimate, $\hat{\mathbf{q}}_{\text{con}}$, can be estimated from:

$$\hat{\mathbf{q}}_{\text{CON}} = \underset{\mathbf{q} \in Q}{\operatorname{argmax}} E_{\text{CON}}(\mathbf{q}). \quad (3.23)$$

The corresponding consistent time delay estimates are calculated using:

$$\hat{\tau}_{\mathbf{p}_{i1}, \mathbf{p}_{i2}} = T(\{\mathbf{p}_{i1}, \mathbf{p}_{i2}\}, \hat{\mathbf{q}}_{\text{CON}}). \quad (3.24)$$

With this structure, any objective function for which the maximum theoretically occurs at the true time delay may be substituted for $\hat{R}_{x_{\mathbf{p}_{i1}}, x_{\mathbf{p}_{i2}}}$ in Equation 3.22, including a

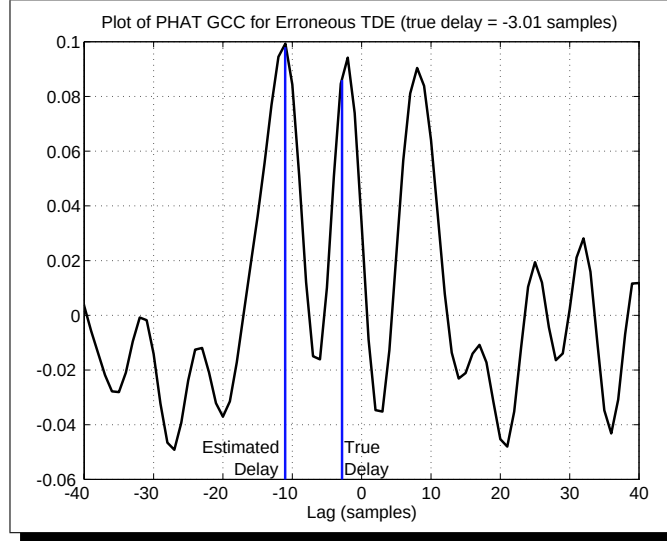


Figure 3.3: Plot of the PHAT GCC for the rightmost microphone pair in Figure 3.2. The true delay is -3.01 samples, while the estimated delay is -11.2 samples.

function appropriate for Benesty’s eigenvalue-based TDE. Note that in later results using the generalized cross-correlation, frequency bins outside the typical range of speech (i.e., 50-4000 Hz) are excluded to reduce the effects of noise on the resulting time delay estimates.

Figure 3.4 shows bearing lines that were achieved by using the consistent PHAT GCC estimator and they therefore intersect at one (x, y) location in the room. Note that in this example, because the array aperture is small relative to the source range, it is difficult to accurately estimate the talker’s range. This is a classic linear array processing result. In fact, for two microphones, the variance of the range error is proportional to the fourth power of the range of the source [Carter (1981)]. However, with this geometry, it is still possible to accurately obtain the source bearing angle. One way to provide accurate bearing and range estimates is to distribute the microphones throughout the enclosure rather than along a single wall [Brandstein et al. (1996)].

A less general frequency domain algorithm similar to this consistent PHAT estimator

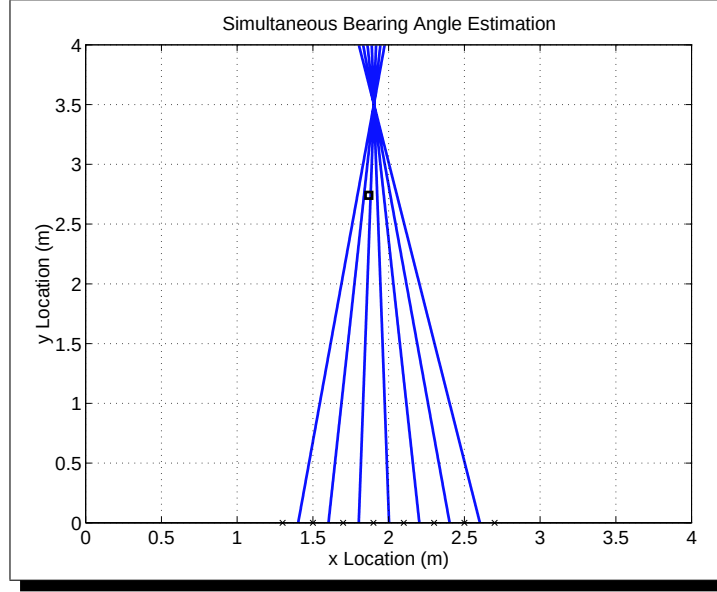


Figure 3.4: *Bearing angle estimates for consistent PHAT GCC time delay estimation. The true source location is indicated by $\square$ and the 8 microphones are indicated by $\times$.*

was presented in [DiBiase (2000)]. DiBiase used all pairwise combinations of microphones in the array. In the consistent time delay estimator described above, only the set of all neighboring microphones is used (i.e., horizontally or vertically connected). This leads to 4 pairs ($P = 4$) for a 2×2 array of 4 microphones and 10 pairs ($P = 10$) for a 2×4 array of 8 microphones. This is faster than using all possible pairwise combinations, which would be 6 pairs for 4 microphones and 28 pairs for 8 microphones. Also, microphones that are far apart tend to produce less reliable estimates than those located near each other so in some cases, using all pairwise combinations of microphones is neither efficient nor desirable. Another less general technique was presented in [Nishiura et al. (2000)] which assumed that all microphone pairs are centered around the same location. This theoretically leads to a peak in the sum of the individual cross-correlations at the correct time delay.

3.3 Consistent Multi-Channel Maximum Likelihood Time Delay Estimation

In this section, a novel maximum likelihood (ML) technique is developed to simultaneously estimate a vector of time delays for an array of M microphones. It will be shown that for 2 microphones, the resulting maximum likelihood estimator (MLE) of time delays reduces to the result of [Knapp and Carter (1976)]. The following derivation, though considerably more complicated, is roughly patterned after the 2 microphone, additive noise only case originally presented in [Knapp and Carter (1976)]. As discussed previously, the observed signal at the i^{th} microphone can be written as:

$$x_i(t) = s(t - \tau_i) + u_i(t) \quad i = 1, \dots, M \quad (3.25)$$

where $s(t)$ is the speech source signal, τ_i is the time delay for the source signal to arrive at the i^{th} microphone (relative to some reference time), and u_i is assumed to be zero-mean Gaussian noise with variance σ_u^2 . The noise samples are identically and independently distributed and are also independent of the source signal, s . Note that this implies a non-reverberant room and ignores the scaling effects of the environment. The goal is to estimate the true time delay vector $\boldsymbol{\tau} = [\tau_1 \dots \tau_m]$. It is more convenient to work in the frequency domain, where the observations take on the form:

$$X_i(\omega) = S(\omega)e^{-j\omega\tau_i} + U_i(\omega) \quad i = 1, \dots, M. \quad (3.26)$$

As before, due to the independence properties of the noise signal, the following equation holds:

$$G_{x_i, x_j}(\omega) = G_{s, s}(\omega)e^{-j\omega(\tau_i - \tau_j)} + G_{u_i, u_j}(\omega) \quad (3.27)$$

where

$$G_{u_i, u_j}(\omega) = \begin{cases} \sigma_u^2 & i = j \\ 0 & i \neq j. \end{cases} \quad (3.28)$$

To find the ML estimate of the time delay vector, $\boldsymbol{\tau}$, it is assumed that the speech signal, s , is Gaussian. The ML estimate of $\boldsymbol{\tau}$ is that which maximizes the following probability:

$$p(\mathbf{X}(\omega)|\boldsymbol{\Sigma}(\omega)) = p(\mathbf{X}(\omega)|G_{s,s}(\omega), \sigma_u^2, \boldsymbol{\tau}) \quad (3.29)$$

where the vector $\mathbf{X}(\omega) = [X_1(\omega) \dots X_m(\omega)]^T$, and $\boldsymbol{\Sigma}$ is the frequency-dependent cross-power spectral density matrix defined by:

$$\begin{aligned} \boldsymbol{\Sigma}(\omega) &= \begin{bmatrix} E[X_1(\omega)X_1^*(\omega)] & E[X_1(\omega)X_2^*(\omega)] & \dots & E[X_1(\omega)X_M^*(\omega)] \\ E[X_2(\omega)X_1^*(\omega)] & E[X_2(\omega)X_2^*(\omega)] & \dots & E[X_2(\omega)X_M^*(\omega)] \\ \vdots & \vdots & \ddots & \vdots \\ E[X_M(\omega)X_1^*(\omega)] & E[X_M(\omega)X_2^*(\omega)] & \dots & E[X_M(\omega)X_M^*(\omega)] \end{bmatrix} \\ &= \begin{bmatrix} G_{x_1,x_1}(\omega) & G_{x_1,x_2}(\omega) & \dots & G_{x_1,x_M}(\omega) \\ G_{x_1,x_2}^*(\omega) & G_{x_2,x_2}(\omega) & \dots & G_{x_2,x_M}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ G_{x_1,x_M}^*(\omega) & G_{x_2,x_M}^*(\omega) & \dots & G_{x_M,x_M}(\omega) \end{bmatrix}. \end{aligned}$$

By using Equation 3.27, this matrix becomes:

$$\boldsymbol{\Sigma}(\omega) = \begin{bmatrix} G_{s,s}(\omega) + \sigma_u^2 & G_{s,s}(\omega)e^{-j\omega(\tau_1-\tau_2)} & \dots & G_{s,s}(\omega)e^{-j\omega(\tau_1-\tau_M)} \\ G_{s,s}(\omega)e^{j\omega(\tau_1-\tau_2)} & G_{s,s}(\omega) + \sigma_u^2 & \dots & G_{s,s}(\omega)e^{-j\omega(\tau_2-\tau_M)} \\ \vdots & \vdots & \ddots & \vdots \\ G_{s,s}(\omega)e^{j\omega(\tau_1-\tau_M)} & G_{s,s}(\omega)e^{j\omega(\tau_2-\tau_M)} & \dots & G_{s,s}(\omega) + \sigma_u^2 \end{bmatrix}. \quad (3.30)$$

This can be rewritten as $\boldsymbol{\Sigma}(\omega) = \boldsymbol{\Sigma}_x(\omega) + G_{s,s}(\omega)\boldsymbol{\Sigma}_\tau(\omega)$ where

$$\boldsymbol{\Sigma}_x(\omega) = (G_{s,s}(\omega) + \sigma_u^2)\mathbf{I} \quad (3.31)$$

and

$$\mathbf{\Sigma}_{\boldsymbol{\tau}}(\omega) = \begin{bmatrix} 0 & e^{-j\omega(\tau_1-\tau_2)} & \dots & e^{-j\omega(\tau_1-\tau_M)} \\ e^{j\omega(\tau_1-\tau_2)} & 0 & \dots & e^{-j\omega(\tau_2-\tau_M)} \\ \vdots & \vdots & \ddots & \vdots \\ e^{j\omega(\tau_1-\tau_M)} & e^{j\omega(\tau_2-\tau_M)} & \dots & 0 \end{bmatrix}. \quad (3.32)$$

Given the definition of the matrix $\mathbf{\Sigma}$ and the assumption that s is Gaussian, the optimal delay vector $\boldsymbol{\tau}$ maximizes:

$$p(\mathbf{X}|\mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^M |\mathbf{\Sigma}(\omega)|}} \exp \left\{ -\frac{1}{2} \int_{-\infty}^{\infty} \mathbf{X}^H(\omega) \mathbf{\Sigma}^{-1}(\omega) \mathbf{X}(\omega) d\omega \right\} \quad (3.33)$$

where H represents the conjugate transpose operation, and $|\mathbf{\Sigma}(\omega)|$ is the determinant of $\mathbf{\Sigma}(\omega)$. Equivalently, the log-likelihood function can be maximized:

$$\ln\{p(\mathbf{X}|\mathbf{\Sigma})\} = k - \frac{1}{2} \ln |\mathbf{\Sigma}(\omega)| - \frac{1}{2} \int_{-\infty}^{\infty} \mathbf{X}^H(\omega) \mathbf{\Sigma}^{-1}(\omega) \mathbf{X}(\omega) d\omega \quad (3.34)$$

where k is a constant not involving $\boldsymbol{\tau}$ and can therefore be ignored.

The resulting maximum likelihood time delay estimator, $\hat{\boldsymbol{\tau}}_{ML}$, for a microphone array with M sensors is:

$$\hat{\boldsymbol{\tau}}_{ML} = \underset{\boldsymbol{\tau} \in \mathbb{R}^M}{\operatorname{argmax}} \sum_{i=1}^M \sum_{j=1, j \neq i}^M \left[\int_{-\infty}^{\infty} \frac{(\sigma_u^2)^{M-2} G_{s,s}(\omega)}{|\mathbf{\Sigma}_M(\omega)|} X_i(\omega) X_j^*(\omega) e^{j\omega(\tau_i-\tau_j)} d\omega \right], \quad (3.35)$$

the derivation of which is presented in Appendix A. That is, the ML estimator is the sum of $\binom{M}{2}$ generalized cross-correlation functions between all pairwise combinations of microphones, each using the same weighting function:

$$W(\omega) = \frac{G_{s,s}(\omega)}{(\sigma_u^2)^2 + M \sigma_u^2 G_{s,s}(\omega)}. \quad (3.36)$$

For the two channel case, $M = 2$, we have:

$$W(\omega) = \frac{G_{s,s}(\omega)}{(\sigma_u^2)^2 + 2(\sigma_u^2) G_{s,s}(\omega)}. \quad (3.37)$$

which, after some algebraic manipulation, is precisely the ML estimator derived by Knapp and Carter [Knapp and Carter (1976)]. Furthermore, the consistent TDE technique can be applied to this simultaneous ML estimation technique. That is, in Equation 3.35, the search for optimal delay vectors can be restricted to $\boldsymbol{\tau} \in T_{CON}$ where T_{CON} is the set of all consistent time delay vectors.

This ML time delay estimation technique is theoretically optimal provided the assumptions it makes are true. Therefore, each assumption deserves some consideration:

1. The speech utterance, $s(t)$, is a Gaussian source. Naturally, this model is not exactly appropriate for speech signals. However, due to its mathematical tractability, it often is assumed to be true, sometimes leading to successful approaches to speech processing [Sohn and Sung (1999), Knapp and Carter (1976)].
2. The scaling effects of the sound medium (i.e., the factors α_i in Equation 3.2) were assumed to be the same for each channel. Although this is not generally true, closely spaced microphones will have very similar scaling factors, particularly for farfield sources. Also, when searching for the optimal consistent delay vector, for each candidate source location these scaling terms can be predicted and can effectively be inverted, causing the new scaling factors to be constant. This inverse scaling operation would have to be completed after subtracting the noise estimate from the speech spectrum.
3. The model assumes that the noise can be estimated, or is known *a priori*. To this end, one might make use of a speech/non-speech detector to locate non-speech intervals during which the background noise level can be estimated. However, the Minimum Statistics noise estimation technique developed by [Martin (1994), Martin (1993)] can estimate the background noise even during speech activity and obviates the need for

such speech detection algorithms. This technique will be detailed in Chapter 4 on background noise reduction. In the simplified model of Equation 3.25, the same variance is assumed for all channels, a reasonable assumption for closely spaced microphones. Theoretically, the power spectrum for additive white noise has an expected constant value equal to the noise variance. The noise can be estimated by averaging the Minimum Statistics noise spectrum estimate over all frequency bins. Averaging over all channels can make the estimate even more accurate.

4. The effects of the channel (i.e., room reverberation) were ignored in this model. The effort to analytically solve for the optimal ML consistent delay vector in the presence of room reverberations is not worthwhile, since accurate estimates of the room impulse responses, a natural requirement for the resulting ML estimate, are generally unavailable.

3.4 Practical Time Delay Estimation Systems

3.4.1 Discrete Time Delay Estimation

One practical matter of time delay estimation deals with the implied discrete nature of any consistent TDE technique. First of all, the equations discussed previously would be replaced by their discrete-time counterparts. For example, Equation 3.3 becomes:

$$\begin{aligned} x_1[n] &= h_1[n] * s[n] + u_1[n] & 0 \leq n \leq N \\ x_2[n] &= h_2[n] * s[n] + u_2[n] \end{aligned} \quad (3.38)$$

where N is the time domain support of the discrete-time signals. To arrive at this discrete formulation, the standard sampling procedure is used, where the prerequisite prefiltering and a sufficiently high sampling rate are used to avoid aliasing. Because the significant

energy of the speech spectrum lies below 4 kHz, a sampling rate above 8 kHz is typically used to avoid aliasing effects. One problem with the required discrete formulation and implementation of these TDE algorithms is that the resolution of the resulting delay estimates is limited by the sampling rate used to convert the analog signal to a digital signal. Therefore, a parabolic fit is performed about the peak in the TDE objective function to achieve sub-sample resolution. That is, suppose the maximum value of the objective function, ϕ , occurs at integer time n^{opt} so that the maximum value is $\phi[n^{opt}]$. Parabolic interpolation would estimate the new maximum to be located at time:

$$n_{interp}^{opt} = n^{opt} + \frac{1}{2} \frac{\phi[n^{opt} - 1] - \phi[n^{opt} + 1]}{\phi[n^{opt} - 1] - 2\phi[n^{opt}] + \phi[n^{opt} + 1]}. \quad (3.39)$$

3.4.2 Voice Activity Detection

Voice activity detection (VAD) is important in many speech processing applications, including time delay estimation. A survey of VAD algorithms will be presented in Chapter 4 on denoising, as VAD allows for noise-only states to be detected. For TDE algorithms, the procedure should not be allowed to update TDEs or source location estimates during non-speech moments as these estimates will be searching for the noise source location.

3.4.3 Numerical Techniques

There are numerous ways to approach the solution of the consistent time delay estimation algorithm proposed in this chapter. A straightforward, yet very inefficient method would be to perform a grid search. That is, form a grid of possible source locations, each of which implies a theoretical time delay vector. Each $\hat{R}$ function in Equation 3.22 must be interpolated to find its value at the appropriate time delay implied by the current candidate source location. The precision of this technique depends on the distance between the room grid elements. This method could be improved by using a coarse-to-fine algorithm

to zoom in on the optimal delay vector until meeting a certain iterative criterion. As an alternative, one could consider any of a number of optimization techniques such as gradient descent [Etter and Stearns (1981), van Laarhoven and Aarts (1989)]. The particular technique used would depend on knowledge about the speech environment (i.e., one vs. multiple talkers, number of noise sources, etc.). In this thesis, a coarse-to-fine approach was taken to provide the source location estimates.

3.5 Experimental Results

3.5.1 Simulations Results

Figure 3.5 shows the negative LS space from ($-E_{LS}$ from Equation 3.4) and the E_{CON} consistency function (from Equation 3.22) for a source near the array, a reverberation time of 400 ms, and a 15 dB SNR level. There were 8 microphones (2 rows of 4), represented by $\times$, with their orientation indicated in the plot. This yields a total of $P = 10$ microphone pairs for 8 microphones ($M = 8$) when considering only neighboring microphone pairs. However, in the case of the ML estimate, theory dictates that all pairwise combinations be used, yielding $P = 28$ for 8 microphones. The true source location is shown as a $\square$ in each plot, while the respective location estimates are plotted as $+$. Throughout this section, impulse responses were generated using the image model of [Allen and Berkley (1979)] with subsample interpolation [Peterson (1986)]. The microphones themselves were modeled as having a cardioid response of $0.5 + 0.5 \cos(\theta)$ where θ is the angle from the orientation vector of the microphone. Therefore, any sound arriving perpendicular to the microphone sensor from the front will be passed in its entirety, while from any other angle it would be attenuated. The cardioid response pattern is shown in Figure 3.6.

Returning to Figure 3.5, the negative least-squares error (LS) and consistency crite-

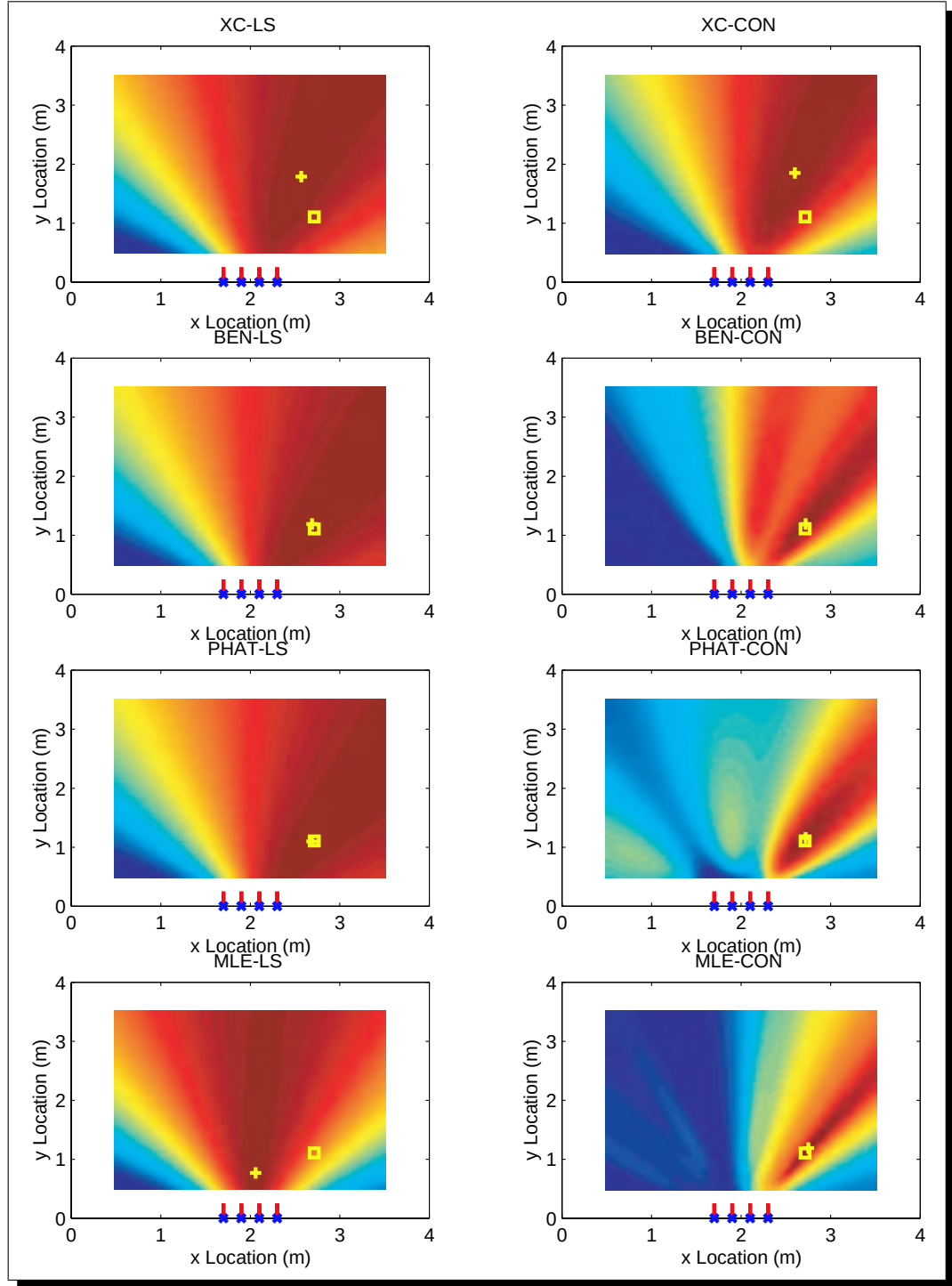


Figure 3.5: Qualitative plots of the $-E_{LS}$ and E_{CON} criteria typical of a source near the array for the four TDE methods considered. The source location is indicated by $\square$, the source location estimate by $+$, and the 8 microphones (two rows of 4) by $\times$.

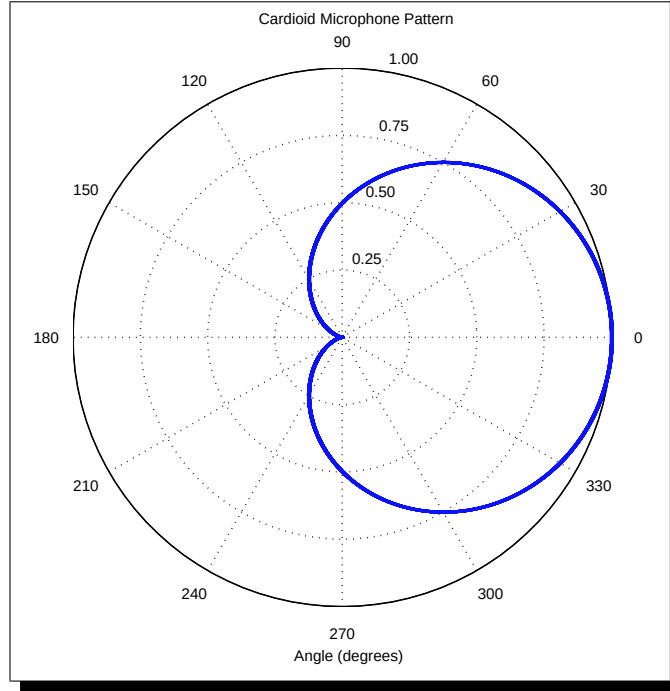


Figure 3.6: *Response pattern for a cardioid microphone*

rion (CON) are both plotted for four different TDE methods: cross-correlation (XC), Benesty's algorithm (BEN), the Phase Transform (PHAT), and the multi-channel maximum likelihood estimation (MLE) method. Red indicates areas of highest amplitude. The consistency plots in the right-hand column all show a superior ability to detect the source location relative to their LS counterparts in the left-hand column. In some cases, the source location estimate is much more accurate (e.g., MLE-CON). In all cases, the energy is more focused in the CON cases than it is in the LS cases. The PHAT-CON and MLE-CON cases exhibit the best performance. Figure 3.7 shows similar results for a source far from the array. Figure 3.8 repeats this experiment with two simultaneous talkers. As expected, the LS criterion is unable to deal with the multi-talker scenario and the extrema in the corresponding plots tend to be somewhere in the middle of the two desired locations. The XC-CON,

BEN-CON, and MLE-CON represent only mild improvements over their LS counterparts, while the PHAT-CON estimation method clearly exhibits two distinct direction of arrival estimates. Because the XC, BEN, and MLE methods all break down in moderate to high reverberation conditions, they tend to be less accurate than the relatively robust PHAT. This is due to the fact that the PHAT tends to exhibit local (if not global) maxima near the true time delay while the other methods tend to be too corrupted by reverberations for this to hold.

To quantify the utility of the proposed localization schemes, simulations were performed in a $4\text{ m} \times 4\text{ m} \times 3\text{ m}$ room using 10 randomly generated source locations (see Figure 3.9) and reverberation times ranging from 0 to 1000 ms. A speech signal 10 seconds in length was used as the original clean utterance. Eight cardioid microphones arranged as two 4-element linear arrays (i.e., a 2×4 grid) with 0.2 m spacing in both dimensions were positioned on one wall of the enclosure as shown in Figure 3.9. For the results of the cross-correlation, PHAT, and Benesty's algorithm, only horizontally or vertically neighboring microphone pairs were used to decrease the computation time. As before, impulse responses were generated using the image model with subsample interpolation. White noise resulting in an SNR of 15 dB was added to each microphone signal to make the test more challenging and realistic.

Figure 3.10 presents an analysis of the 8 localization algorithms evaluated for the 10 random source locations and varying reverberation levels with 4 (a 2×2 array) and 8 microphones (a 2×4 array). For this experiment, the line from the center of the array to the estimated source location served as the estimated bearing line, and the line from the center of the array to the true source location was the true bearing line. If the angle between these two bearing lines was less than 5° , the location estimate was considered to be correct.

As shown in the plots, the consistent versions of each algorithm always performed

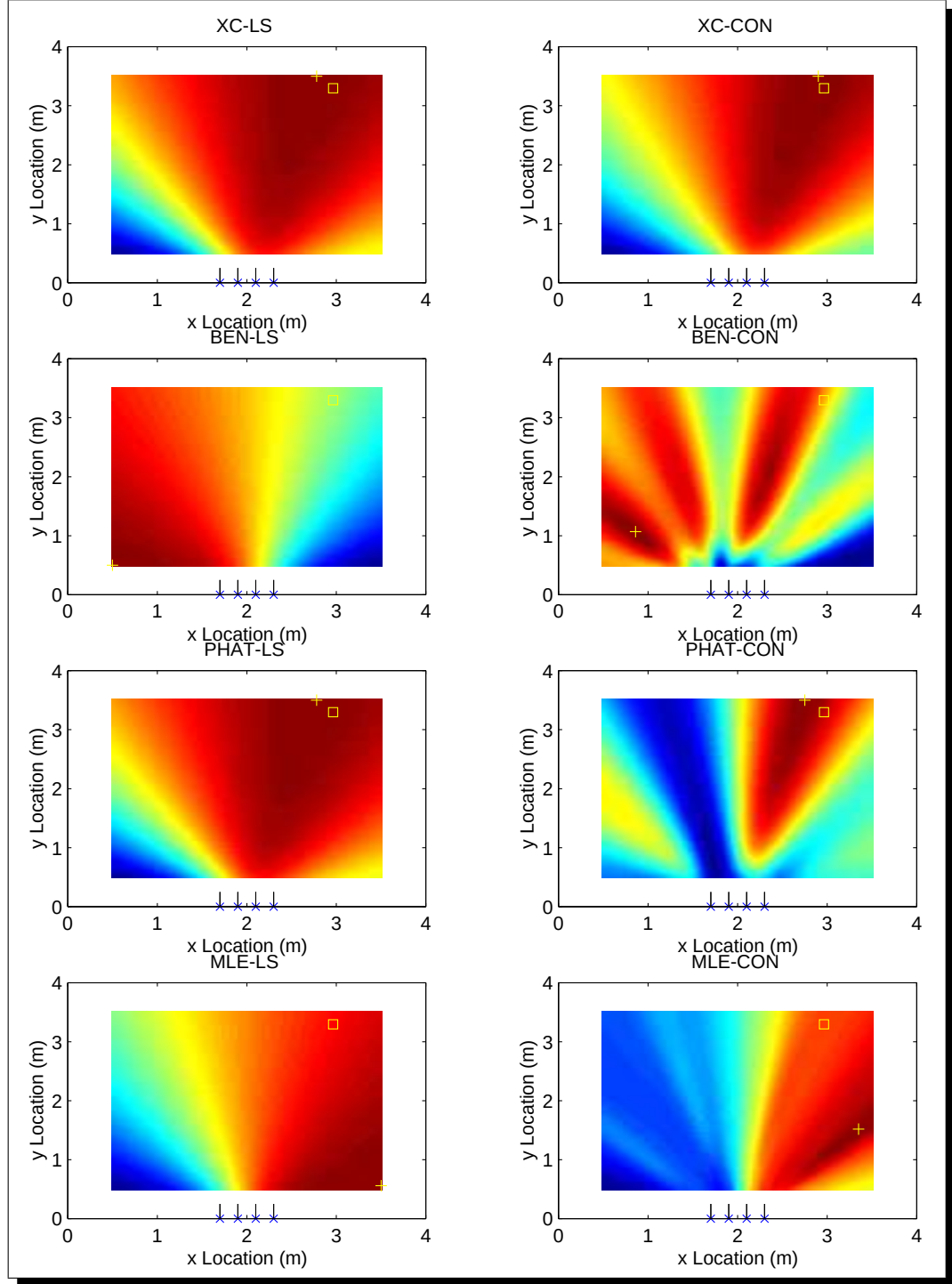


Figure 3.7: Qualitative plots of the $-E_{LS}$ and E_{CON} criteria typical of a source far from the array for the four TDE methods considered. The source location is indicated by $\square$, the source location estimate by $+$, and the 8 microphones (two rows of 4) by $\times$.

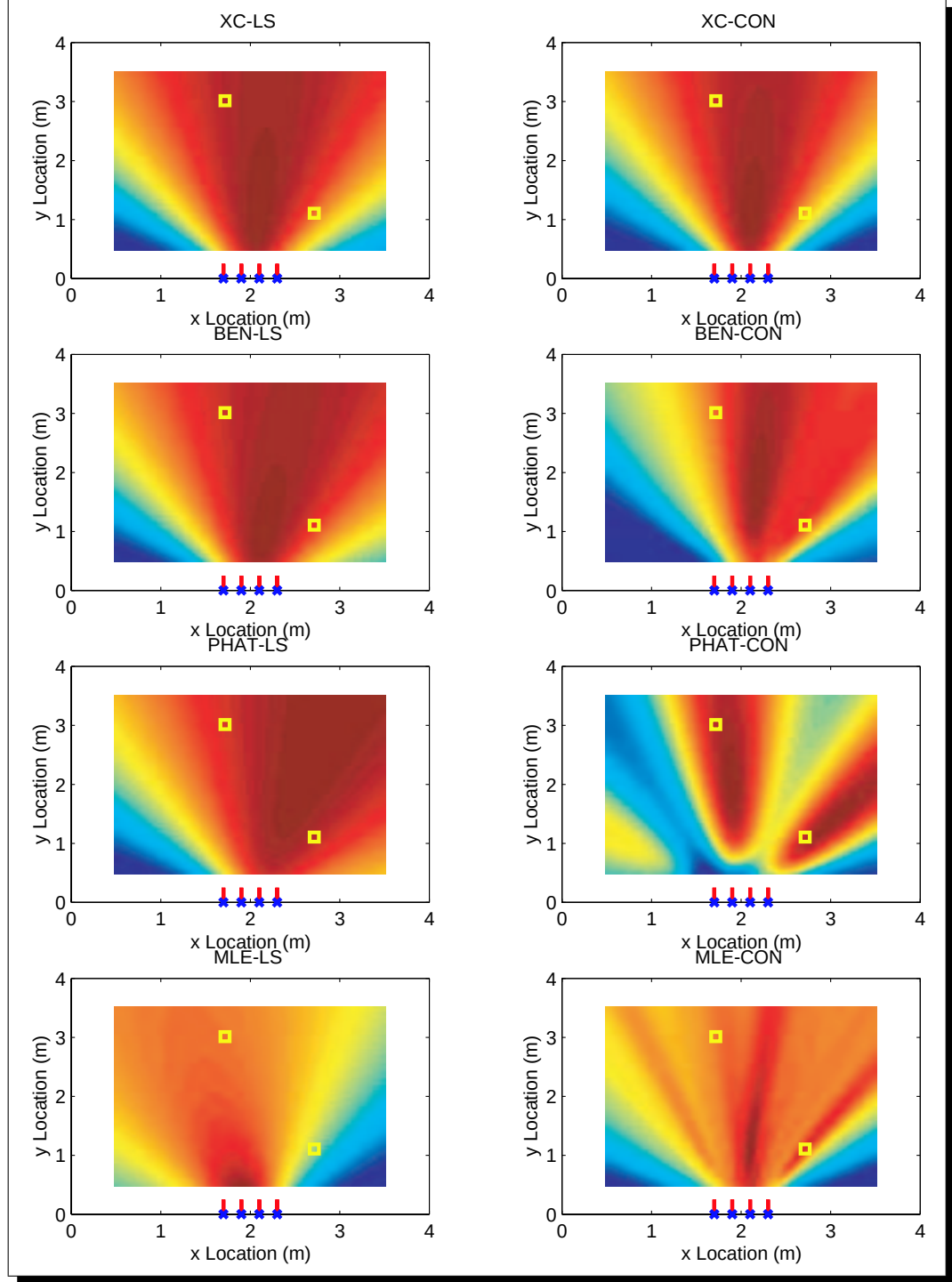


Figure 3.8: Qualitative plots of the $-E_{LS}$ and E_{CON} criteria for 2 simultaneous sources for the four TDE methods considered. The source location is indicated by $\square$ and the 8 microphones (two rows of 4) by $\times$.

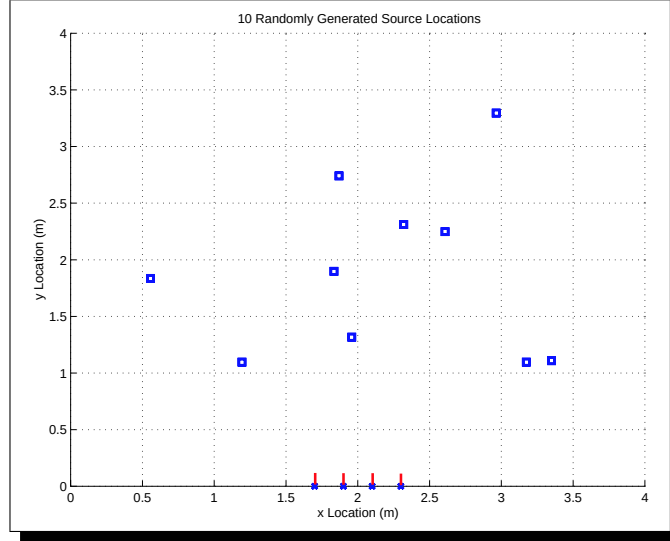


Figure 3.9: 10 random source locations, indicated by $\square$ used for TDE simulations. The microphones are positioned in two rows of 4, indicated by $\times$.

at least as well as their LS counterparts - an intuitive result. Also note that the XC function does not gain much from enforcing consistency. This suggests that when errant time delays are observed, the correlation peaks corresponding to the true delay were very small. However, the PHAT, Benesty, and ML techniques show a dramatic improvement when enforcing consistency. The trends and observations for 8 microphones are similar to those for 4 microphones, but the % accuracy was higher for all reverberation levels, as expected. As an interesting side note, if we were to randomly pick a bearing angle, the result would fall within 5° of the true source location $10/180 = 0.055$ or 5.5% of the time. At high levels of reverberation, the XC method is so corrupted by reverberation effects that it is no better than randomly picking time delay estimates.

Figure 3.11 compares the different consistent TDE algorithms to each other. It appears that the Benesty algorithm performs better for lower reverberation times (0-200 ms) while the PHAT and ML algorithms perform better for higher reverberation times (300-1000 ms).

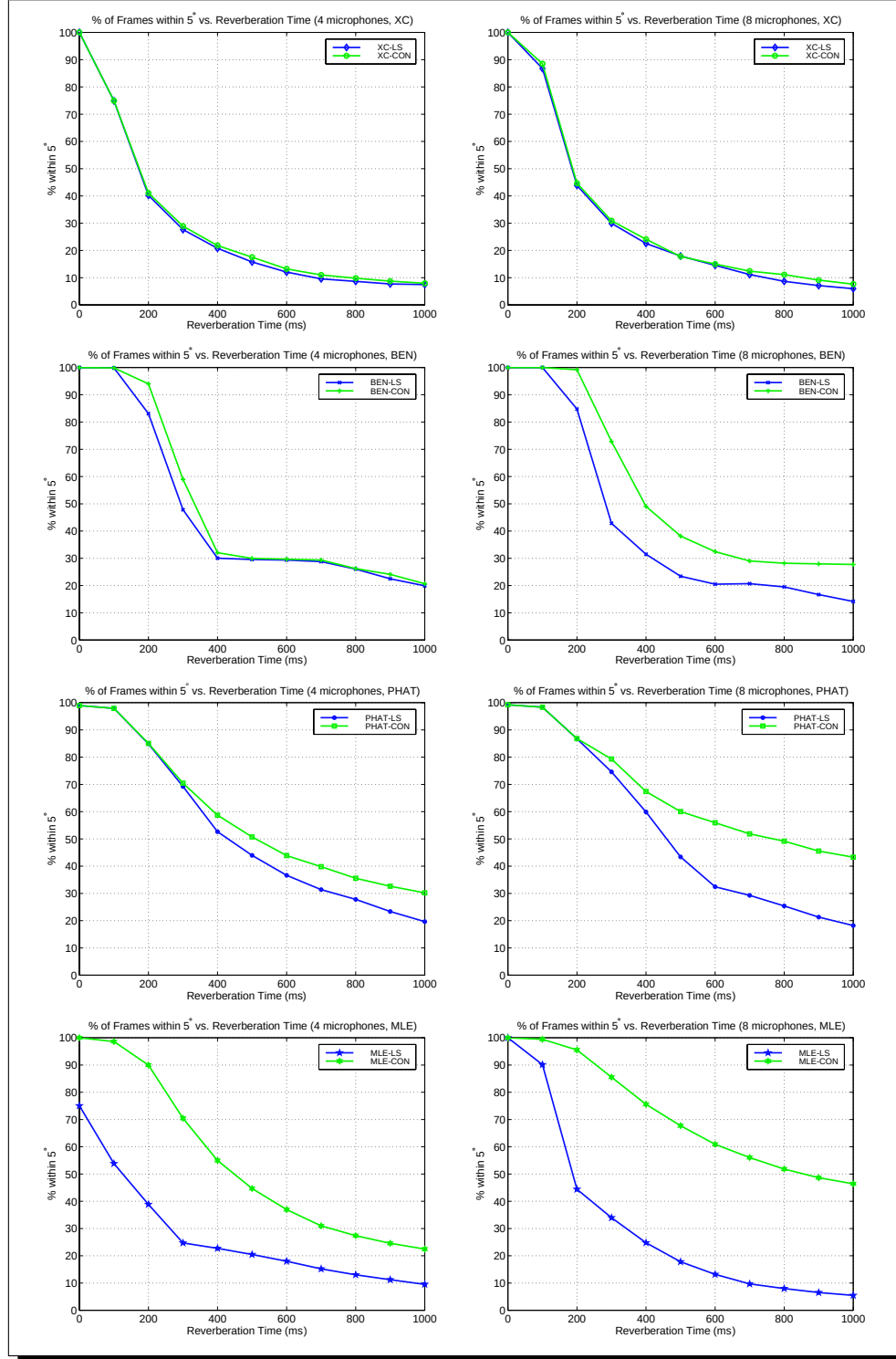


Figure 3.10: Analysis of the LS (XC-LS, BEN-LS, PHAT-LS, MLE-LS) and consistency-based (XC-CON, BEN-CON, PHAT-CON, MLE-CON) locators for various reverberation levels and 15 dB SNR additive white noise

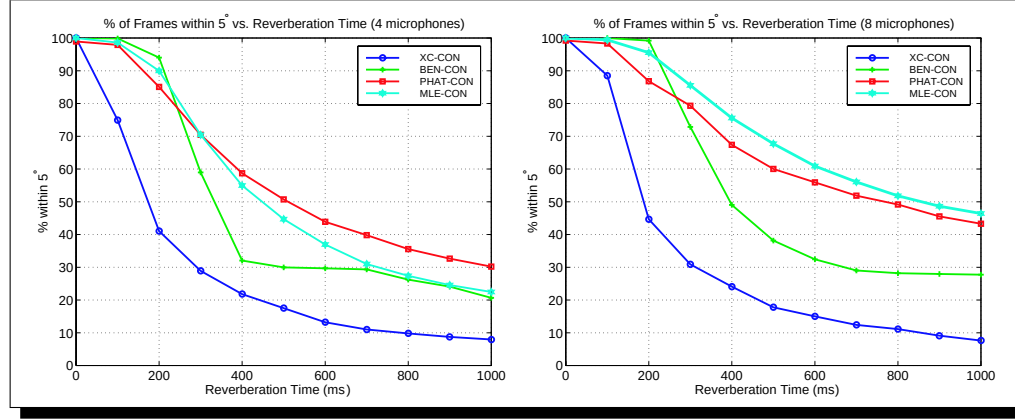


Figure 3.11: Comparison of the consistency-based (*XC-CON*, *BEN-CON*, *PHAT-CON*, *MLE-CON*) locators for various reverberation levels and 15 dB SNR additive white noise

Figure 3.12 shows how the PHAT-CON technique performs using varying values of the angle required to achieve a “correct” estimate. The point of this figure is to show that any reasonable angle could be used for this criterion with similar results.

Because the MLE-CON TDE uses all possible microphone-pair combinations, the previous set of results does not provide the fairest of comparisons. In Figure 3.13, the angle results are compared for PHAT-CON, MLE-CON, and PHAT-CON-ALL, where PHAT-CON-ALL uses all pairwise microphone combinations (as does MLE-CON) rather than just adjacent microphone pairs. These plots show that most of the improvement of the MLE-CON estimator comes from its use of additional microphone pairs. Theoretically, the MLE-CON estimator should perform best for low reverberation levels only, as reverberations were not included in the model. This does indeed seem to be the case.

3.5.2 Real Room Results

Finally, qualitative plots are presented for two real room recordings performed in an enclosure of size 5 m $\times$ 5 m $\times$ 3 m. Figures 3.14 and 3.15 show the results for two different

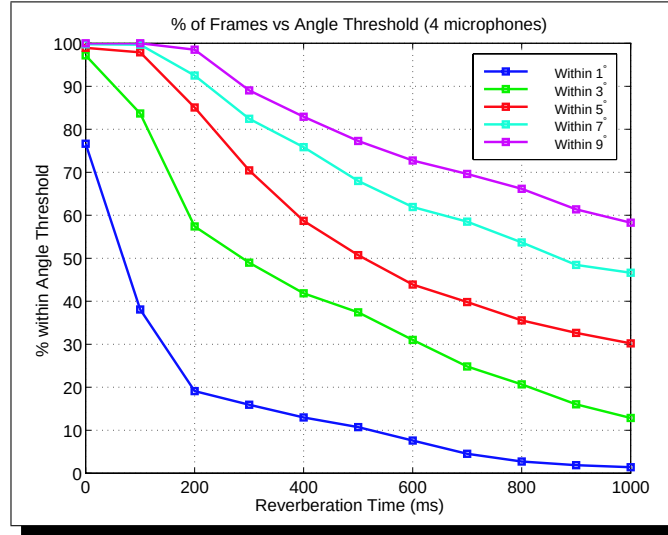


Figure 3.12: Comparison of PHAT-CON accuracy results for different angle criteria

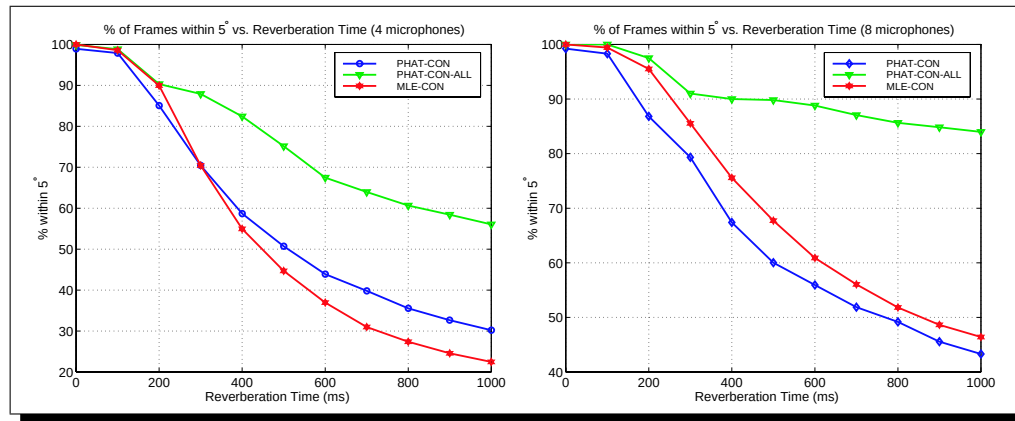


Figure 3.13: Comparison of PHAT-CON, PHAT-CON-ALL, and MLE-CON locators for various reverberation levels and 15 dB SNR additive white noise

source locations. The room contained 16 cardioid microphones arranged as four 2×2 microphone arrays positioned along the walls (as indicated in the plots) at a height of 1.5 m. Within each array, neighboring microphones had a separation of 6.5 cm. Note that in these plots, there are 16 microphone pairs (4 for each microphone quad) used in the consistent estimation for the cross-correlation, Benesty, and PHAT TDEs. However, the ML TDE uses 120 pairs. This explains the very sharp peak in the MLE-CON objective functions. Both figures show effective results for all algorithms since the reverberation level in the real room was relatively low (approximately 200 ms). However, the consistent estimators do tend to exhibit a tighter peak around the source, indicating an ability for more accurate localization when multiple sources are present.

3.6 Source Localization Summary

As the results in this chapter suggest, the traditional two-step localization procedure clearly suffers from its intermediate use of explicit time delay estimates. The consistency-based localization scheme more effectively exploits the information available in the cross-correlation function by incorporating knowledge of the physical constraints dictated by the geometry of the source and microphones.

The proposed method may be employed with any objective function where the maximum or minimum value indicates the optimal time delay estimate. The advantages of the PHAT weighting over cross-correlation and Benesty approaches, namely the PHAT's de-emphasis of signal content and robustness to reverberations, are also apparent (and enhanced) in this application. The function derived from the Benesty algorithm, while not as effective at moderate to high reverberation levels, does offer advantages under low reverberation conditions where the additive noise is the dominant source of degradation. Finally, although the ML method exhibited impressive performance, at high reverberation conditions, it was

primarily due its inclusion of more microphone pairs in the localization process. Using such a high number of microphone pairs quickly becomes impractical.

From a computational standpoint, the proposed consistent localization scheme represents a middle ground between the highly efficient closed-form localizers and the burdensome beamsteering methods. As the plots in Figure 3.5 and Figure 3.7 indicate, the search space is very amenable to iterative approaches which can be performed in an efficient manner.

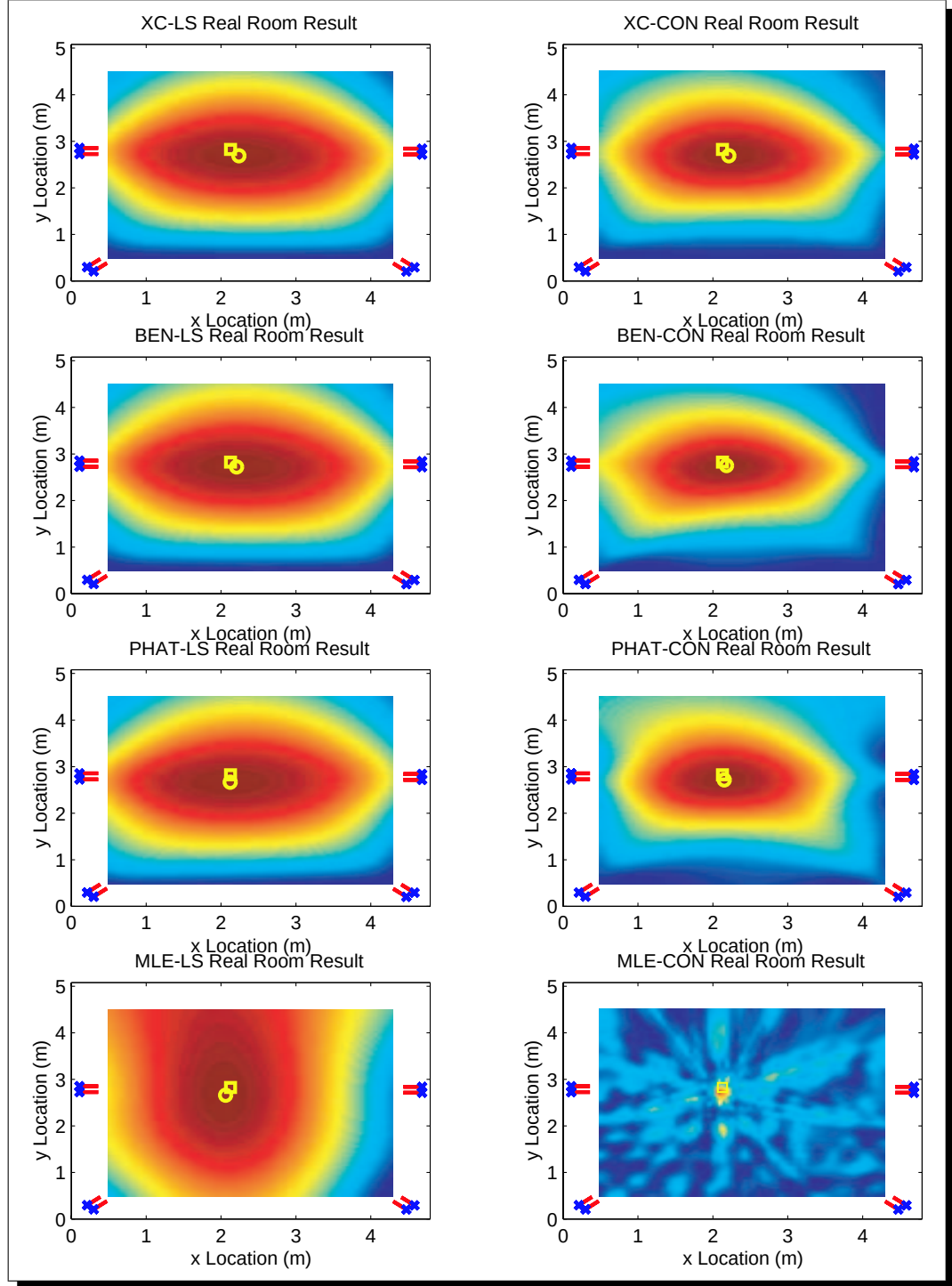


Figure 3.14: Real room plots of the $-E_{LS}$ and E_{CON} criteria. The source location is indicated by $\square$, the source location estimate by $\circ$, and the 8 microphones (two rows of 4) by $\times$

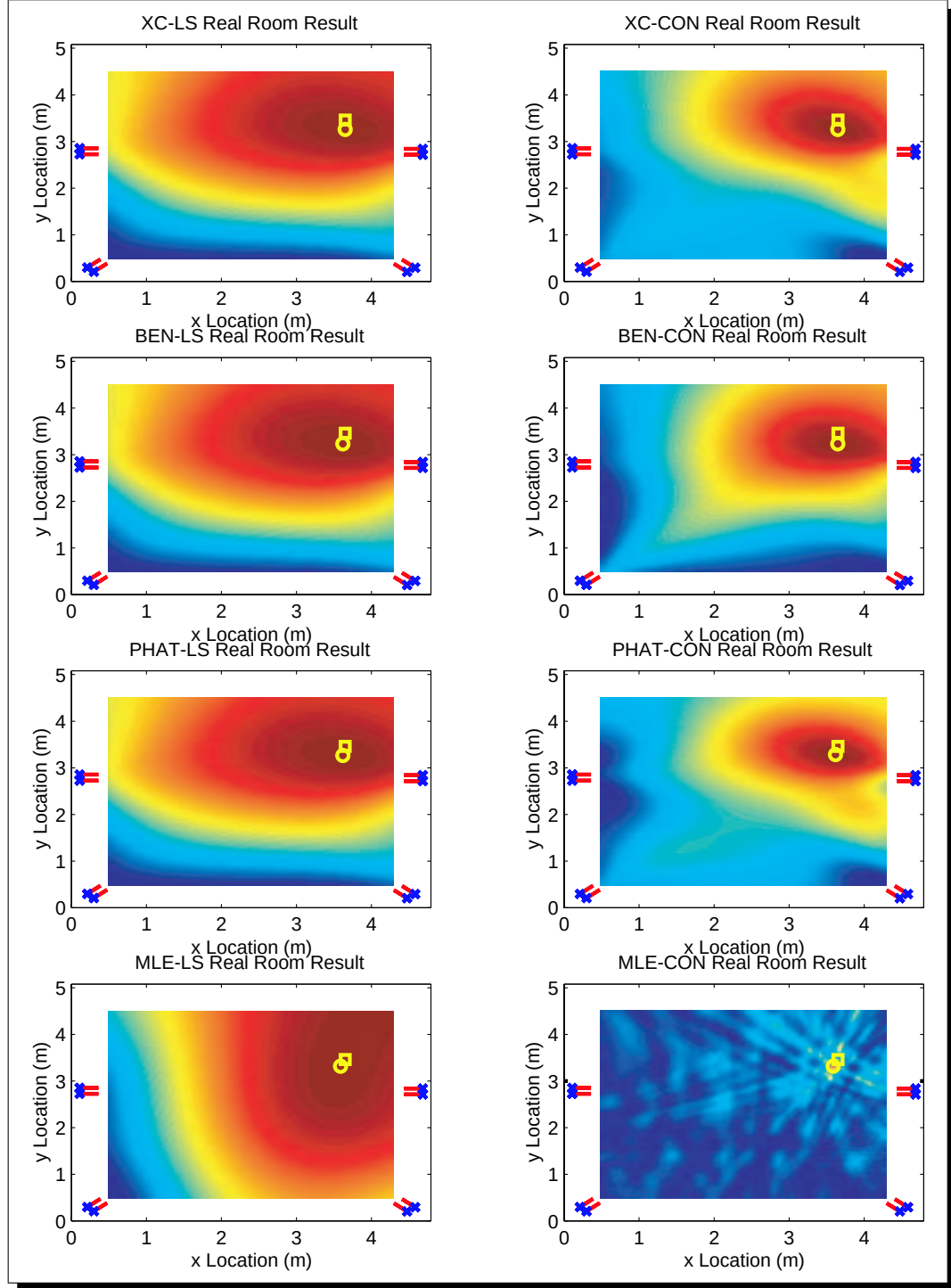


Figure 3.15: Real room qualitative plots of the $-E_{LS}$ and E_{CON} criteria. The source location is indicated by $\square$, the source location estimate by $\circ$, and the 8 microphones (two rows of 4) by $\times$

Chapter 4

Background Noise Reduction

This chapter focuses on the second stage of the proposed speech enhancement system - the reduction of background noise. A survey of both single channel and multi-channel techniques is presented, followed by the presentation of the Multi-Channel Minimum Statistics Denoising algorithm (MC-MSD). A suite of results will be presented to indicate the effectiveness of this algorithm for both simulated and real room scenarios.

4.1 Single Channel Approaches

Of the single channel techniques, spectral modification techniques such as Spectral Subtraction [Boll (1979)] are very common. This approach and its numerous variations (e.g., [McAulay and Malpass (1980)]) attempt to directly alter the STFT of the noise-degraded speech. In the case of Spectral Subtraction, this involves subtracting an estimate of the background noise spectrum from the observed short-term spectrum to obtain an estimate of the speech spectrum. Wiener filtering and its variations offer another single channel approach to additive noise reduction. These techniques minimize the mean squared difference between the enhanced speech signal and the underlying speech utterance [Quatieri (2002)]

based on an estimate of the noise power spectrum. These techniques vary in their domain of application, estimation procedure for the background noise, and various subtleties required to maintain artifact-free enhanced speech.

4.1.1 Spectral Subtraction

In a single channel setting where the original speech signal, $s[n]$, is corrupted only by additive noise, the observed signal, $x[n]$, is given by:

$$x[n] = s[n] + u[n] \quad (4.1)$$

where $u[n]$ is the background noise signal. Taking the Fourier transform results in:

$$X(\omega) = S(\omega) + U(\omega). \quad (4.2)$$

If an estimate of the background noise spectral amplitude, $|\hat{U}(\omega)|$, is available, then the enhanced speech spectrum, $\hat{S}(\omega)$, achieved by Spectral Subtraction was originally expressed by [Boll (1979)] and is given by:

$$\hat{S}(\omega) = \left[|X(\omega)| - |\hat{U}(\omega)| \right] e^{j\phi_X(\omega)} \quad (4.3)$$

where $\phi_X(\omega)$ is the phase function of the observed signal, $x[n]$. Note that only the spectral magnitude is modified. The phase is left unaltered. The Spectral Subtraction operation can also be viewed as the filtering operation:

$$\hat{S}(\omega) = X(\omega) H_{SS}(\omega) \quad (4.4)$$

where the zero-phase Spectral Subtraction filter $H_{SS}(\omega)$ is given by:

$$H_{SS}(\omega) = 1 - \frac{|\hat{U}(\omega)|}{|X(\omega)|}. \quad (4.5)$$

Note that this filter automatically adjusts the amount of attenuation at a given frequency according to the Signal-to-Noise Ratio at that frequency. Typically, in the event that

$$\left[|X(\omega)| - |\hat{U}(\omega)| \right] < 0, \quad (4.6)$$

the difference is set to 0. A well-known side effect of Spectral Subtraction is the introduction of tonal artifacts into the resulting enhanced speech signal. Many attempts have been made to modify the Spectral Subtraction procedure to reduce these artifacts, with limited success.

4.1.2 Wiener Filtering

The Wiener filter is another single channel approach to the problem of background noise reduction [Lim and Oppenheim (1978), Quatieri (2002)]. Returning to the additive noise case of Equation 4.1, suppose an optimal filter, H_W , is to be derived to minimize the mean squared error, e , between the clean speech and the filtered speech. That is, the enhanced speech signal spectrum, $\hat{S}$, will be obtained by a filtering of the observed signal spectrum, X from:

$$\hat{S}(\omega) = H_W(\omega)X(\omega) = H_W(\omega)[S(\omega) + U(\omega)]. \quad (4.7)$$

The mean squared error is then given by:

$$e = E \left[\left(\hat{S}(\omega) - S(\omega) \right)^2 \right] \quad (4.8)$$

where $E[\cdot]$ is the expectation operator. Using the definition of $\hat{S}$ in Equation 4.7, Equation 4.8 can be rewritten as:

$$e = E \left[\left(H_W(\omega)[S(\omega) + U(\omega)] - S(\omega) \right)^2 \right]. \quad (4.9)$$

Taking derivatives with respect to $H_W(\omega)$ leads to:

$$\begin{aligned} \frac{\partial e}{\partial H_W(\omega)} &= 2E \left[(H_W(\omega)[S(\omega) + U(\omega)] - S(\omega)) (S(\omega) + U(\omega)) \right] \\ &= H_W(\omega)E \left[S^2(\omega) + 2S(\omega)U(\omega) + U^2(\omega) \right] - E \left[S^2(\omega) - S(\omega)U(\omega) \right]. \end{aligned} \quad (4.10)$$

However, assuming that the speech and noise are uncorrelated, or:

$$E[S(\omega)U(\omega)] = 0 \quad (4.11)$$

and setting the expression of Equation 4.10 equal to 0 leads to:

$$\frac{\partial e}{\partial H_W(\omega)} = H_W(\omega)E[S^2(\omega) + U^2(\omega)] - E[S^2(\omega)] = 0 \quad (4.12)$$

which yields the optimal Wiener filter:

$$H_W(\omega) = \frac{E[S^2(\omega)]}{E[S^2(\omega)] + E[U^2(\omega)]}. \quad (4.13)$$

Practically speaking, the expected values are replaced by power spectra, which are typically estimated in a short-time varying sense to arrive at:

$$H_W(\omega) = \frac{\hat{G}_{ss}(\omega)}{\hat{G}_{ss}(\omega) + \hat{G}_{uu}(\omega)} \quad (4.14)$$

where $\hat{G}_{ss}(\omega)$ and $\hat{G}_{uu}(\omega)$ are estimates of the power spectra of the speech and noise respectively. The speech and noise statistics necessary for implementing the Wiener filter are typically generated using *maximum a posteriori* (MAP) estimation of an LPC model on a frame by frame basis [Lim and Oppenheim (1978)].

4.1.3 Other Techniques

In recent years, more sophisticated speech models have been applied to the enhancement problem. Wavelet speech enhancement techniques [Pintér (1996), Seok and Bae (1997)] provide a multi-resolution analysis which inherently takes advantage of the likeness of wavelet octave-band filtering to the perception characteristics of the human ear. Other systems, in addition to utilizing the periodic features of the speech, exploit the signal's mixture of harmonic and stochastic components [McAulay and Quatieri (1986), Hardwick (1992), Laroche et al. (1993), Yoo and Lim (1995)]. Such model-based techniques offer an improved performance in speech quality. Additionally, these methods, by virtue of their specific parameterization of the speech signal, offer some applicability to the more general acquisition problem. Hidden Markov Models (HMM), HMM-based Minimum Mean

Squared Error (MMSE), and MAP approaches are discussed in [Ephraim (1992)]. Currently, however, these model-based estimation schemes are limited to single channel applications.

4.2 Multi-Channel Approaches

By employing spatial filtering in addition to temporal processing, multi-channel systems (i.e., microphone arrays) offer a distinct performance advantage over single-channel techniques in the presence of additive noise. See [Johnson and Dudgeon (1993)] for a review of beamforming methods. The simplest microphone array method is the Delay-and-Sum Beamformer which derives its output via an averaging of the time-aligned microphone data. This technique reduces the noise level roughly by a factor of M where M is the number of microphones in the array. A variety of more sophisticated algorithms exists for adaptively “steering” the array in the direction of the desired source and simultaneously adjusting the microphone weightings to minimize the contributions of noise sources [Hoshuyama and Sugiyama (2001)]. These techniques usually assume the desired source is slowly varying and at a known location. While dynamic localization schemes like those of [Brandstein and Silverman (1997)] and weighting constraints may be incorporated into the adaptation procedure, these methods are very sensitive to steering errors which limit their noise source attenuation performance and frequently distort or cancel the desired signal.

Another class of algorithms attempts to enhance the speech signal by employing any of the beamforming methods described above, followed by a post-filter. The post-processor may take the form of multi-channel Wiener filtering [Simmer et al. (2001), Zelinski (1988), Marro et al. (1998), Yang et al. (1993), Simmer and Wasiljeff (1992)], a Spectral Subtraction technique [Gierl (1990), Dahl et al. (1997)], or a combination of the two approaches

[Meyer and Simmer (1997)]. These algorithms are limited in their assumption that the additive noise is uncorrelated. A multi-channel wavelet Wiener filter and coherence filter method is presented in [Mahmoudi and Drygajlo (1998)].

4.3 Background Noise Estimation and Voice Activity Detection

Traditional denoising algorithms such as Spectral Subtraction and Wiener filtering require an estimate of the noise spectrum or power spectrum, which is typically assumed to be known *a priori*. However, in real systems, this prior knowledge is often not available. Therefore, the noise spectrum is estimated during non-speech intervals. Automatically estimating these intervals is a very challenging problem. Alternatively, the signal subspace method presented in [Ephraim and Trees (1995)] could be used. This method separates the signal into a signal-plus-noise subspace and a noise subspace, allowing for the extraction of an estimate of the underlying speech signal.

In order to determine non-speech intervals for background noise estimation, a Voice Activity Detection (VAD) algorithm is often used. VAD methods often make use of one or more of the following crude parameters suggested in [Atal and Rabiner (1976)]: signal amplitude/energy, zero-crossings, one sample delay correlation coefficient, pitch, the first linear prediction coefficient (LPC), and the LPC residual energy. Most methods make use of only the first three of these measures [Savoji (1989)]. In [Lee and Un (1986)], a VAD method that makes use of the average signal magnitude and zero crossing rates of each frame is presented. Although these methods are relatively easy to implement in real time, they are sensitive to changes in noise/non-speech statistics [Haigh and Mason (1993)]. In [Savoji (1989)], a robust algorithm for speech endpoint detection using energy and zero-

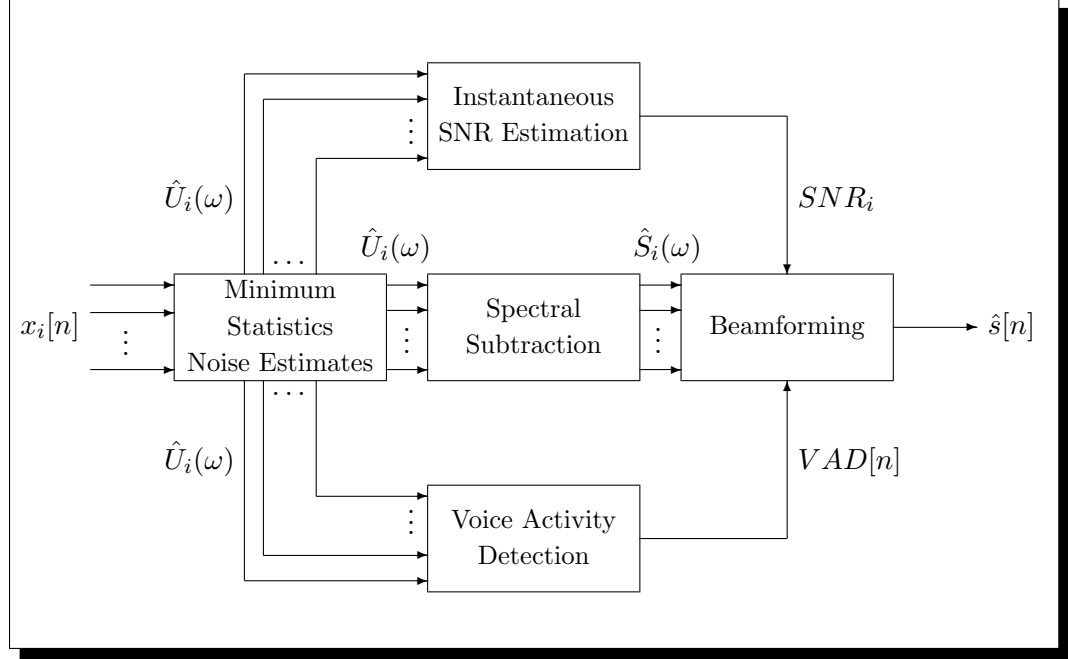
crossing measures was presented. This method assumes a three-state model (i.e., speech, non-speech, transition) rather than the two-state model (i.e., speech, non-speech) usually employed. This technique is adaptive and can handle varying signal to background noise ratios in a real-time setting. In [Floc'h (1992)], a cepstral distance measure is used to make speech/silence classifications. In [DeSouza (1983)], a statistical approach to silence detection is proposed which forms a vector of potential speech/non-speech discriminating measurements. Another approach, a maximum likelihood VAD, was presented in [Sohn and Sung (1999)]. Any useful technique for estimating background noise must be adaptable (i.e., allow for changing noise statistics, SNR, energy, etc.) and be simple enough to implement in real time. In the next section, the Minimum Statistics background noise estimation algorithm [Martin (2001)], which does not require explicit non-speech interval detection, will be discussed and used for a multi-channel Spectral Subtraction technique.

4.4 Multi-Channel Minimum Statistics Denoising (MC-MSD)

The denoising component of the proposed speech enhancement system is outlined in Figure 4.1. It will be referred to as the Multi-Channel Minimum Statistics Denoising (MC-MSD) algorithm. Each of these stages will be discussed below in more detail.

Multi-Channel Minimum Statistics

The Minimum Statistics method developed in [Martin (2001), Martin (1993), Martin (1994)] provides a simple alternative for background noise estimation. This noise estimation procedure, unlike those that estimate the background noise during non-speech intervals, does not require explicit speech/non-speech interval detection and is both intuitive and effective. The power spectrum (i.e., squared spectral magnitude) of an observed signal consists of the sum of the background noise power spectrum and speech power spectrum, assuming

Figure 4.1: *Outline of the proposed multi-channel denoising algorithm*

uncorrelated noise and speech signals. Martin observed that by tracking the minimum of each spectral bin over a sufficient number of time periods, it is possible to estimate the noise power spectrum. Throughout a speech signal, there are natural pauses during which the power spectrum falls to the noise floor. By tracking the minimum of each spectral bin over a time period long enough to encompass speech pauses, an effective background noise estimate can be determined. The algorithm is capable of estimating the possibly changing noise signal “on the fly.” As stated above, the minimum spectral magnitude of each frequency bin over a predetermined recent number of frames, N_F , is used as an estimate of the background noise at that frequency. For implementation purposes, in order to decrease the frame-to-frame difference in the noise estimate, the power spectrum of the current frame, n , is calculated with a memory term, α , as in:

$$\hat{G}_{x_i x_i}^n(\omega) = \alpha \hat{G}_{x_i x_i}^{n-1}(\omega) + (1 - \alpha) |X_i(\omega)|^2 \quad (4.15)$$

where $\hat{G}_{x_i x_i}^k(\omega)$ is the power spectrum estimate for the i^{th} channel at frame k , and $X_i(\omega)$ is the Fourier Transform of the current frame for the i^{th} channel. The memory term α satisfies $0 \leq \alpha \leq 1$. This power spectrum estimate can then be popped on to a stack of power spectrum estimates for the previous N_F frames to form a buffer. The minimum of each frequency bin of this stacked buffer leads to the noise power spectral estimate, $\hat{G}_{u_i u_i}(\omega)$, for each frequency. In this thesis, the DFT size was 512 bins and the buffer stored the previous 3 seconds worth of speech frames.

Figure 4.2 shows a speech signal with additive noise along with the time-varying energy in one DFT frequency bin (586 Hz) along with the Minimum Statistics noise estimate for 30 seconds of noisy speech. This estimate is able to adapt to changing noise, and is most effective when the buffer of frames used to find the minimum bin energy is long enough to span non-speech regions. The required buffer size can be predicted fairly well by examining the statistics of conversational speech. One problem with this technique is that the noise estimates have a downward bias due to the minimum tracking since the minimum of a noise process will generally be smaller than its expected value. Therefore, the estimate must be multiplied by a scalar to achieve a less biased noise spectrum. Typical values are between 1.1 and 1.5 [Martin (1993)]. This estimation procedure is repeated for each channel independently, giving an estimate of the noise power spectrum, $\hat{G}_{u_i u_i}(\omega)$ for each channel i .

When calculating the ML TDE in Equation 3.25, the same variance was assumed for all channels, a reasonable assumption for closely spaced microphones. Theoretically, the spectrum for additive white noise has an expected constant value equal to the noise variance. The noise can be estimated by averaging the Minimum Statistics noise spectrum estimate over all DFT bins. Averaging over all channels can make the estimate even more robust. In Figure 4.3, differing levels of white noise were added to the same 30 seconds of speech

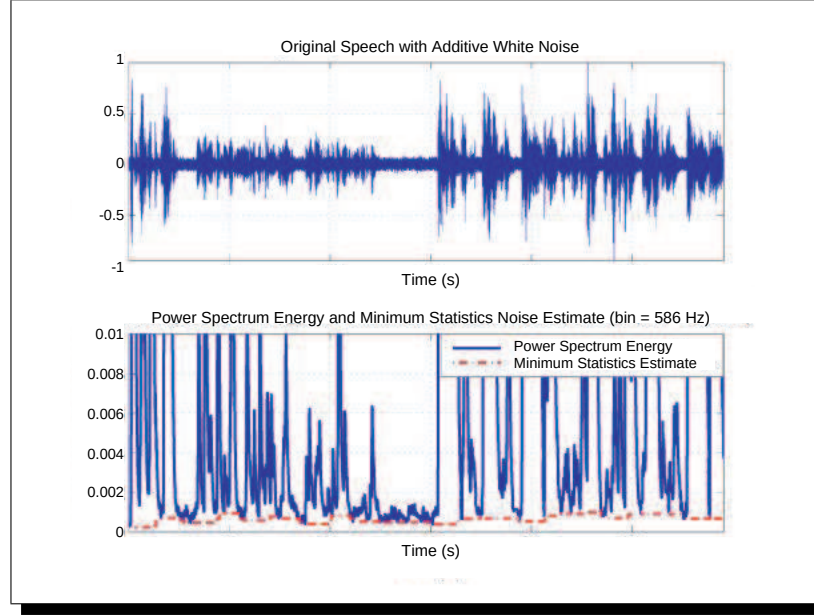


Figure 4.2: *Speech signal with additive noise and the time-varying DFT energy of the 586 Hz frequency bin and resulting Minimum Statistics noise estimate.*

used for Figure 4.2. The top plot shows the true noise variance and the estimated noise variance using the Minimum Statistics estimation technique just described. As expected, there is a downward bias in this plot. The second plot shows the true noise variances and the estimated noise variances when scaled by a factor of 1.3. This same factor will be used for the simulations presented at the end of this chapter.

Instantaneous Signal-to-Noise Ratio Estimation

The frequency-dependent instantaneous SNR can be estimated for each channel, i , and for each frame from [Martin (1993)]:

$$SNR_i(\omega) = \frac{|X_i(\omega)|^2 - \hat{G}_{u_i u_i}(\omega)}{\hat{G}_{u_i u_i}(\omega)} \quad (4.16)$$

where $\hat{G}_{u_i u_i}$ is the Minimum Statistics noise estimate for channel i . This SNR estimate will be used for both microphone subset selection and for Spectral Subtraction.

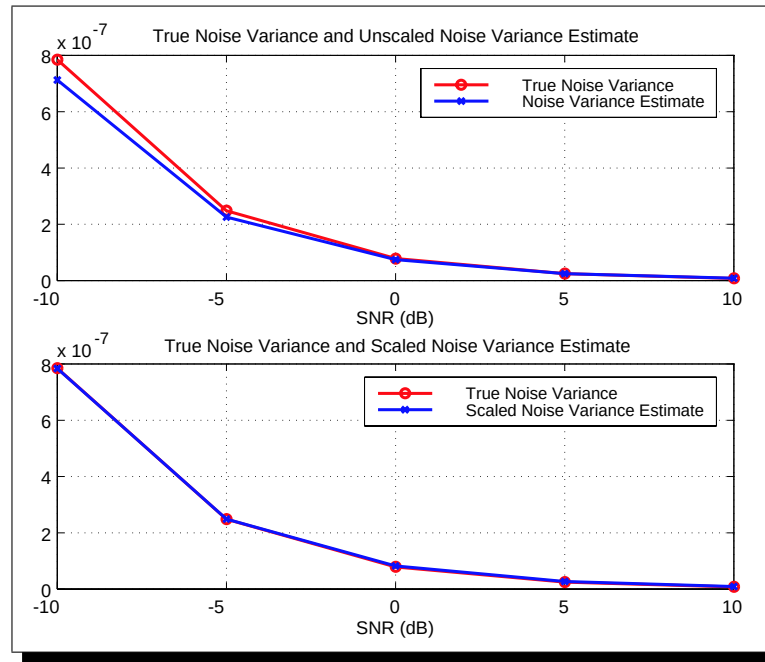
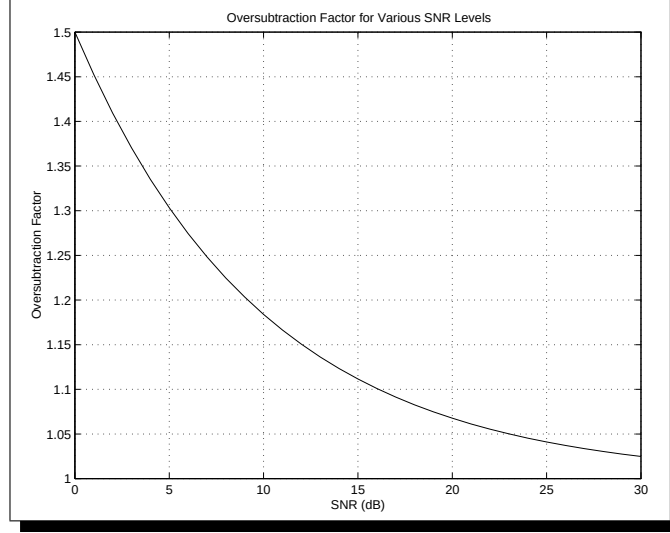


Figure 4.3: *True and estimated noise variances before and after scaling to correct for downward bias of Minimum Statistics noise estimate*

Figure 4.4: *Oversubtraction factor used for Spectral Subtraction*

Multi-Channel Spectral Subtraction

The first stage of this algorithm involved the estimation of the background noise for each channel independently using the Minimum Statistics noise estimation technique. The noise estimates are then used for Spectral Subtraction on each channel. Our experiments suggested that even after the scale factor to account for the downward bias of the Minimum Statistic noise estimate, an additional SNR-dependent oversubtraction factor, $O(\omega)$, offered an improvement to the resulting enhanced speech quality. Therefore, using the SNR estimated from Equation 4.16, the frequency-dependent oversubtraction factor for channel i is calculated from:

$$O_i(\omega) = \frac{e^{-SNR_i(\omega)/\gamma}}{2} + 1 \quad (4.17)$$

which is roughly in the range 1 to 1.5 for SNR levels between 0 and 30 as shown in Figure 4.4 for $\gamma = 10$.

The Spectral Subtraction filter analogous to Equation 4.5 for this algorithm for channel

i is given by:

$$H_i(\omega) = 1 - O_i(\omega) \frac{\sqrt{\hat{G}_{u_i u_i}(\omega)}}{|X(\omega)|}. \quad (4.18)$$

To prevent an excessive amount of tonal artifacts in the spectral subtraction, a modified filter $\tilde{H}_i(\omega)$ is calculated from:

$$\tilde{H}_i(\omega) = \begin{cases} H_i(\omega) & \text{if } H_i(\omega) > T_{SS} \\ T_{SS} & \text{if } H_i(\omega) < T_{SS} \end{cases} \quad (4.19)$$

where T_{SS} is a threshold below which the Spectral Subtraction filter is not allowed to fall and which must satisfy $0 \leq T_{SS} \leq 1$. In this thesis, a value of $T_{SS} = 0.2$ is used. This Spectral Subtraction procedure is repeated for each channel with that channel's noise power spectrum estimate.

Voice Activity Detection

The VAD decision is not required to produce the Minimum Statistics noise estimate unlike other noise estimation procedures, but it can be used to further dampen the signal during non-speech activities should the application warrant it. The instantaneous SNR of Equation 4.16 provides an intuitive metric for Voice Activity estimation [Wang et al. (2001)]. The frame is assumed to contain speech when the distance between the current short-time spectrum and the noise spectrum estimate is large (i.e., large SNR) and non-speech for small spectral distances (i.e., low SNR). Here, the SNR measure is applied in the frequency domain using:

$$SNR_{freq} = \frac{1}{M} \sum_{i=1}^M \int_{-\pi}^{\pi} SNR_i(\omega) d\omega \quad (4.20)$$

where $SNR_i(\omega)$ is the instantaneous SNR given in Equation 4.16 and M is the number of microphones in the array. If this metric is below a specified threshold, the current frame could be set to zero or some other desired non-speech condition. In many applications,

silence attenuation may be deemed undesirable, as listeners tend to prefer speech signals with a constant level of background noise to the same speech signal with non-speech intervals set to silence (with some level of background noise occurring during the speech intervals). This is especially true for speech signals with relatively short silence intervals within a speech utterance. In this case, there is a fluttering effect as the background noise appears (during speech) and disappears (during non-speech). However, in some applications such as Automatic Speech Recognition, setting non-speech intervals to zero (i.e., forcing them to be silent) prevents the recognizer from adding unwanted text to the transcribed output, a desired result.

Beamforming with Optional Microphone Subset Selection

The beamforming operation of this section is the simple Delay-and-Sum Beamformer discussed previously. However, with the instantaneous $SNR_i(\omega)$ information present for each microphone, only microphone signals with the highest SNR need to be included in the beamforming process. For an array of closely spaced microphones, the SNR values tend to be very similar and this additional step can be safely eliminated. However, in a room where the microphones are separated by great distances, it is possible for some microphones to have a much higher SNR than other microphones. This is indeed the case when there is a noise or speech source closer to some microphones than to others. In a real environment with possibly moving noise sources and probably moving speech sources, the subset of microphones with the highest SNR values will tend to change over time. Therefore, only these microphones would be used in the beamforming process. One simple statistic is to use all microphones for which the SNR value is at least 75% of the maximum SNR of all microphone channels. This statistic will tend to include all microphones in a closely spaced array while excluding low SNR microphone signals in an array of distant microphones.

4.5 Experimental Results

4.5.1 Simulations

Figure 4.5 shows the results of the proposed denoising algorithm outlined in Figure 4.1 for multiple Signal-to-Noise ratios. Room impulse responses were generated for 10 random source locations in a $4\text{ m} \times 4\text{ m} \times 3\text{ m}$ room with no reverberation effects. That is, the impulse response for each channel contains only one impulse for the direct path sound waveform which is delayed and attenuated by the appropriate amounts given the distance between the source and microphone. The random source locations and the 8 microphone locations shown in Figure 3.9 were also used in the time delay experiments from the previous chapter. At each location, a 45 second segment of speech was filtered with the appropriate generated room impulse responses. This simulated non-reverberant multi-channel speech data was then degraded by additive white noise of varying SNR levels. Due to the close proximity of these microphones, it was not necessary to choose microphone subsets prior to beamforming, as the SNR levels of the microphones were very close in value. As can be seen in the plots in Figure 4.5, the improvement of the proposed MC-MSD algorithm over a single recorded channel and traditional Delay-and-Sum Beamforming is substantial for both the segmental SNR (SEGSNR) and Bark Spectral Distortion (BSD) metrics which are computed relative to the original clean speech signal. Also shown are the improvements of Delay-and-Sum Beamforming and MC-MSD over the single channel recording.

4.5.2 Real Room Array

Figure 4.6 plots the results for a $5\text{ m} \times 5\text{ m} \times 3\text{ m}$ real room recording using 16 spatially diverse cardioid microphones with SNR-based microphone subset selection. The speech utterance was “A B C D, one two three four.” Additionally, recorded computer fan noise

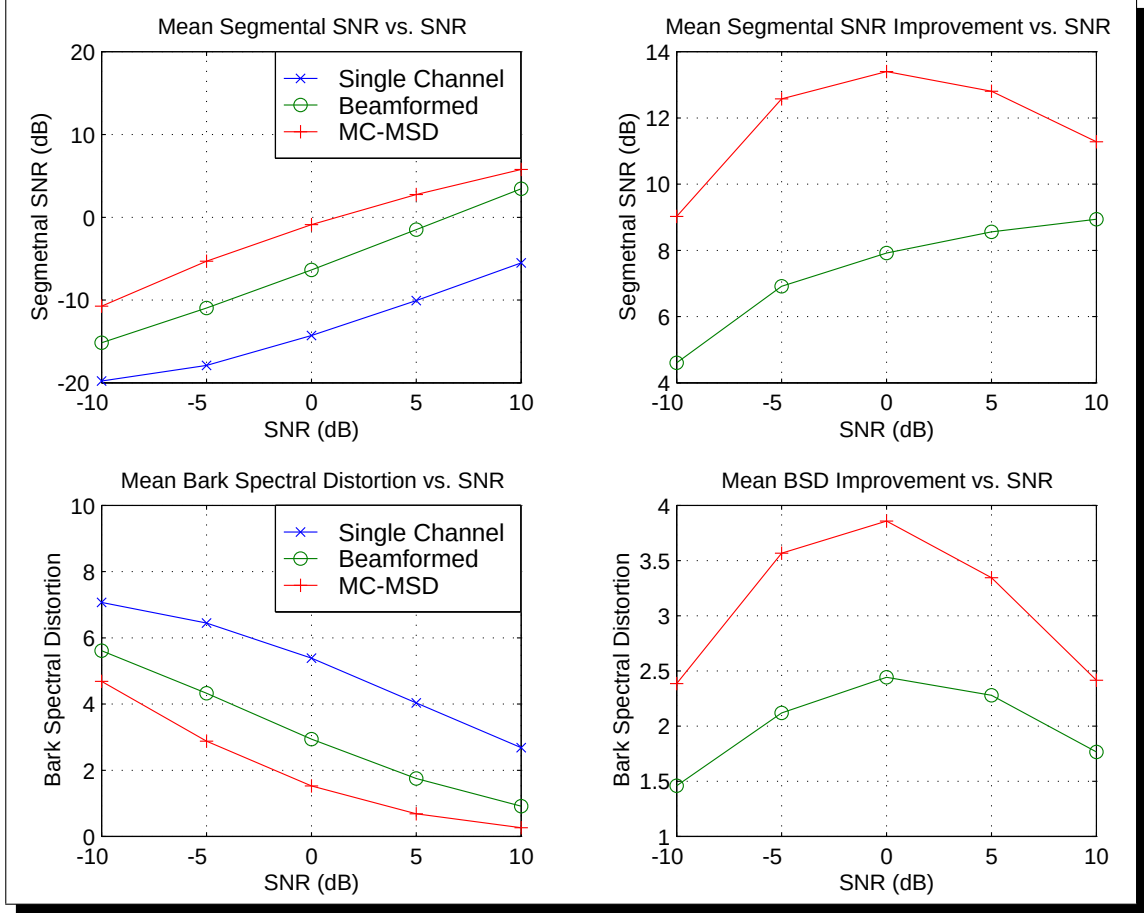
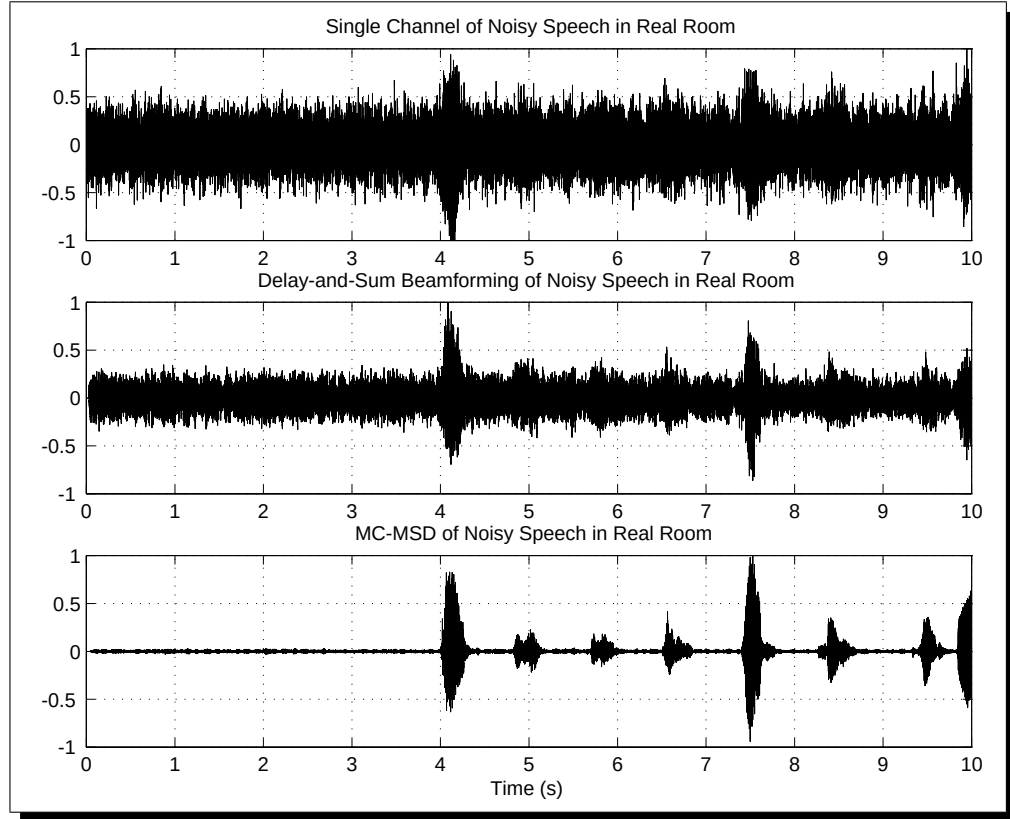


Figure 4.5: *Multi-channel denoising results for various SNR levels*

was loudly played out of a speaker during the speech recording process. The speech source, noise source, and microphone locations are shown in Figure 4.7. The acoustic conditions in this room are far from ideal. Computer fans and air conditioning contribute to an ambient noise factor over 70 dB. The floors are tiled and the walls are made of a non-absorbing material.

Figure 4.6 shows the results of the MC-MSD algorithm along with a single recorded channel and a Delay-and-Sum Beamforming of the different channels. The MC-MSD algorithm is clearly less noisy than the other techniques. The SEGSR scores for these three

Figure 4.6: *Real room denoising results*

plots are 0.13, 0.26, and 0.52 dB respectively. The BSD scores for the three plots are 12.6, 7.2, and 0.96 respectively.

4.5.3 Real Room Circular Array

The final denoising results make use of a circular microphone array, shown in Figure 4.8. This array has a total of 16 cardioid microphones (2 vertically stacked circles of 8) wrapped around a mesh cylinder. In the center of this cylinder is a foam core for acoustic dampening purposes. This array was placed on the center of a conference table in an office environment. The test consisted of a talker counting from 1 to 10 while walking around the perimeter of the table approximately 7 feet from the circular array. No attempt was made to face

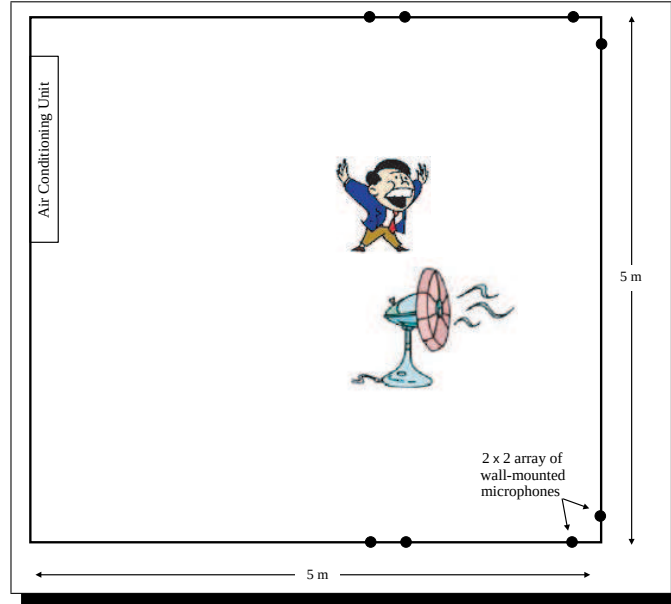


Figure 4.7: *Location of talker, fan noise source, and 16 microphones in real room denoising experiment*

the array when talking - the speech waveform was spoken at a natural angle tangent to the circular microphone array. The $5\text{ m} \times 5\text{ m} \times 3\text{ m}$ room setup is shown in Figure 4.9 which also shows the location of a speaker which was playing a constant level of fan noise throughout the 10 second recording. Note that only 1 “row” of 8 microphones is shown in Figure 4.9 although there are actually 2 rows for a total of 16 microphones, as shown in Figure 4.8.

The directionality and placement of the microphones provide an opportunity to illustrate when the SNR-based microphone subset selection strategy would be advantageous. Due to the array geometry and the fact that the speech source is in motion around the array, the SNR levels of different microphones will vary widely. For this array, it was decided that microphones would be grouped in 8 subsets of 6 microphones (3 adjacent columns of 2) and only the subset with the highest average SNR for the current frame would be used in

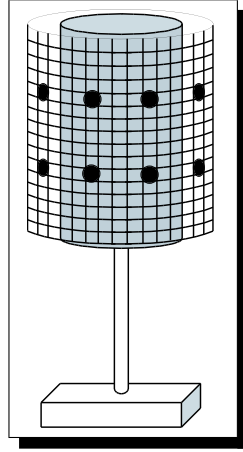
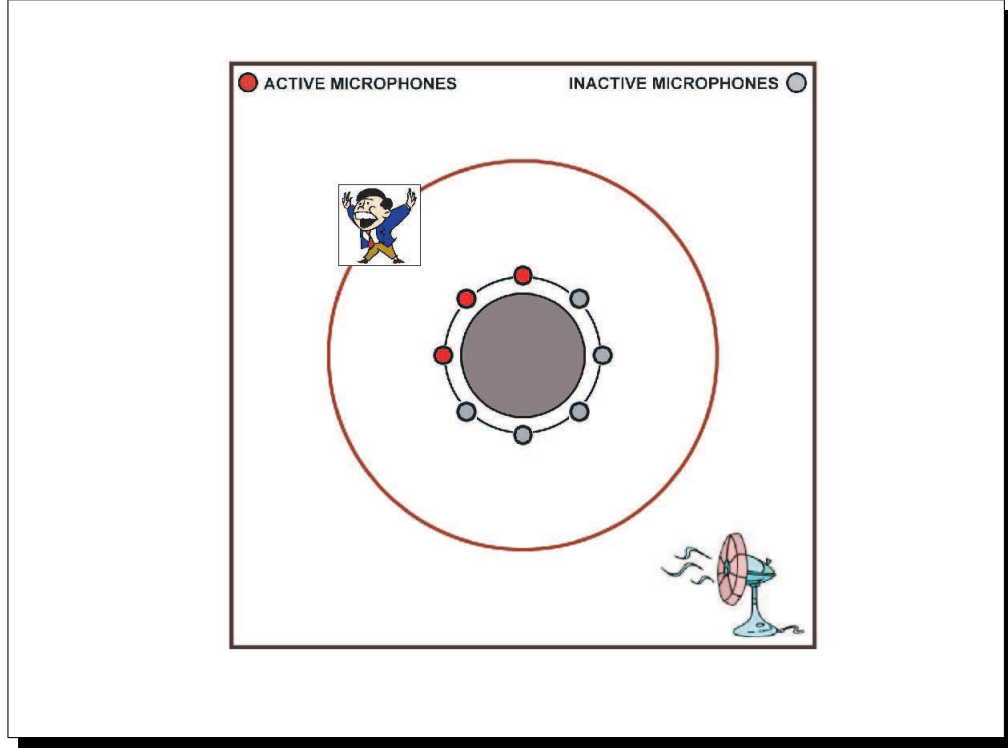


Figure 4.8: *Circular array of 16 microphones used in real room denoising experiment*

the beamforming if the following condition held: the subset used for beamforming could only change from the previous frame's subset if a significantly large change in SNR was realized or if the two subsets were not adjacent and the previous frame's subset had a very low SNR level for the new frame (indicating a new, distant speech source). This prevents a jittering between adjacent subsets for the same or close sources. This switching tends to be an annoyance to listeners. In Figure 4.9, the upper 3 microphones of the 6 microphone subset are indicated for the current position of the talker in the room.

Figure 4.10 shows how the selected subset of microphones varies as the talker circles the array - an expected result. Figure 4.11 shows the results of the MC-MSD algorithm along with one channel of recorded speech (the microphone at the top of the circle in Figure 4.9) and the beamformed result. The beamformed output is actually noisier than the one channel recording because it averages all microphone signals, including those closest to the fan. Naturally, if the one channel example was for a microphone near the fan noise, it would have been noisier than the beamformed result. In any case, it is interesting to note that for the one channel recording, the amplitude of the speech decreases as the talker

Figure 4.9: *Circular array experiment room setup*

gets farther away when travelling along the opposite side of the array. The speech then increases in amplitude as the circle is completed and the talker nears the first microphone again. In the beamformed output, the noise is higher as explained above, but the talker level is constant, since all microphones are equally weighted. Finally, the results of the MC-MSD are quite effective in eliminating the background noise while maintaining a constant talker level. This is even true when the talker passes in front of the noise source.

4.6 Denoising Summary

This chapter presented a multi-channel denoising algorithm based on the Minimum Statistics noise estimation and Spectral Subtraction procedures. The effectiveness of this method

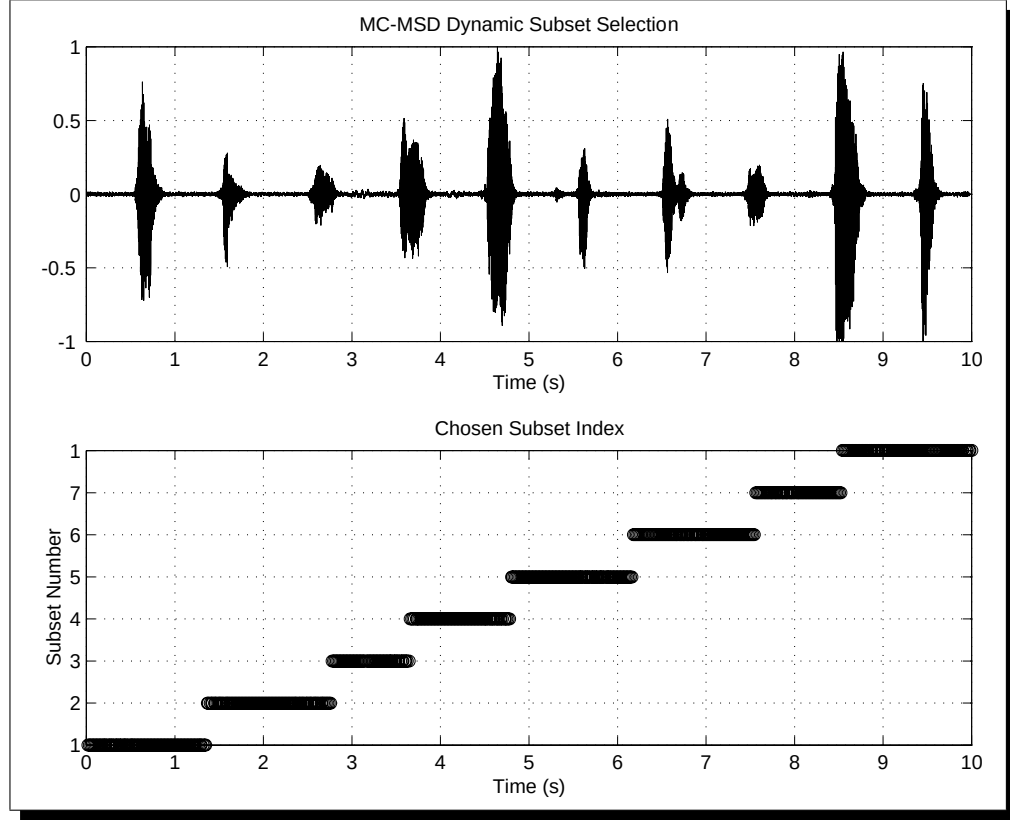
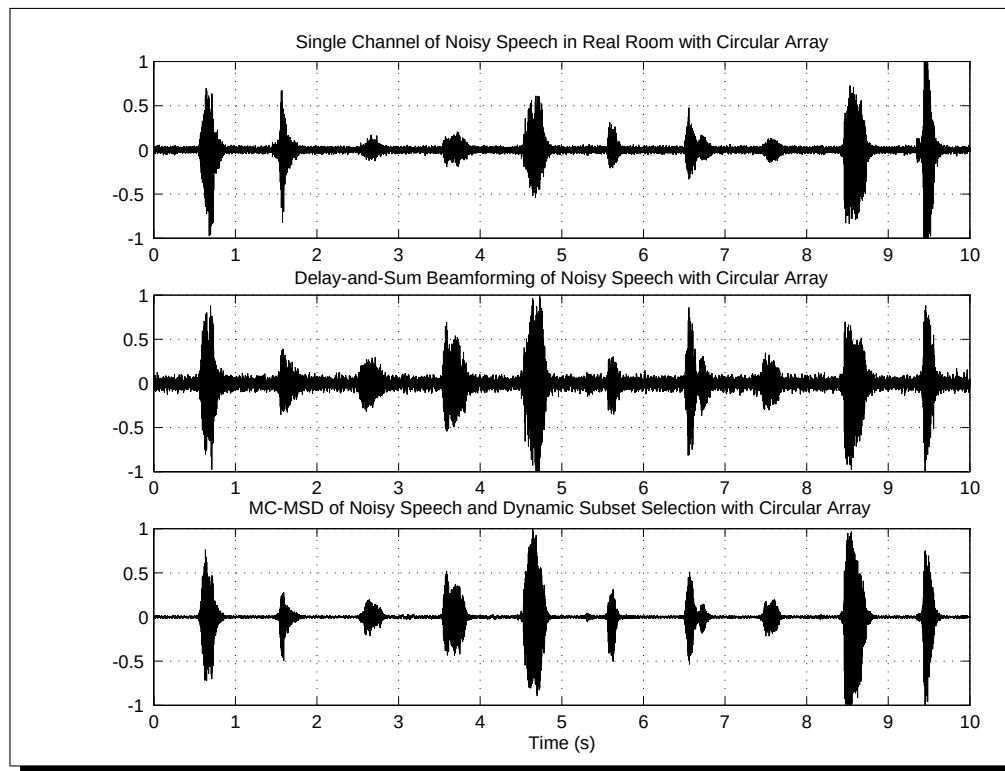


Figure 4.10: *Optimal SNR microphone subset selection where subset 1 is centered around the topmost microphone in Figure 4.9*

in both simulations and real room recordings was demonstrated. The next stage of the proposed enhancement system is multi-channel model-based dereverberation which is meant to remove the convolutional noise due to reverberations, a separate problem from the removal of background noise discussed in this chapter.

Figure 4.11: *Circular array denoising results*

Chapter 5

Multi-Microphone Model-Based Dereverberation

Recorded speech signals are subjected to degradations arising from the environment in which they are produced. In a typical enclosure, speech waveforms will reflect off of the walls, floor, ceiling, and any other items (e.g., furniture) in the room. The way in which the waveform reflects is a function of the properties of the surface in question. Some surfaces such as carpeting are highly absorptive and produce little reverberation, while, for example, a solid metal wall would lead to very high levels of reverberation. In any typical room, there will be at least minimal levels of reverberation for which the intelligibility of the speech waveform will be decreased. These degradations provide the motivation for research in the area of reverberant speech enhancement. Additionally, as seen in Chapter 3, the performance of traditional speech processing algorithms such as time delay estimation and source localization can be drastically reduced in the presence of room reverberations. Typically, single microphone techniques exploit various features of the speech signal while multi-channel methods focus primarily on improving the quality of the spatial filtering

process. Two-channel methods which address both background noise reduction and dereverberation are rare [Dorbecker and Ernst (1996), Martin and Vary (1994)].

As will be shown, it is possible to utilize some highly nonlinear, but still intuitive, techniques to suppress the effects of reverberation. In this chapter, a joint approach will be taken which makes use of explicit models of the original speech utterance and room reverberations in order to produce an enhanced speech signal from the multi-channel degraded data. Before presenting this new approach, a brief survey of the most commonly used single and multiple microphone techniques for dereverberation will be offered.

5.1 Single-Channel Approaches

Past research on single-channel enhancement of reverberant speech has focused on two main approaches: inverse filtering of the channel responses and cepstral averaging. Many approaches have attempted to invert the channel function between the speech source and the single microphone [Neely and Allen (1979), Tohyama et al. (1993)]. Alternatively, this may be performed in the cepstral domain where the inverse filtering operation becomes a subtraction operation. In practice, while the channel function is generally not minimum phase, only the minimum phase portion can be inverted. Although the inverse filtering approach is appealing in theory, the room impulse responses are difficult to estimate in practice. The channel responses are affected by changes in speaker and/or microphone locations, as well as changes in the environment itself [Subramaniam et al. (1996), Liu et al. (1995)]. The inversion technique can actually further degrade the reverberant speech if the room response is not accurately estimated [Mourjopoulos (1985)]. In general, this method is only effective in highly controlled environments for which the channel response is known or measured with great accuracy - an improbable event.

The Real Cepstrum, $c_x[n]$, of a signal, $x[n]$, is found from the inverse Fourier Transform

of the log magnitude spectrum, i.e., [Deller et al. (1987)]:

$$c_x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |X(\omega)| e^{j\omega n} d\omega \quad (5.1)$$

where $X(\omega)$ is the Fourier Transform of $x[n]$. It is well known that the convolution operation in the time domain results in a multiplication operation in the frequency domain. Therefore, the log operator in the above equation leads to a mapping of convolution in the time domain to an addition operation in the cepstral domain. Because the recorded speech signal is formed from a convolution of the original speech waveform and the room impulse response, the resulting recorded signal's cepstrum will be the sum of the cepstra of the speech signal and room response. The room impulse response is made up of a sum of delayed impulses, causing its log spectrum to have rapid variations [Bees et al. (1991)], unlike the speech signal which has a relatively slowly varying spectrum. Therefore, the speech spectrum is mainly located in the lower quefrecies of the cepstrum [Liu et al. (1996), Oppenheim and Shafer (1989), Deller et al. (1987), Rabiner and Shafer (1978)], while the reverberation effects tend to appear far away from the origin in the cepstrum. By windowing (i.e., “liftering”) the cepstrum, low-quefrecy coefficients corresponding to the speech spectrum are maintained, while the reverberant effects are attenuated. However, this method is complicated by the fact that the speech and echo portion of the cepstrum are often not well separated, and the optimal length of window in the liftering process is difficult to estimate [Cole et al. (1997), Liu et al. (1996), Liu et al. (1995)]. Therefore, this technique is generally useful only for the simplest of echoes.

5.2 Multi-Channel Approaches

Multiple microphone processing which takes advantage of spatial filtering opportunities is known as beamforming. Perhaps the simplest of all beamformers is the non-adaptive, data-

independent beamformer that is a linear combination of the microphone signals, which, in the frequency domain, is expressed as:

$$B(\omega) = \sum_{i=1}^M W_i^*(\omega) X_i(\omega) \quad (5.2)$$

where $B(\omega)$ is the beamformed spectrum, $W_i(\omega)$ is the frequency-dependent weight associated with the i^{th} sensor, $*$ is the complex conjugation operator, M is the number of microphones, and $X_i(\omega)$ is the spectrum of the i^{th} sensor signal. The Delay-and-Sum Beamformer is defined by the delay compensated, uniform weighting:

$$W_i(\omega) = \frac{1}{M} e^{-j\omega\tau_i} \quad (5.3)$$

where τ_i is the source's time delay to the i^{th} microphone relative to some reference point.

Typically, microphone array research deals with farfield sources, for which the arrival angle, θ , measured relative to broadside (i.e., perpendicular to the array), is the same for each pair of microphones in a linear array (see Chapter 2). With a uniformly spaced linear array and using microphone 1 as the phase reference point (i.e., $\tau_1 = 0$), the resulting weighting vector for steering the beam to angle θ is:

$$W(\omega) = \frac{1}{M} e^{-j\omega(i-1)d/c\sin(\theta)}, \quad i = 1, \dots, M \quad (5.4)$$

where d is the separation of the microphones and c is the speed of sound in the environment (typically 342 m/s).

Figure 5.1 illustrates the beampatterns associated with an 8-element, linear array with $d = 10$ cm for all potential source locations in a horizontal slice of a $4 \text{ m} \times 4 \text{ m}$ room. The beam is steered to a source located at various angles (0° , 15° , 30° , 45° , 60° , and 75°) at a range of 1.5 cm. The specified frequency is 2000 Hz. Because soundwaves drop off at a rate proportional to the distance from the source, the beampattern decreases in amplitude along any direction of arrival with increasing distance from the array. These plots graphically

illustrate the level to which beamsteering filters out reflections from angles outside the main lobe of the resulting beam pattern.

Although this simple beamforming procedure does reduce reverberations and background noise to some degree, many attempts have been made to improve upon this performance. By beamforming to the direct path and the major images, it is possible to use the multipath reflections constructively to increase the SNR well beyond those achieved with a single beamformer. The result is a matched filtering process [Flanagan et al. (1993)] which is effective in enhancing the quality of reverberant speech and attenuating noise sources. Unfortunately, this technique has a number of practical shortcomings. The matched filter is derived from the source location-dependent room response and is therefore difficult to estimate dynamically. The channel responses obtained in this manner do not address the issue of non-stationary or unknown source locations or changing acoustic environments. These problems are addressed in [Affes and Grenier (1997)] which attempts to adaptively estimate the channel responses and incorporate the results into an adaptive beamforming process.

In general, beamforming research has dealt with algorithms to attenuate undesired sources and noise, track moving sources, and deconvolve channel effects. These approaches, while effective to some degree, are fundamentally limited by the nature of the distant talker environment. Array design methods are overly sensitive to variations in their assumptions regarding source locations and radiation patterns, and are inflexible to the complex and time-varying nature of the enclosure's acoustic field. Motion as little as a few centimeters or a talker's turning his or her head is frequently sufficient to compromise the optimal behavior of these schemes in practical scenarios [Radlovic et al. (1999), Brandstein (1998)]. Similarly, matched filter processing, while shown to be capable of tracking source motion to a limited degree, requires significant temporal averaging and

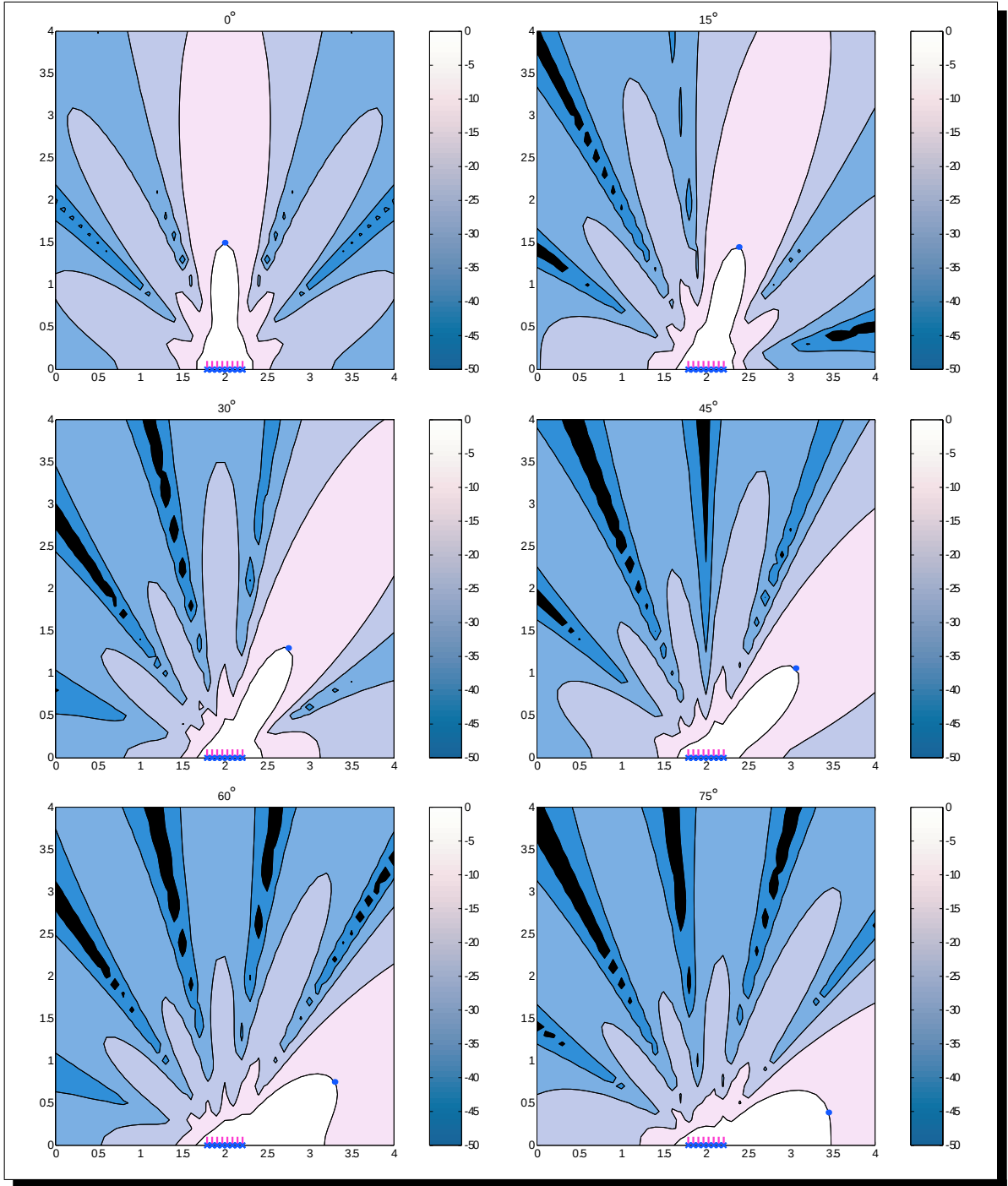


Figure 5.1: Array response for a horizontal slice of a $4\text{ m} \times 4\text{ m}$ room for $\theta = 0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ,$ and 75° from vertical at a frequency of 2000 Hz where the 8 microphones are represented by $\times$ and the point of focus is indicated by $\circ$

is not adaptable at rates sufficient to effectively capture the motions of a realistic talker. However, some success can be found when producing only a coarse model of the reverberant channels [Griebel and Brandstein (2001)]. In that work, the incorporated channel model required only approximate times and amplitudes of the initial multi-path reflections. In practice, these impulses are responsible for the bulk of the reverberant energy in the received speech signal and can be estimated to a sufficient degree on a time-varying basis.

In [Brandstein (1998)] an alternative to these methods was proposed that explicitly incorporated the Dual Excitation Speech Model [Hardwick (1992)] into the beamforming process. The work in [Brandstein (1999a)] extended this idea using a multi-channel version of the Multi-Pulse LPC model [Singhal and Atal (1984), Atal and Remde (1982)] and a nonlinear event-based processing method to discriminate impulses in the received signals due to channel effects from those present in the desired speech. These approaches were shown to suppress the deleterious effects of both reverberations and additive noise without explicitly identifying the channel, and were adaptive on a frame by frame basis. However, this research was restricted to a small class of speech signals and was limited to individual analysis frames. In this thesis, these concepts are taken a step further by employing the wavelet domain for a multi-resolution signal analysis and reconstruction of the LPC residual. This allows the nonlinear, model-based processing paradigm to be applied for effective speech enhancement under general conditions.

5.3 Multi-Channel Model-Based Dereverberation Algorithm (MC-MBD)

As introduced in Chapter 2, the reverberant speech signal, $x_i[n]$, observed at the i^{th} microphone ($i = 1, 2, \dots, M$) can be modeled as:

$$x_i[n] = h_i[n] * s[n] \quad (5.5)$$

where $s[n]$ is the clean speech utterance and $h_i[n]$ is the room impulse response between the speech source and the i^{th} microphone. Under all but ideal circumstances, the room impulse response is a complicated function of the enclosure environment.

A very general model for speech production approximates the vocal tract as a time-varying all-pole filter [Deller et al. (1987)], as described in Chapter 2. In the case of voiced speech, the filter excitation is modeled as a quasi-periodic impulse train where the time between consecutive impulses is the pitch period. For unvoiced speech, the excitation signal is approximated by random noise. The reverberant channel response itself is a sum of weighted impulses, one for the direct path and one for each multi-path echo arriving at the microphone. The proposed algorithm relies on the assumption that the detrimental effects of reverberations introduce only zeros into the overall system and will primarily affect only the nature of the speech excitation sequence, not its all-pole filter. It is also assumed that the impulses due to reverberations contributed to the excitation sequences are relatively uncorrelated across the individual channels, while the excitation impulses due to the original speech are invariant to the environmental effects. Essentially, the approach will be to identify the clean speech excitation signal from a set of corrupted excitations. The enhanced speech is then reconstructed by using the enhanced excitation signal as input to an estimate of the all-pole filter representing the vocal system.

The idea of enhancing the excitation signal was explored in [Brandstein (1999a)] using

the LPC residual derived from multiple microphone channels. The algorithm proposed in this chapter represents a more effective method for estimating and then reconstructing the excitation signal by employing a class of wavelets to decompose the LPC residuals. In [Mallat and Zhong (1992)], quadratic spline wavelets were shown to be effective at detecting singularities in a signal. As discussed in Chapter 2, because these quadratic spline wavelets are the derivatives of smoothing functions, significant Wavelet Transform (WT) coefficients, or wavelet extrema, prove to be appropriate indicators of discontinuities in a signal. The set of extrema at each wavelet scale in the WT of the LPC residual signals is an effective representation of the impulses (i.e., discontinuities) in the excitation signals. By locating the extrema which are well clustered across all channels, we attempt to capture the underlying impulsive structure (for voiced speech) of the original non-reverberant LPC residual signal. The estimated “clean” multi-scale extrema can be used to reconstruct the desired excitation signal with minimal approximation error [Mallat and Zhong (1992)].

The overall procedure is outlined in Figure 5.2. The received signals, $x_i[n]$, are assumed to have been time-aligned either through *a priori* knowledge of the source and microphone geometry or as the result of applying a relative time-delay estimation procedure to the original channel data (see Chapter 3). This procedure is repeated for half-overlapping analysis frames which, after processing, are added after being multiplied by the appropriate length Hanning window.

For the proposed algorithm, the microphone signals are first converted to LPC residuals and transformed to the wavelet domain. For each wavelet scale, the multi-channel extrema are clustered and scaled by a local coherence measure. The resulting extrema at each scale are used to generate a multiresolution representation from which the enhanced excitation signal is produced. This synthesized excitation (presumably with less of the detrimental reverberation and noise effects) is then fed into the estimated LPC filter. The algorithm

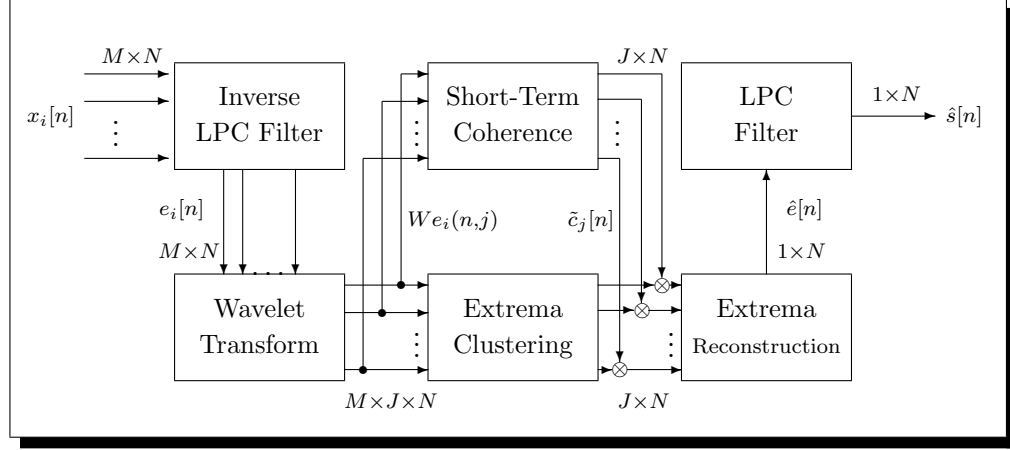


Figure 5.2: Outline of the proposed dereverberation algorithm

stages will now be discussed in more detail. For the illustrative plots that follow in the remainder of this section, speech with a reverberation time of 400 ms is employed.

5.3.1 Joint Inverse LPC Filter

A joint LPC filter is derived from N -point windows of the M channels of speech. The linear prediction formalism assumes the speech source signal can be modeled as [Markhoul (1993)]:

$$x_i[n] = \sum_{k=1}^K a_k x_i[n-k] + e_i[n] \quad (5.6)$$

where $e_i[n]$ is the residual or excitation signal discussed previously for the i^{th} microphone and K is the LPC model order. The LPC coefficients, $\{a_1, a_2, \dots, a_K\}$, are estimated by minimizing the average energy of the residual signal. For this multi-channel setting the criterion is expressed as:

$$\frac{1}{M} \sum_{i=1}^M \sum_{n=1}^N e_i^2[n] = \frac{1}{M} \sum_{i=1}^M \sum_{n=1}^N \left\{ x_i[n] - \sum_{k=1}^K a_k x_i[n-k] \right\}^2. \quad (5.7)$$

The multi-channel autocorrelation function will be defined as:

$$R[l] = \frac{1}{M} \sum_{i=1}^M \sum_{n=l}^N x_i[n] x_i[n-l]. \quad (5.8)$$

The minimum energy solution is achieved by solving the same system of linear equations as in the single-channel case, namely:

$$\begin{bmatrix} R(0) & R(1) & \dots & R(K-1) \\ R(1) & R(0) & \ddots & R(K-2) \\ \vdots & \ddots & \ddots & \vdots \\ R(K-1) & R(K-2) & \dots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_K \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(K) \end{bmatrix} \quad (5.9)$$

where $R(l)$ is as defined in Equation 5.8.

The resulting LPC filter coefficients estimates $\{\hat{a}_1, \hat{a}_2, \dots, \hat{a}_K\}$ can be used to generate a set of residual signals, $e_i[n]$, via inverse filtering of each microphone signal, $x_i[n]$. The goal of the later wavelet extrema reconstruction will be to use the multiple residual signals to create a single enhanced residual signal, $\hat{e}[n]$. This residual will then be used as input to the forward LPC filter using the estimated coefficients $\{\hat{a}_1, \hat{a}_2, \dots, \hat{a}_K\}$ to generate an enhanced speech signal, $\hat{s}[n]$.

For the work presented in this chapter, the analysis frame is 32 ms ($N = 256$ samples at 8 kHz sampling rate), the number of channels is $M = 8$, and the LPC model order is $K = 13$.

5.3.2 Wavelet Transform of LPC Residual Signals

The first J scales of the Wavelet Transform are calculated for each of the M residual signals. In this thesis, the WT is calculated for the first $J = 5$ scales. The WT of the residual signal of the i^{th} channel, $e_i[n]$, is denoted by $We_i(n, j)$ where j is the scale parameter and n is the temporal index within that scale. The non-decimated quadratic spline wavelet

discussed in Chapter 2 is used. Because the wavelet is the derivative of a smoothing function, the resulting WT coefficients will contain local extrema at discontinuities in the signal. Therefore, impulses in the residual signal of voiced speech segments cause large WT coefficients at different scales.

In the following stages, this WT is used to reconstruct an enhanced residual signal, where wavelet extrema at different scales correspond to impulses in the voiced excitation or to impulses from the room channel response due to reverberations. While the significance of the wavelet extrema in the unvoiced case is not as clearly delineated, this approach is also effective for reconstructing an appropriate noise-like LPC residual signal.

5.3.3 Wavelet Transform Coefficient Extrema Clustering

Based upon the observation that residual impulses due to the original speech tend to be clustered across channels and scales while those due to reverberation effects are relatively uncorrelated in both amplitude and time, a clustering technique is now applied to estimate the “clean” extrema. In [Brandstein (1999a)] a nonlinear clustering criterion was presented for this application, while in [Griebel and Brandstein (1999)] we employed a Gaussian smoothing approach. However, the local extrema of WT coefficients of the Delay-and-Sum Beamformed speech have been found to perform sufficiently well in this context. Because of its computational advantages and near comparable performance, this delay-and-sum strategy will be employed for extrema clustering in what follows. The plot in Figure 5.3 illustrates WT coefficients of scale 2^5 for 8 channels and the resulting clustered extrema. In general, for voiced speech, extrema due to glottal excitations - the rising slopes near 2.5, 12.5, 22.5, and 32.5 ms - tend to be more tightly clustered than those due to additional impulses from the reverberant channel responses. This causes the less clustered peaks to have a relatively small amplitude. Figure 5.4 shows the clustered extrema of scale

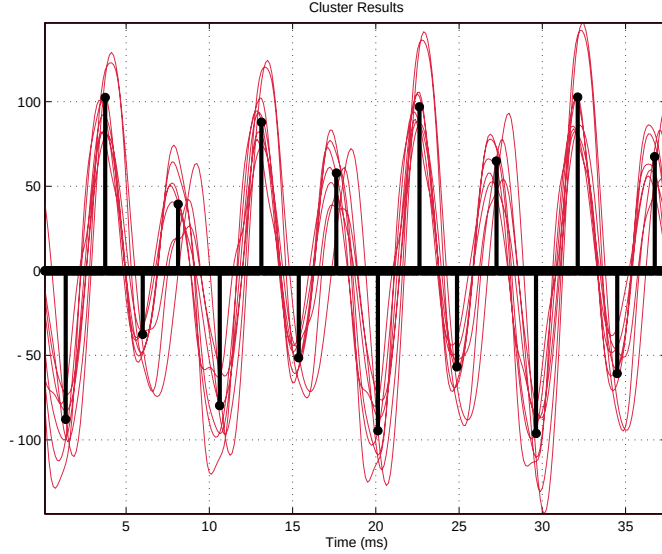


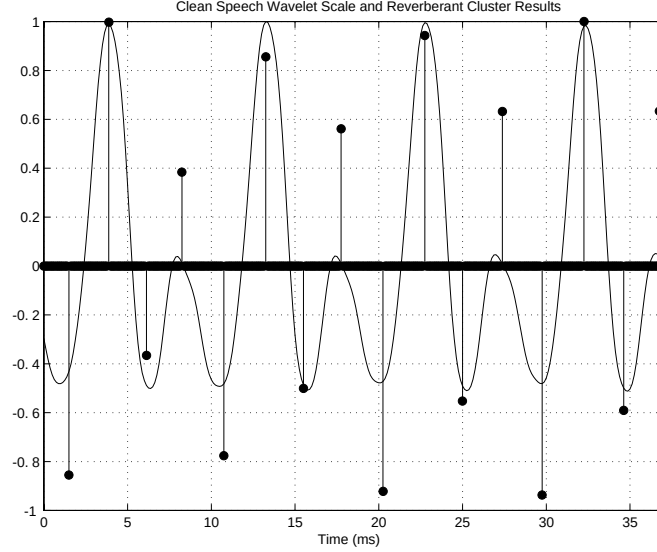
Figure 5.3: *Plot of 8 reverberant channels at scale 2^5 and clustered peaks*

2^5 along with scale 2^5 of the clean speech. Note that some of the extrema have amplitudes that are too large compared to the clean speech. These are the peaks occurring late in the pitch period where the low energy of the clean speech is corrupted by the reverberant channel energy. The clustering operation maps the multi-channel WT extrema data to a single WT extrema representation suitable for input to the WT extrema reconstruction algorithm described previously.

5.3.4 Short-Term Coherence Scaling

The purpose of the coherence-based amplitude modulation scheme detailed in this section is to further attenuate the clustered peaks due to room reverberations beyond what is achieved by the averaging process associated with the extrema clustering.

The following short-term coherence window is multiplied by the wavelet extrema at

Figure 5.4: *Clustered reverberant peaks and scale 2^5 of clean speech*

each scale j :

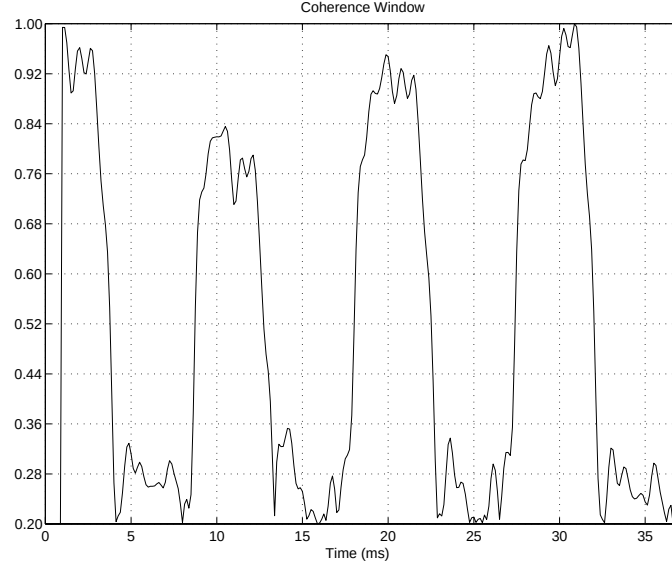
$$c_j[n] = \frac{2}{M-1} \frac{\sum_{i=1}^M \sum_{k=i+1}^M \sum_{l=-L}^L W e_i(n+l, j) W e_k(n+l, j)}{\sum_{i=1}^M \sum_{l=-L}^L W e_i(n+l, j)^2} \quad (5.10)$$

where $2L+1 \ll N$ is the length of a small subwindow within the analysis frame which is scale dependent. For scales 1, 2, 3, 4, and 5, the value of L was set to 1, 1, 2, 4, and 8 respectively.

With this window, any clustered extrema in areas of low coherence will be de-emphasized, while highly coherent extrema will be relatively unaffected. The coherence function as given above can assume values between -1 and 1 . A thresholding scheme is incorporated to maintain non-negative scaling terms and to normalize the coherence values within the subwindow. For each scale 2^j , a modified coherence window is calculated from:

$$\tilde{c}_j[n] = \begin{cases} c_j[n] / \max_n c_j[n] & c_j[n] / \max_n c_j[n] \geq T \\ T & c_j[n] / \max_n c_j[n] < T \end{cases} \quad (5.11)$$

where $0 \leq T \leq 1$. A reasonable value of T was determined experimentally to be 0.2. With


 Figure 5.5: *Short-term coherence function for scale 2^3*

$T > 0$, the thresholding ensures that the coherence window does not completely eliminate any wavelet extrema, which could result in excessive smoothing of the resulting speech signal. The normalized coherence function for scale 2^3 is shown in Figure 5.5. For coarser scales, the coherence is closer to 1 throughout the frame. The results of weighting the clustered peaks by the coherence function are shown in Figure 5.6. The points of excitation (i.e., the large positive peaks) are still approximately equal in amplitude to the peaks of the clean speech while the extra reverberant energy of the other peaks in Figure 5.4 has been attenuated due to low coherence values.

Finally, the coherence window at each scale is multiplied by the average coherence for that frame and that scale. That is, the wavelet coefficients at scale 2^j , are multiplied by the constant:

$$C_j = \frac{1}{N} \sum_{i=1}^N c_j[n] \quad (5.12)$$

which is obviously calculated before the coherence window was normalized to unity maxi-

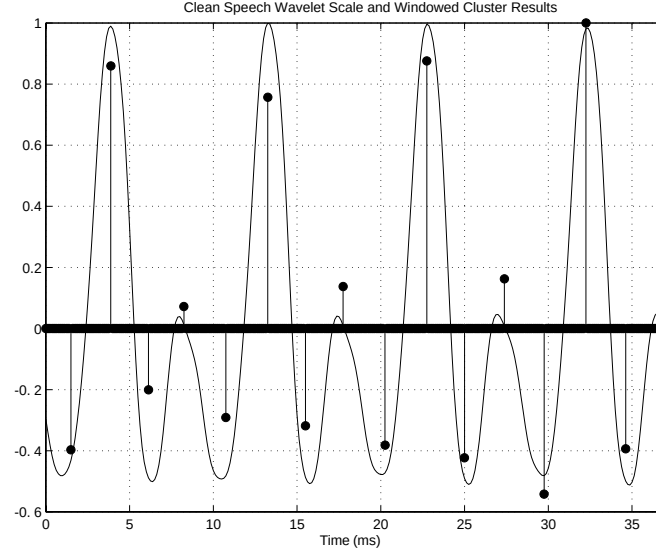


Figure 5.6: *Weighted peaks and scale 2^5 of clean speech*

imum amplitude. The normalization process of Equation 5.11 preserves the most coherent peaks of the current frame. However, if these peaks are less coherent than those of the neighboring frames, they should be attenuated accordingly. This final scale multiplication factor of Equation 5.12 achieves this. This is useful, for example, with speech tails where reverberant energy continues after the speech utterance has stopped. Since this energy tends to be much less coherent from channel to channel, this final multiplication factor will attenuate these reverberant tails.

5.3.5 Wavelet Transform Coefficient Extrema Reconstruction

Another advantage of working in the wavelet domain is the ability to approximately reconstruct a signal from the local extrema of its WT coefficients at each scale. An alternating projection algorithm for WT extrema reconstruction [Mallat and Zhong (1992)] was presented in Chapter 2. Recall that this method tends to approximate a signal to within a very high frequency error term, which, for the speech signals examined in this research, was

inaudible.

The clustered peaks for each wavelet scale are multiplied by the appropriate scale-dependent coherence envelope as determined in the previous step. The low resolution approximation coefficients are simply beamformed over all channels. This new low resolution approximation and the multiresolution coherence-weighted WT extrema are used to reconstruct the enhanced LPC residual using the reconstruction algorithm discussed in Chapter 2. Figure 5.7 shows an example of the LPC residual for the clean and single channel reverberant speech frames followed by the residual signal resulting from the WT extrema reconstruction algorithm. The reconstructed residual signal has distinct peaks with amplitudes and indices approximately equal to those of the clean residual. The proposed method also considerably attenuates the reverberant energy which appears between the peaks in the middle plot.

5.3.6 Forward LPC Filter

Finally, for each analysis frame, the reconstructed residual signal, $\hat{e}[n]$, is filtered by the previously estimated LPC coefficients to generate the enhanced speech signal, $\hat{s}[n]$. Figures 5.8, 5.9, and 5.10 show the results for three different frames of speech for a reverberation time of 400 ms and additive noise at an SNR level of 10 dB. Figure 5.8 displays a single analysis frame corresponding to a little over one period of a voiced speech segment. The four plots compare the reconstructed speech signal to the original speech utterance, a single reverberant channel, and a simple Delay-and-Sum Beamforming of the M reverberant speech channels. It is apparent that the primary effect of the reverberant distortion is to inflate the signal energy between 987 ms and 997 ms. The beamformer is ineffective at attenuating these distortions. The proposed algorithm is capable of detecting the excitation energy due to the clean speech and produces a synthesized signal very similar to the original. Fig-

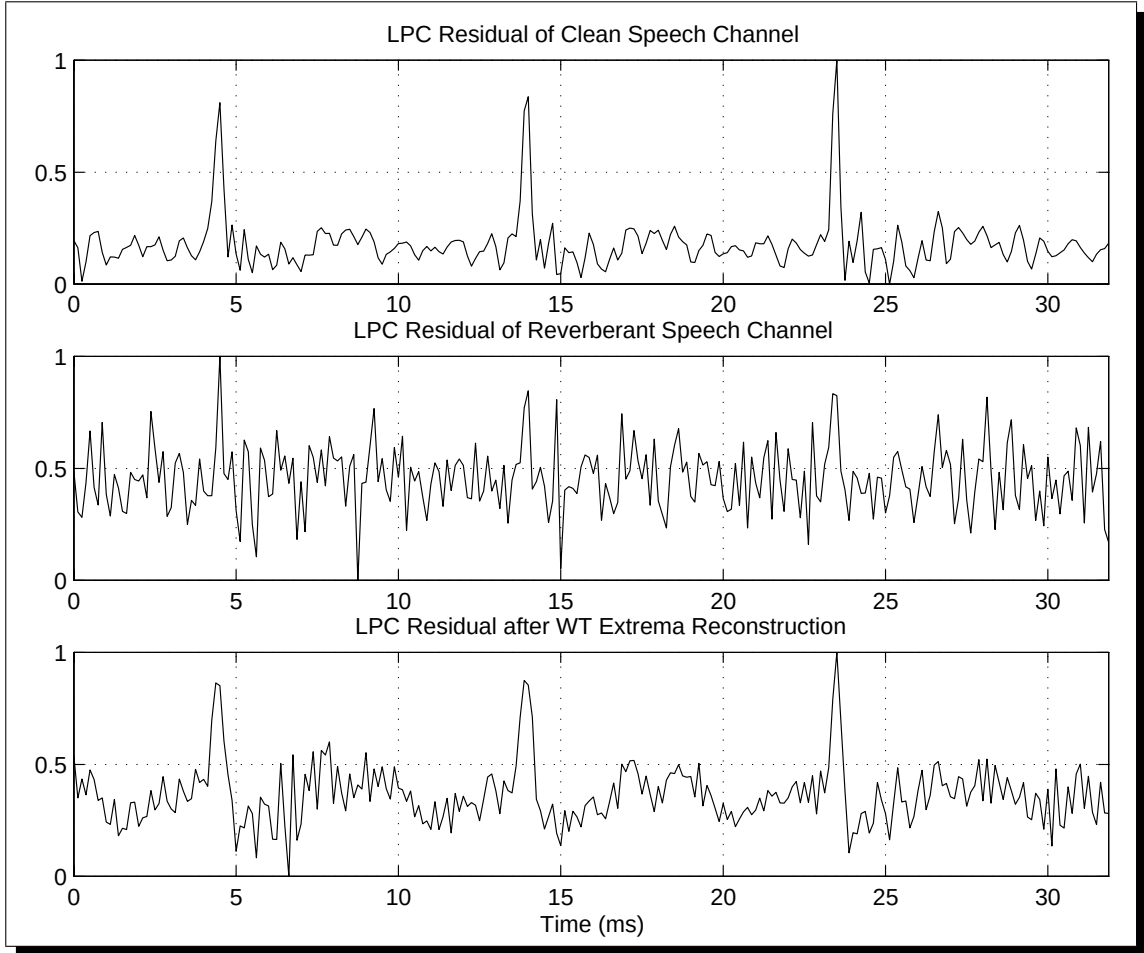


Figure 5.7: *LPC residual of clean speech, reverberant speech, and after wavelet clustering technique*

ure 5.9 demonstrates a voiced speech segment from early in a vowel segment. At this point, the speech is more affected by additive noise than reverberations, and the enhancement procedure diminishes its presence as well. Figure 5.10 shows another vowel portion heavily affected by reverberations for which the results are equally pleasing.

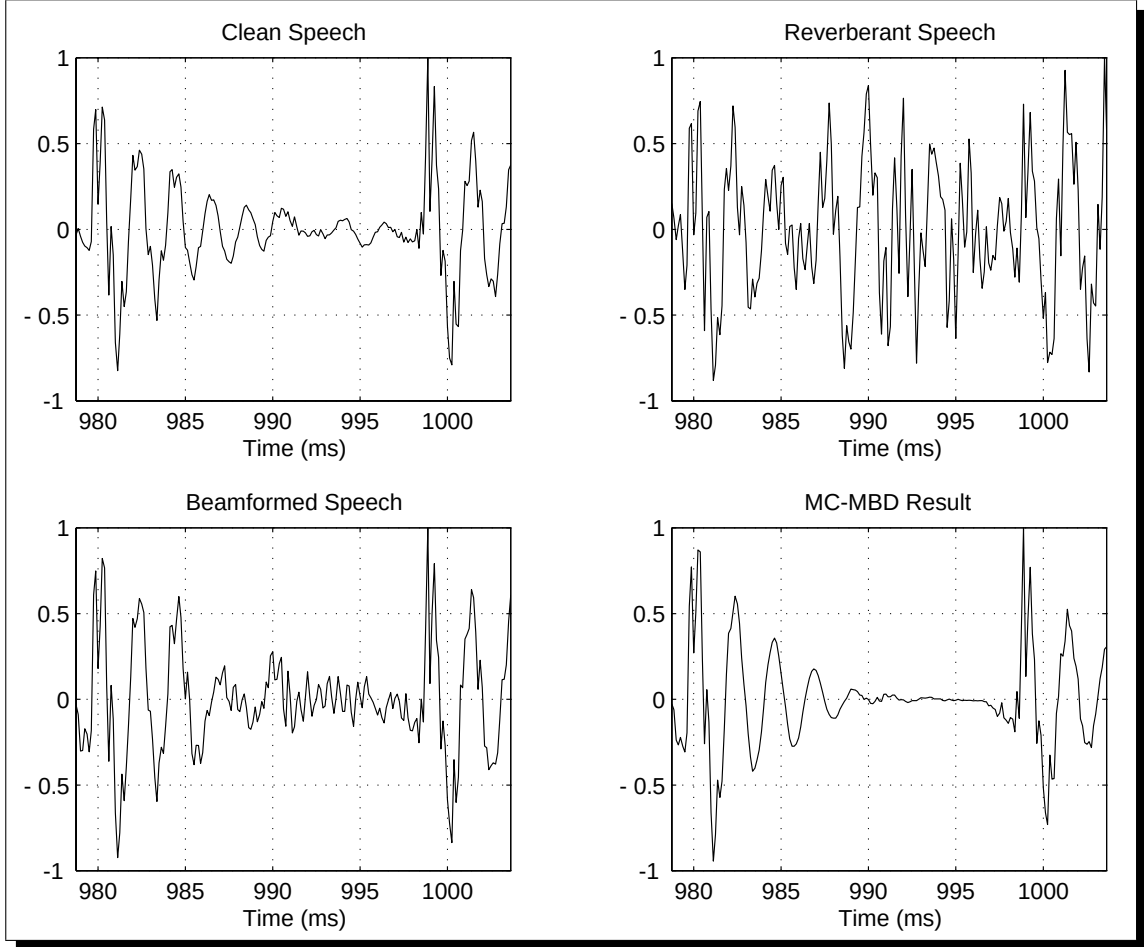


Figure 5.8: *Comparison frame 1: clean, reverberant, beamformed, and MC-MBD result*

5.4 Experimental Results

Reverberant speech segments were generated as in previous chapters in the simulated room of size $4 \text{ m} \times 4 \text{ m} \times 3 \text{ m}$. Recall that the random source locations and the 8 microphone locations were shown in Figure 3.9. Room reverberation times, T_r , ranged from 0 ms to 400 ms. The corresponding reflection coefficients used by the image model were calculated via Eyring's formula given in Equation 2.30. Again, the room was a rectangular enclosure assumed to have plane reflective surfaces and uniform, frequency-independent reflection

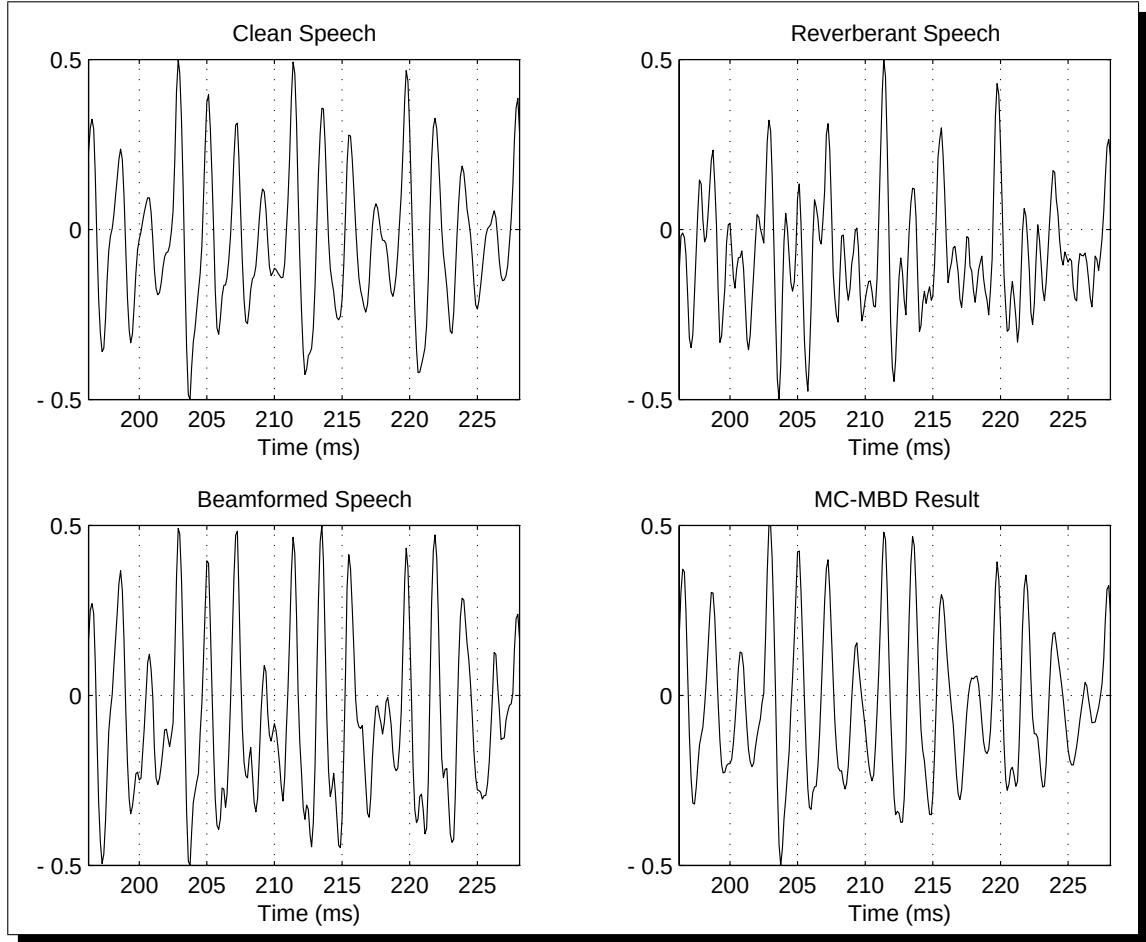


Figure 5.9: *Comparison frame 2: clean, reverberant, beamformed, and MC-MBD result*

coefficients.

Figures 5.11 illustrates a series of three second speech segments which have been simulated in the reverberant enclosure described above with a reverberation time of 400 ms. The figure plots the original speech, the signal received at a single microphone, the outcome of Delay-and-Sum Beamforming, and the result achieved using the proposed MC-MBD enhancement algorithm. The SEGSNR scores for these three plots are 0.75, 1.32, and 2.18 dB respectively. The BSD scores for the three plots are 0.76, 0.54, and 0.21 respectively. These scores show that the MC-MBD procedure is effective, a conclusion that is further

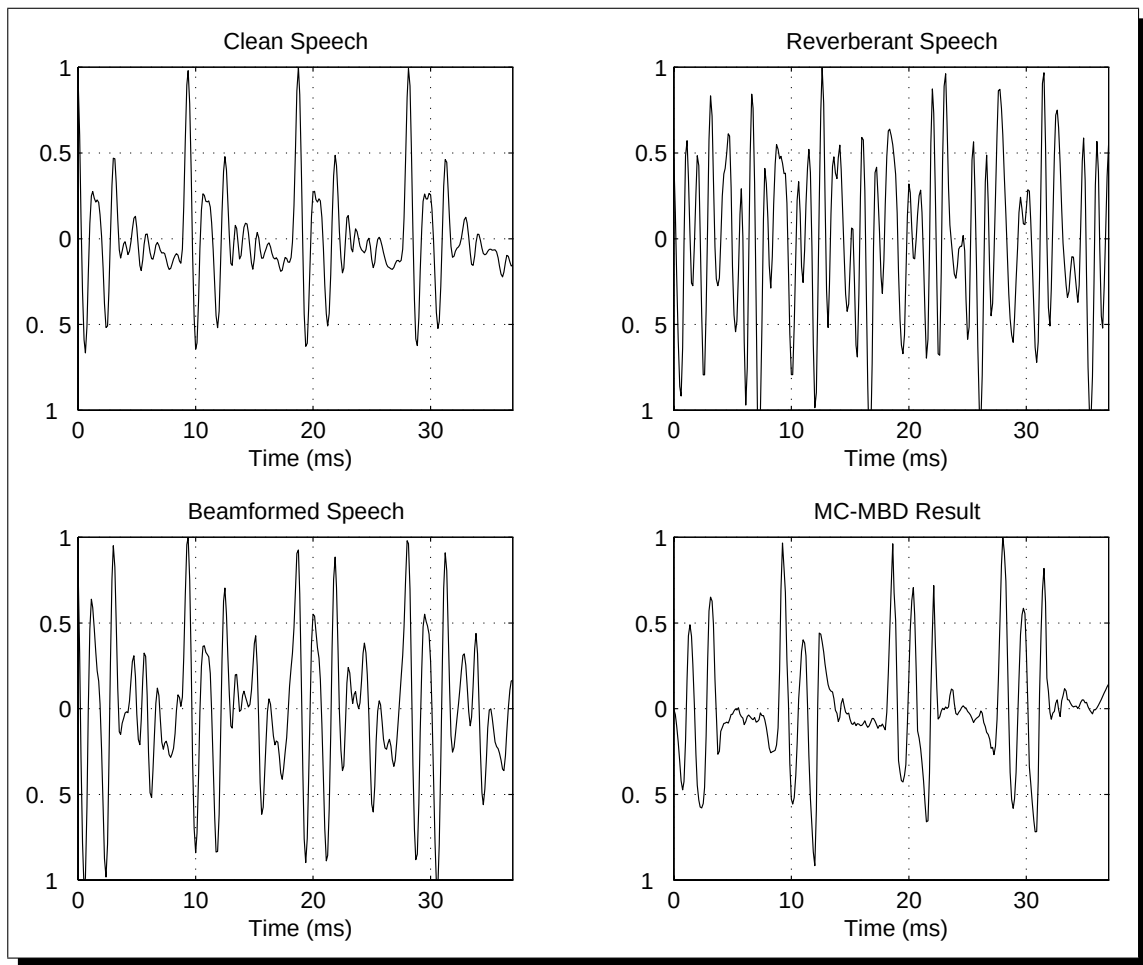


Figure 5.10: *Comparison frame 3: clean, reverberant, beamformed, and MC-MBD result*

supported by the visible and audible improvements of the resulting output waveform.

To quantify the improvements of the proposed algorithm, the average Segmental SNR and average Bark Spectral Distortion were calculated over all frames of speech for a 45 second speech signal simulated at each of the 10 random room locations and for each reverberation level. No additive noise was present for these tests so that the effects of just the MC-MBD algorithm could be examined. Figure 5.12 shows the SEGSNR and BSD results for reverberation times from 0 ms to 400 ms. Although the SEGSNR scores are very similar for beamforming and MC-MBD, the proposed algorithm shows a considerable improvement over Delay-and-Sum Beamforming when using the more perceptually motivated BSD measure. These figures are consistent with informal listening tests.

5.5 Dereverberation Summary

This chapter presented a multi-channel dereverberation algorithm which made use of a speech model and a room reverberation model. By attenuating Wavelet Transform coefficient extrema due to reverberations relative to those due to speech events, the algorithm was capable of outperforming standard Delay-and-Sum Beamforming for all examined reverberation levels.

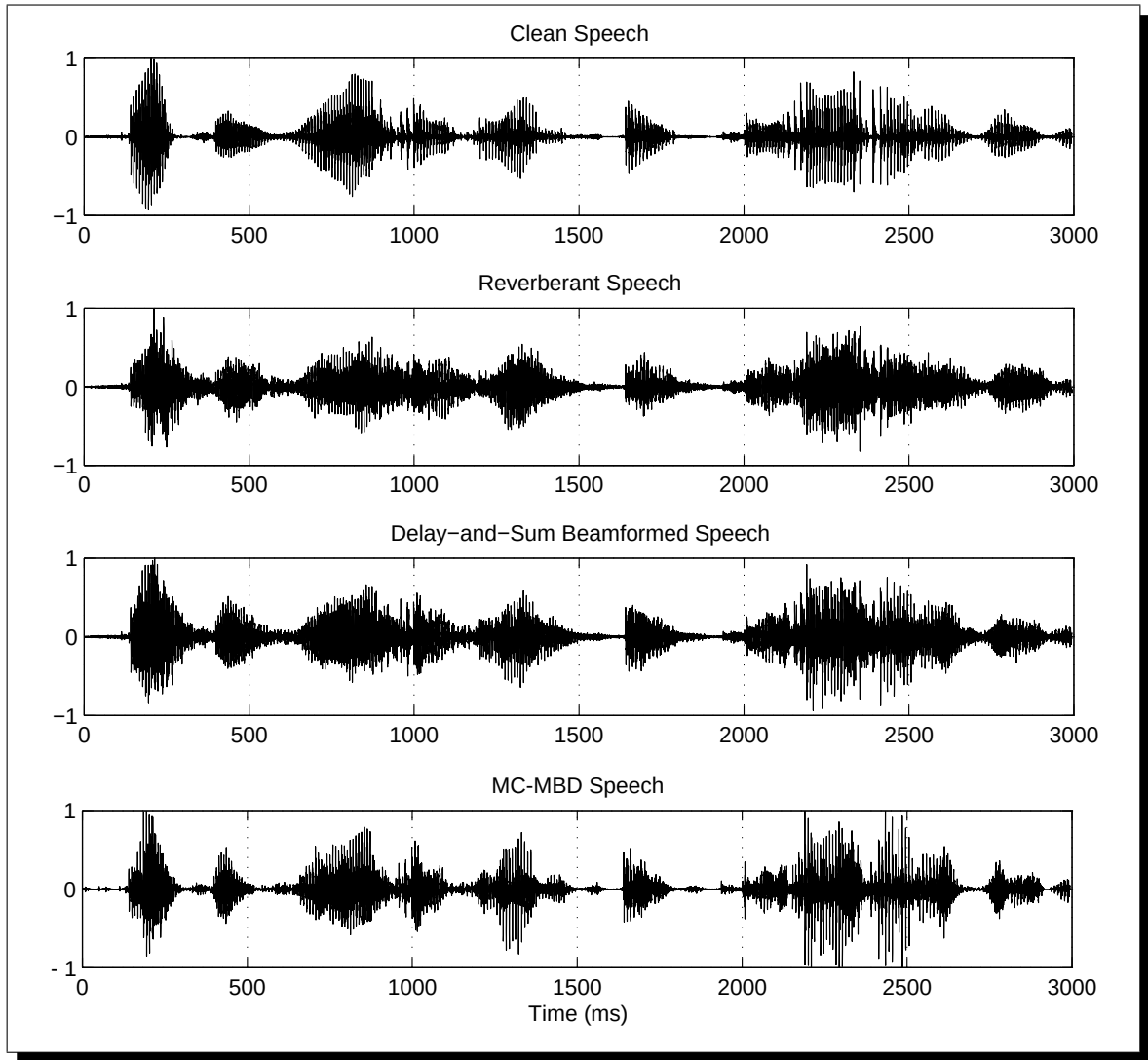
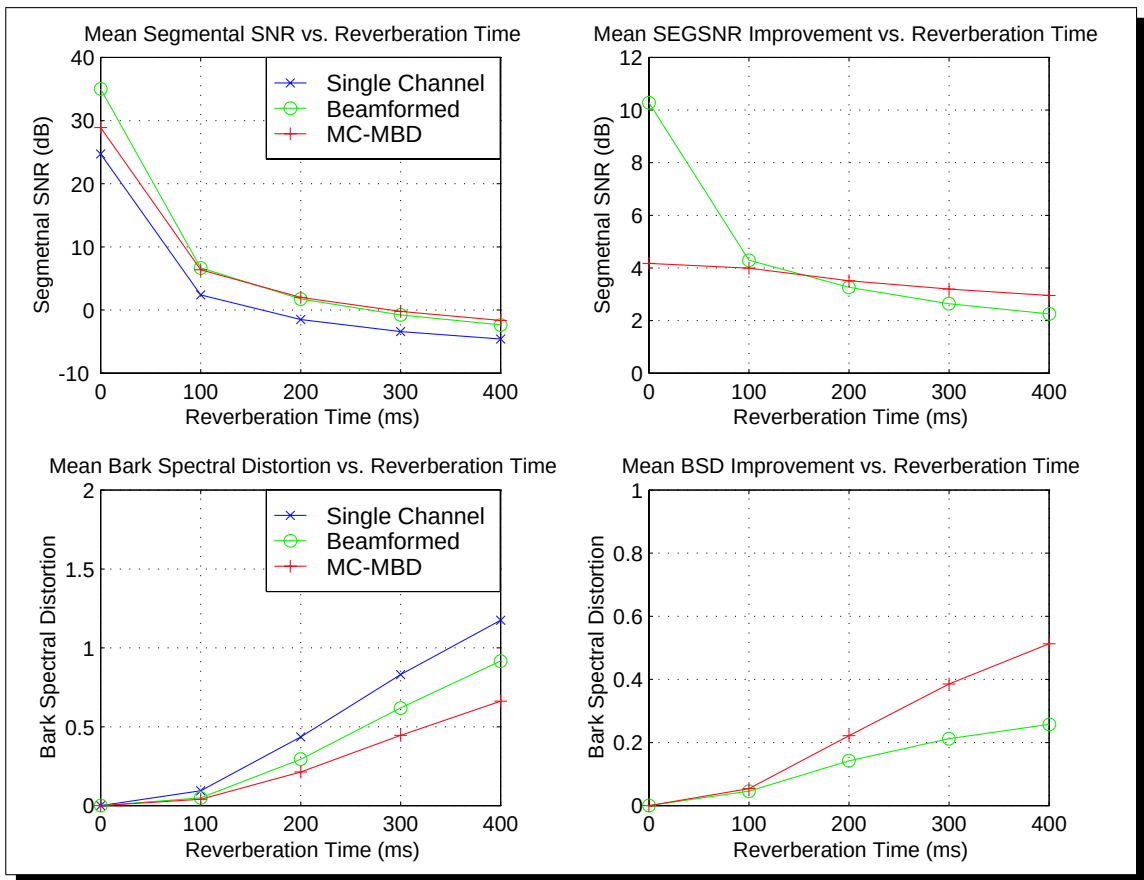


Figure 5.11: *Comparison of clean, reverberant, beamformed, and the proposed MC-MBD algorithm (Reverberation-Only Case)*

Figure 5.12: *Multi-channel dereverberation results for various reverberation times*

Chapter 6

System Performance

In this chapter, two topics involving the proposed system will be discussed. The first is more simulations similar to those presented in Chapters 4 and 5 on denoising and dereverberation. However, these new tests simultaneously corrupt the clean speech utterance with additive noise and reverberation effects. Next, a description of the Harvard University multimedia laboratory's videoconferencing system will be offered, of which the proposed speech enhancement system is a key component.

6.1 Simulations of Denoising and Dereverberation

Figure 6.1 shows the results of simultaneous denoising and dereverberation for 400 ms reverberant speech plus additive white Gaussian noise. The superior performance of the proposed algorithm relative to Delay-and-Sum beamforming at removing both long and short term reverberation effects is apparent. The algorithm helps to eliminate the additive noise as well. Listening tests confirm the effectiveness of the procedure. The synthesized speech reduces the audible reverberation and noise effects which are clearly present in both the degraded and Delay-and-Sum Beamformed speech. The SEGSNR scores for these three

plots are 0.61, 1.31, and 1.99 dB respectively. The BSD scores for the three plots are 1.63, 0.65, and 0.22 respectively.

To quantify the performance of simultaneous denoising and dereverberation, two tests were performed. The first suite of tests for speech degraded by additive noise and reverberations is shown in Figure 6.2. As in Chapters 4 and 5, these tests were run for the same 10 random source locations and for 45 seconds of speech at each location. For this test, the SNR of the additive noise was held constant at 10 dB for reverberation times varying from 0 to 400 ms. The results of the multi-channel Minimum Statistics Denoising and Model-Based Dereverberation algorithm (MC-MSD-MBD) show a dramatic improvement in both Segmental SNR and Bark Spectral Distortion scores.

The second set of tests for speech degraded by additive noise and reverberations is shown in Figure 6.3. For this test, the reverberation time was held constant at 200 ms for SNR levels varying from -10 to 10 dB. The results of the MC-MSD-MBD algorithm again show a dramatic improvement in both segmental SNR and Bark spectral distortion scores.

One point to consider is whether or not one part of the algorithm is responsible for most of the segmental SNR or Bark spectral distortion improvement in these joint denoising and dereverberation tests. Table 6.1 shows varying levels of additive noise and reverberation for one source location. Note that both the denoising (MC-MSD) and dereverberation (MC-MBD) show improvements in segmental SNR over the beamformed result. Including both the denoising and dereverberation stages results in the greatest improvement. Therefore, it can be said with confidence that both components of the system offer their own benefits to these segmental SNR score improvements. A similar statement can be made about the improvement in the Bark Spectral Distortion scores after viewing Table 6.2.

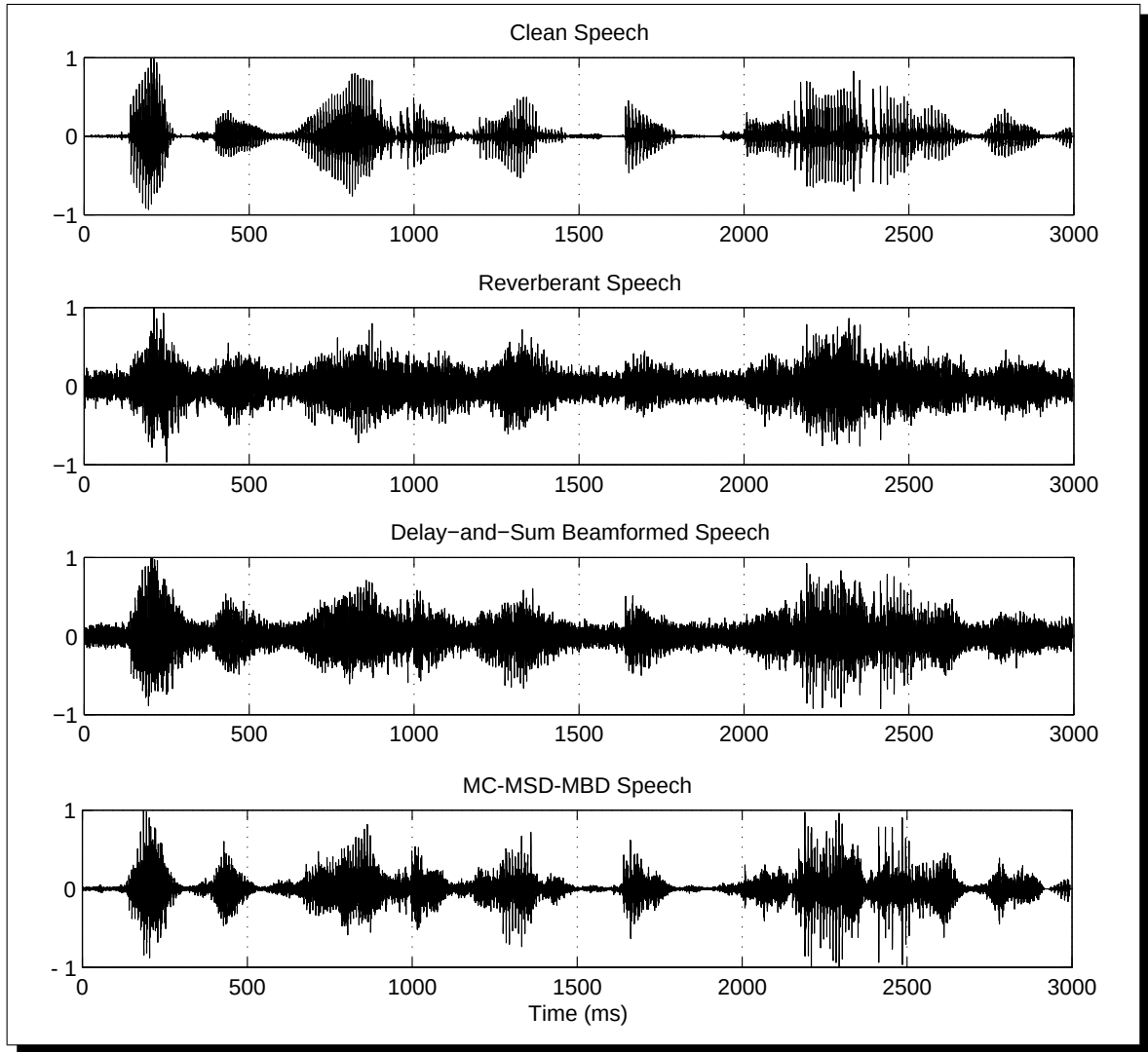


Figure 6.1: *Comparison of clean, reverberant, beamformed, and the proposed MC-MSD-MBD algorithm (reverberation plus noise case)*

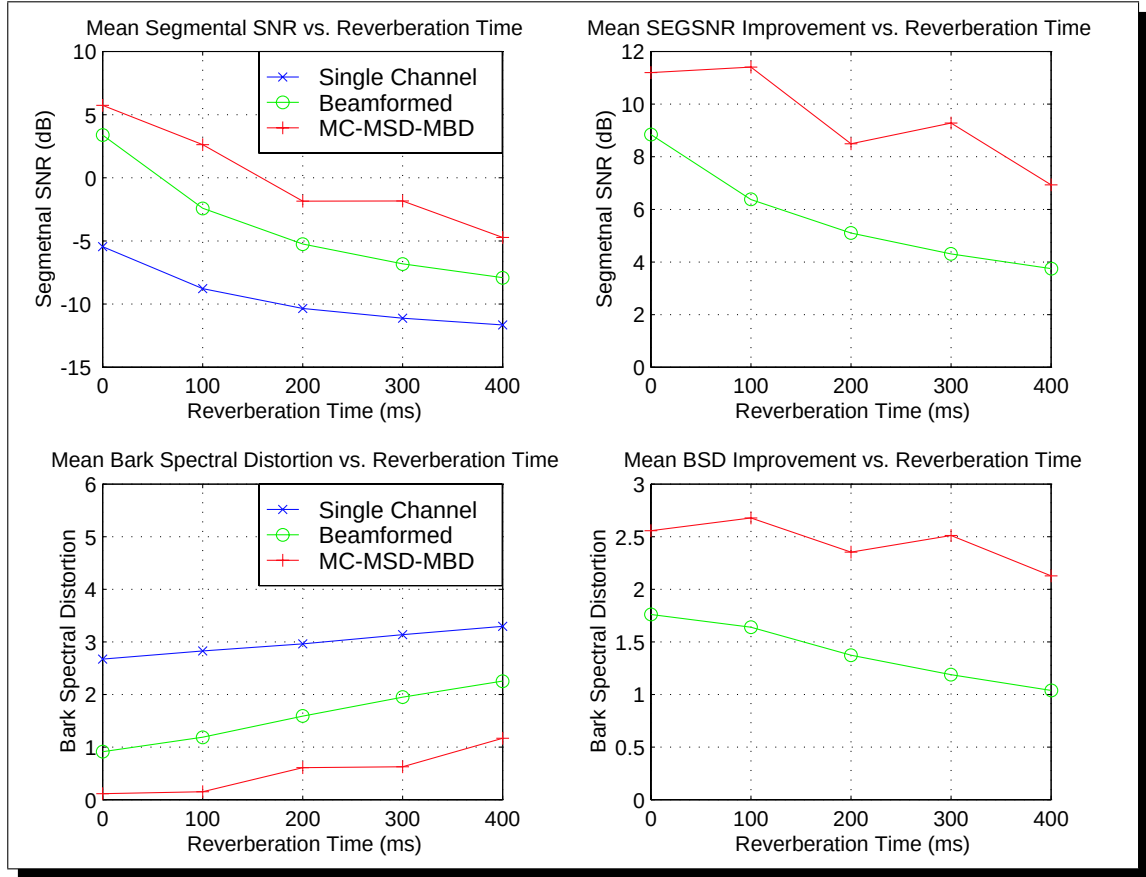


Figure 6.2: Results for varying levels of reverberation for an additive noise level of 10 dB SNR

6.2 Speech Processing for Automated Videoconferencing

The overall goal of the work at the Harvard Intelligent Multi-Media Environments Laboratory (HIMMEL) is to produce automated and relevant high-quality signal capture in environments populated with microphones and video cameras without the active supervision or distraction of its human users. In this work, the details of this system are presented for acoustic and visual-based tracking and audio signal acquisition of individuals in a practical room. See [Wang et al. (2001)] for more details.

The ability to do real-time talker tracking and audio enhancement is an important req-

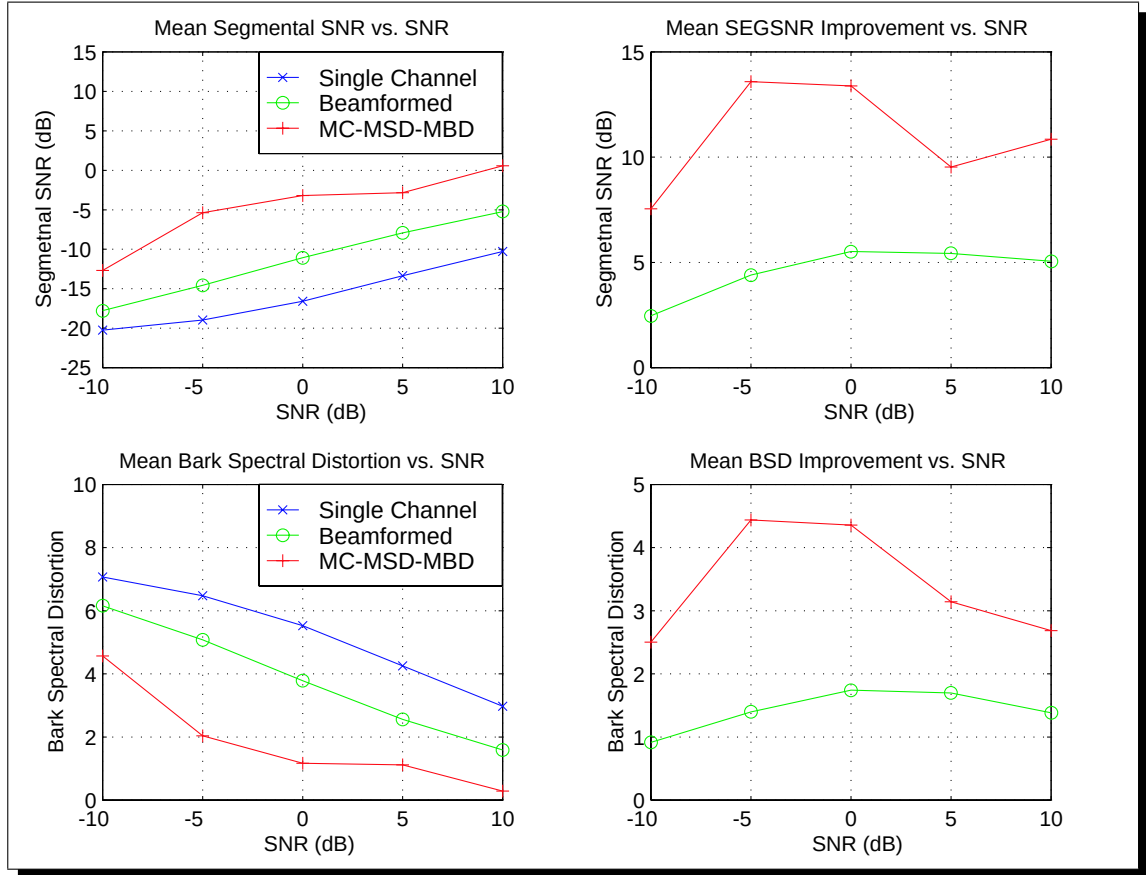


Figure 6.3: Results for varying levels of SNR for a reverberation time of 200 ms

SNR (dB)	Reverberation Time (ms)	Beamforming Result	MC-MSD Only	MC-MBD Only	MC-MSD-MBD Result
10	400	-9.91	-4.63	-5.94	-2.13
5	300	-10.93	-5.30	-6.28	-2.51
0	200	-12.69	-6.43	-7.97	-3.36
-5	100	-13.80	-8.62	-9.44	-5.11
-10	400	-19.87	-17.81	-15.95	-14.19

Table 6.1: Segmental SNR scores for varying values of SNR and reverberation levels for beamforming, independent denoising and dereverberation, and joint denoising and dereverberation

SNR (dB)	Reverberation Time (ms)	Beamforming Result	MCMS Only	MCMBD Only	MC-MSD-MBD Result
10	400	2.32	0.93	0.94	0.44
5	300	3.54	1.51	1.57	0.64
0	200	4.81	2.57	2.20	1.04
-5	100	6.20	4.07	3.75	2.17
-10	400	7.90	6.96	5.95	4.95

Table 6.2: *BSD scores for varying values of SNR and reverberation levels for beamforming, independent denoising and dereverberation, and joint denoising and dereverberation*

quisite of many developing technologies, including video-conferencing, interactive human-computer communications, automated surveillance systems, and virtual reality interfaces. As an example application of this technology, consider video-conferencing. Until recently, multi-participant systems required a human camera operator to track speakers and select shots. Earlier work involved acoustically-based talker tracking to provide rudimentary automated camera control [Brandstein and Silverman (1997), Brandstein et al. (1997)]. A number of commercial video-conferencing systems now employ these and similar technologies. However, the reliance on acoustic data alone limits performance. The system described here fuses the video and audio data to dramatically improve source tracking capability, allowing for the effective isolation of one or more desired talkers within the video frame. The quality of the associated audio signal is also important. In addition to ease of use and information conveyance considerations, the quality of the speech signal plays a key role in the performance of related analysis systems such as speech recognition and speaker identification. To this end, microphone array methods are employed to reduce the effects of ambient noise, interfering sources, and room acoustics in the resulting audio signal.

The overall system is outlined in Figure 6.4. A microphone array detects active acoustic sources in the environment and provides initial talker locations to the video portion of the system. Video processing methods are then employed to allow the computer-controlled

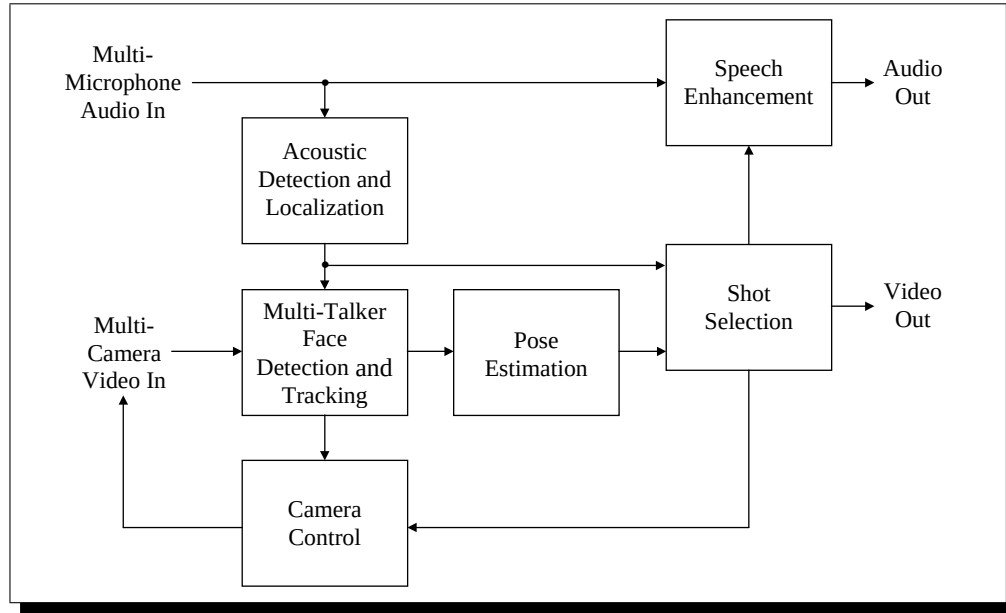


Figure 6.4: *Outline of the overall automated videoconferencing system*

cameras to isolate the image of one or more individuals within a frame, adjust for head orientation and range, and to compensate for any source motion. Source tracking information is fed to the acoustic processing element to allow for enhancement of the appropriate audio signal. The system is designed to be robust over a wide range of acoustic and visual environmental conditions. The video algorithms are computationally efficient, allowing for real-time operation on standard personal computer systems.

The remainder of this section describes the algorithms employed in the current hybrid system and illustrates their functionality. Sections 6.2.1 and 6.2.2 discuss details of the acoustic and visual components, respectively.

6.2.1 Audio Processing

The goal of the multi-channel audio system is twofold:

1. to detect acoustic activity and determine the position of a talker based on the signals

observed at microphones located throughout the room

2. to use the multi-channel audio data to produce an enhanced audio signal.

The positional information is used to determine the initial selection and steering of the appropriate video camera. The camera then takes over and tracks the active talker using the video techniques to be discussed in the video section. As time progresses, the source locator is used to detect the presence of new dominant speech sources, which leads to image re-initialization and possibly the selection of a new camera. An enhanced speech signal is estimated for the talker being tracked to accompany the video signal. The speech enhancement system proposed in this thesis will provide the acoustic processing used in the HIMMEL system. The time-delay and source localization procedures were described in Chapter 3 while the speech enhancement procedure was discussed in Chapters 4 and 5.

In the current HIMMEL system, there are a total of 16 microphones ($M = 16$) arranged in pairs positioned along the walls and in the ceiling panels of an office room. See Figure 6.5 for details of the room layout. The acoustic conditions in this room are far from ideal. Computer fans and air ducts contribute to an ambient noise factor over 70 dB. The floors are tiled and the walls are made of a non-absorbing material. Reverberation levels are on the order of 200 ms. Illumination is provided by fluorescent lighting mounted in the 3 m ceiling.

Figure 6.6 shows an example of the output of the speech enhancement procedure. The top plot shows a segment of speech received at a single microphone within the room. The bottom plot shows the enhanced signal. As is evident from the graph, the background noise has been reduced considerably.

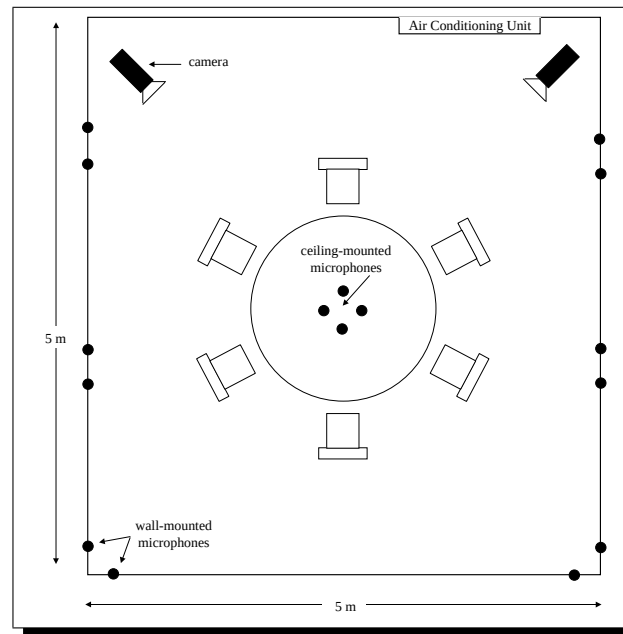


Figure 6.5: *Overhead view of the room showing microphone and camera positions*

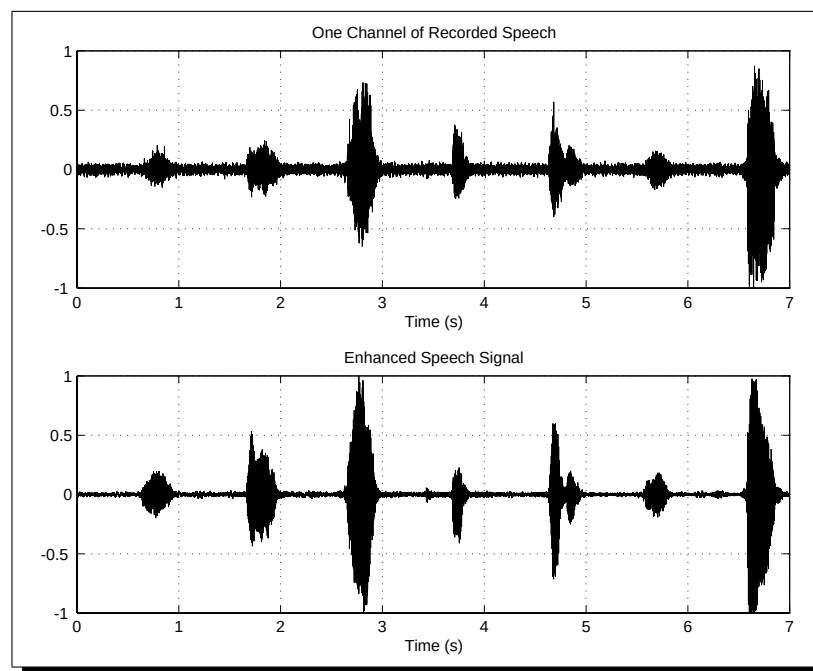


Figure 6.6: *Speech enhancement results*

6.2.2 Video Processing

The visual part of the system, developed by Wang in [Wang et al. (2001)] consists of a set of mobile communication video cameras, each with three degrees of freedom (pan, tilt, and zoom) which may be controlled remotely via a computer interface. The current system consists of two cameras positioned as shown in Figure 6.5. Given the initial camera steering provided by the acoustic localization process, the goals of the video component of the system are to:

1. detect and isolate the talker's head within the acquired image
2. track individuals as they move
3. evaluate the head orientations

The first two items require a feedback control mechanism that allows the cameras to be continually steered to keep the individual in the center of the frame with the appropriate zoom. The head orientation information is used in conjunction with the acoustic localization data to select the camera with the best view of the active talker's face.

The HIMMEL system employs a procedure which uses a variety of features that are easily extracted from the image stream. The result is a multi-person face detection procedure that is robust to a range of environmental conditions while only requiring a limited amount of computation. In order of application, the utilized features are: (1) source motion, (2) contour geometry, (3) color statistics, and (4) facial analysis. Source motion and contour geometry identify potential face regions while color and facial analysis are used as validation criteria. Figure 6.7 displays an image derived from the 4-stage face detection process [Wang et al. (2001)]. The green boxes delineate the segmented regions derived from the motion information (stage 1). In this case, three candidate locations have been identified. The blue ovals are the result of the geometrical analysis (stage 2) which generates

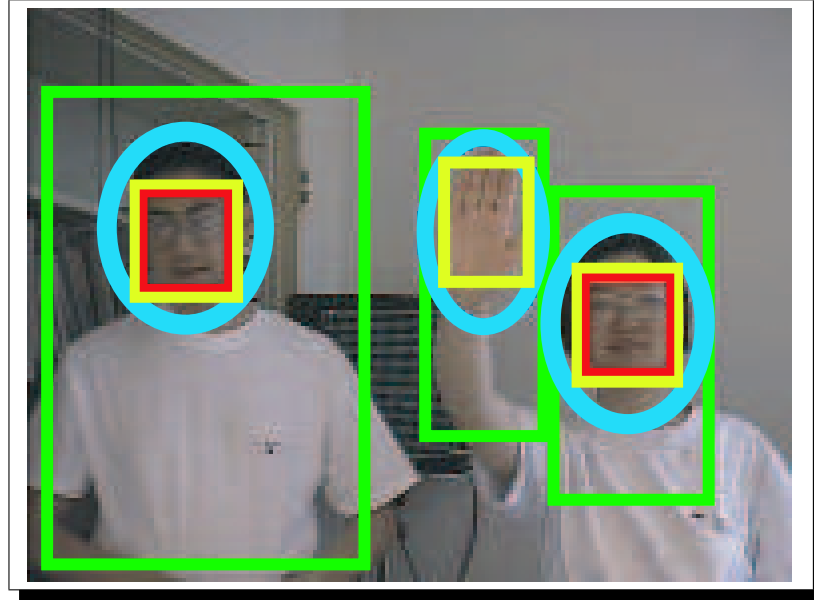


Figure 6.7: *Face tracking example [Wang et al. (2001)]*

a two-sided contour boundary for each region and then estimates a neckline location. For this example, each of the three candidates (two faces and a hand) were found to possess a head-like geometry. The color confirmation test (stage 3) is then applied within each of the ovals. The detected skin color regions are indicated by a yellow box. Up to this point, the hand is still misclassified as a face candidate due to its similar shape and color. Lastly, the facial analysis (stage 4), with positive results shown as a red box within the yellow box, correctly identifies the two faces and discards the hand.

Once a face has been detected, a tracking algorithm is activated to predict the position and velocity of the target in the subsequent frames. The predicted position estimate is used to reduce the tracking search region, thereby decreasing processing requirements and providing a further resistance to noise. A very general model of source motion is used, followed by a visual tracking procedure detailed in [Wang et al. (2001)].

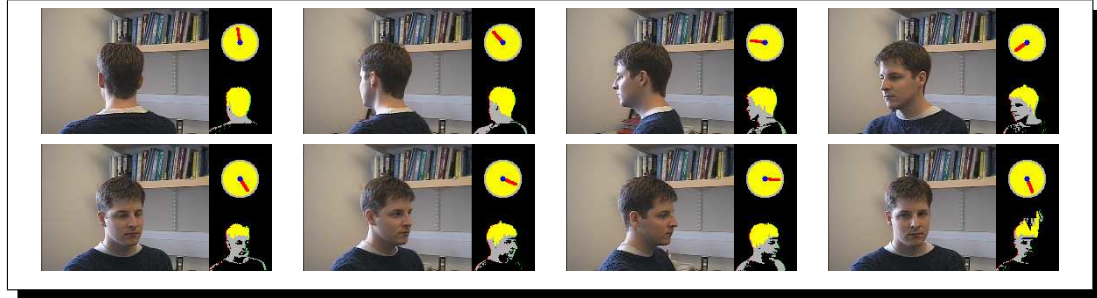


Figure 6.8: *Examples of face orientation estimation where the detected hair region is yellow on the bottom right of each figure and the clock face indicates the estimated pose angle [Wang et al. (2001)]*

The final stage of the video component addresses the problem of automatically estimating an individual's head orientation. This information is useful for determining shot relevance and deciding which camera best captures the scene of interest. The head pose algorithm developed by Wang focuses on the border between an individual's hair and skin as the major feature for estimating head orientation. The hair-skin border is straightforward to estimate across a wide range of viewing angles and is relatively insensitive to lighting conditions and poor image quality. The curve shape of the hairline is sufficient for estimating head orientation to a satisfactory degree and is a robust and consistent statistic across a variety of people and hairstyles. Figure 6.8 illustrates the pose estimation results achieved for one individual [Wang et al. (2001)]. By determining which camera the talker is facing, automatic camera selection can be performed. This process can be made even more reliable by also using the SNR estimates of the various microphones. A higher SNR value generally indicates that a talker is speaking more towards that microphone than towards the others.

6.3 System Summary

This chapter presented results for speech simultaneously degraded by additive noise and room reverberations. An automatic videoconferencing system being developed at Harvard

University's multimedia lab was also discussed which makes use of the speech enhancement system presented in this thesis. The videoconferencing system automatically tracked a speech source while simultaneously providing an enhanced audio signal to accompany the video stream. The audio and video output of this system could be combined for transmission in a videoconferencing or video recording setting.

Chapter 7

Conclusion

7.1 Discussion

In this thesis, a microphone array speech enhancement system was presented for signals acquired in challenging environments (i.e., environments with background noise and reverberations). The proposed system had three main components: source localization, background noise reduction, and dereverberation. The localization algorithm was robust to both background noise and reverberations and simultaneously produced relative time delay estimates and a source location estimate. It provided a procedure applicable to all time delay estimators which either maximize or minimize an appropriate objective function, improving their robustness to environmental degradations. The denoising algorithm was a multi-microphone extension to the Minimum Statistics denoising technique [Martin (2001)]. This algorithm also had an additional and optional SNR-dependent beamforming stage that was shown to be very useful in certain environments. This was especially useful for situations with moving speech and noise sources. The final component of the system was a multi-channel model-based dereverberation algorithm which modeled the speech source and room reverberations independently. By focusing on residual onset data and incorpo-

rating specific knowledge of the nature of the speech signal, the proposed wavelet-domain algorithm is capable of discriminating impulses of the excitation residual generated by the desired speech signal from those brought about by reverberations. The process achieves a significant attenuation of environmental reverberations and additive noise without explicitly requiring estimation of the channel characteristics. This illustrates the effectiveness and practicality of the advocated fusion of nonlinear and model-based processing for microphone array applications. Results for the various components were provided as proof of the effectiveness of the proposed multi-microphone speech enhancement system.

7.2 Future Work

In future work, effort will be focused on three areas. The first is to produce a real-time system incorporating all three stages of the enhancement system presented in this thesis. Although the denoising system is easily implemented in real time, the time delay estimation and dereverberation procedures are quite computationally intensive. These algorithms must be streamlined in order to obtain real-time status.

The second avenue for future work is to test the proposed system with Automatic Speech Recognition (ASR) software. ASR is becoming an increasingly popular topic. The uses vary from speech transcription to phone menu navigation systems. In [Hall (2001)] it was shown that when using ASR scores with Hidden Markov Models on phonemes, the correlation coefficient with Mean Opinion Scores was $\rho = 0.77$. When using commercially available continuous speech recognition software, the correlation was $\rho = 0.76$. Therefore, it would be interesting to set up a testing methodology to compare the MC-MSD-MBD enhancement system to beamforming approaches using this metric.

Finally, the principles presented in this thesis may be extended to address the case of interfering sources. Some qualitative results were shown in Chapter 3 that revealed the

ability of the source localization algorithm to detect two simultaneous sources. The multi-microphone array could then be steered towards the desired source, while placing a null in the beam pattern in the direction of the competing source. In terms of the dereverberation model used in this thesis, the impulse events may be associated with one or more source locations by evaluating their relative delays across the channels given knowledge of the microphone placements. Once this association is performed, the individual speech signals may be independently reconstructed through the synthesis procedure outlined in Chapter 5. This approach to speaker isolation represents a distinct contrast to the spatial filtering paradigm which relies on very specific knowledge and assumptions regarding talkers' locations and radiation patterns to generate appropriate channel weightings. The procedure, by virtue of its underlying speech model and exploitation of impulse data alone, has the potential to be much less sensitive to environmental uncertainties (e.g. imperfect source localization, non-ideal radiator effects, and unknown channels).

Appendix A

The Multi-Channel Maximum Likelihood Time Delay Estimator

To find the multi-channel Maximum Likelihood (ML) estimate of the time delay vector, $\boldsymbol{\tau}$, it is assumed that the speech signal, s , is Gaussian. As derived in Chapter 3, the multi-channel ML estimate of $\boldsymbol{\tau}$ is that which maximizes the following:

$$\ln\{p(\mathbf{X}|\boldsymbol{\Sigma})\} = k - \frac{1}{2} \ln |\boldsymbol{\Sigma}(\omega)| - \frac{1}{2} \int_{-\infty}^{\infty} \mathbf{X}^H(\omega) \boldsymbol{\Sigma}^{-1}(\omega) \mathbf{X}(\omega) d\omega \quad (\text{A.1})$$

where k is a constant not involving $\boldsymbol{\Sigma}$ and can therefore be ignored.

Before proceeding with the maximization process, it must be shown that the determinant of $\boldsymbol{\Sigma}$ is not a function of the delay vector $\boldsymbol{\tau}$. This will allow the term $-\frac{1}{2} \ln |\boldsymbol{\Sigma}(\omega)|$ to be ignored in the likelihood function when searching for the optimal $\boldsymbol{\tau}$. To this end, the following lemma is presented:

Lemma If $\mathbf{A}_1$ is nonsingular, then:

$$\begin{vmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_3 & \mathbf{A}_4 \end{vmatrix} = |\mathbf{A}_1| |\mathbf{A}_4 - \mathbf{A}_3 \mathbf{A}_1^{-1} \mathbf{A}_2| \quad (\text{A.2})$$

for suitably sized matrices $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, and $\mathbf{A}_4$ where $|\cdot|$ is the determinant operator.

Proof Premultiply the matrix as follows:

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{A}_3\mathbf{A}_1^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_3 & \mathbf{A}_4 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{0} & -\mathbf{A}_3\mathbf{A}_1^{-1}\mathbf{A}_2 + \mathbf{A}_4 \end{bmatrix}. \quad (\text{A.3})$$

The determinant of any block triangular matrix is the product of the determinants of the diagonal blocks. Therefore, taking the determinants of both sides of this equation yields:

$$(1) \begin{vmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_3 & \mathbf{A}_4 \end{vmatrix} = |\mathbf{A}^{-1}| |\mathbf{A}_4 - \mathbf{A}_3\mathbf{A}_1^{-1}\mathbf{A}_2| \quad (\text{A.4})$$

which concludes the proof of the lemma. ■

Theorem $|\Sigma|$ is not a function of the delay vector, $\boldsymbol{\tau}$.

Proof Let Σ_n represent the matrix Σ for n channels. Let the hypothesis be that:

$$|\Sigma_n| = (\sigma_u^2)^n + n(\sigma_u^2)^{n-1}G_{s,s} \quad (\text{A.5})$$

where the dependence on ω has been ignored. Clearly, this expression for $|\Sigma_n|$ does not involve $\boldsymbol{\tau}$. The validity of this expression will be shown by induction. First, consider the cases where $n = 1$ and $n = 2$. For $n = 1$, there is only one microphone signal, and so the matrix Σ_1 is actually a scalar equal to $G_{s,s} + \sigma_u^2$ which is not a function of τ_1 . For $n = 2$, Σ_2 has the form:

$$\Sigma_2 = \begin{bmatrix} G_{s,s} + \sigma_u^2 & G_{s,s}e^{-j\omega(\tau_1 - \tau_2)} \\ G_{s,s}e^{j\omega(\tau_1 - \tau_2)} & G_{s,s} + \sigma_u^2 \end{bmatrix}. \quad (\text{A.6})$$

Therefore, the determinant is:

$$|\Sigma_2| = (G_{s,s} + \sigma_u^2)^2 - |G_{s,s}|^2 \quad (\text{A.7})$$

$$= (\sigma_u^2)^2 + 2\sigma_u^2 G_{s,s} \quad (\text{A.8})$$

which does not depend on τ_1 or τ_2 . Now assume that the hypothesis is true for $i = n$ and prove that it must then hold for $i = n + 1$. Form Σ_{n+1} from Σ_n :

$$\Sigma_{n+1} = \begin{bmatrix} \Sigma_n & A_2 \\ A_3 & A_4 \end{bmatrix} \quad (\text{A.9})$$

where

$$A_2 = G_{s,s} \left[e^{-j\omega(\tau_1 - \tau_{n+1})} \dots e^{-j\omega(\tau_n - \tau_{n+1})} \right]^T \quad (\text{A.10})$$

$$A_3 = A_2^H = G_{s,s} \left[e^{j\omega(\tau_1 - \tau_{n+1})} \dots e^{j\omega(\tau_n - \tau_{n+1})} \right] \quad (\text{A.11})$$

$$A_4 = G_{s,s} + \sigma_u^2. \quad (\text{A.12})$$

Using the block matrix determinant lemma from above, we know that:

$$|\Sigma_{n+1}| = |\Sigma_n| |G_{s,s} + \sigma_u^2 - A_3 \Sigma_n^{-1} A_2|. \quad (\text{A.13})$$

Using the closed form expression for Σ_n^{-1} to be presented shortly, it can be shown that $A_3 \Sigma_n^{-1} A_2 = \frac{n|G_{s,s}|^2}{\sigma_u^2 + nG_{s,s}}$. Using this fact and the hypothesized determinant for Σ_n , we have

$$\begin{aligned} |\Sigma_{n+1}| &= ((\sigma_u^2)^n + n(\sigma_u^2)^{n-1}G_{s,s}) \left(G_{s,s} + \sigma_u^2 - \frac{n|G_{s,s}|^2}{\sigma_u^2 + nG_{s,s}} \right) \\ &= ((\sigma_u^2)^n + n(\sigma_u^2)^{n-1}G_{s,s}) \left(\frac{\sigma_u^2 G_{s,s} + (\sigma_u^2)^2 + n|G_{s,s}|^2 + n\sigma_u^2 G_{s,s} - n|G_{s,s}|^2}{\sigma_u^2 + nG_{s,s}} \right) \\ &= (\sigma_u^2)^{n-1} (\sigma_u^2 G_{s,s} + (\sigma_u^2)^2 + n\sigma_u^2 G_{s,s}) \\ &= (\sigma_u^2)^{n+1} + (n+1)(\sigma_u^2)^n G_{s,s} \end{aligned} \quad (\text{A.14})$$

which is the hypothesized expression for $|\Sigma_{n+1}|$, completing the proof by induction. ■

Now it must be shown that the inverse of Σ has a closed form. Before doing so, the following lemma is needed.

Lemma $\Sigma_\tau^2 = \Sigma_\tau \Sigma_\tau = (M-1)\mathbf{I} + (M-2)\Sigma_\tau$ where M is the number of microphone

signals and Σ_τ is as defined in Equation 3.32.

Proof First, calculate the values of the diagonal elements of Σ_τ^2 :

$$[\Sigma_\tau^2]_{i,i} = \sum_{j=1, j \neq i}^M e^{-j\omega(t_i-t_j)} e^{j\omega(t_i-t_j)} = M-1. \quad (\text{A.15})$$

The values of the off-diagonal elements are:

$$[\Sigma_\tau^2]_{i,j} = \sum_{k=1, k \neq i, k \neq j}^M e^{-j\omega(t_i-t_k)} e^{j\omega(t_j-t_k)} \quad (\text{A.16})$$

$$= \sum_{k=1, k \neq i, k \neq j}^M e^{-j\omega(t_i-t_j)} \quad (\text{A.17})$$

$$= (M-2)e^{-j\omega(t_i-t_j)}. \quad (\text{A.18})$$

Therefore, $\Sigma_\tau^2 = (M-1)\mathbf{I} + (M-2)\Sigma_\tau$. ■

Theorem The inverse of Σ_M is given by:

$$\Sigma_M^{-1} = \frac{|\Sigma_{M-1}|}{|\Sigma_M|} \mathbf{I} - \frac{(\sigma_u^2)^{M-2} G_{s,s}}{|\Sigma_M|} \Sigma_\tau \quad (\text{A.19})$$

Proof We have the following:

$$\begin{aligned} \Sigma_M^{-1} \Sigma_M &= \left[\frac{|\Sigma_{M-1}|}{|\Sigma_M|} \mathbf{I} - \frac{(\sigma_u^2)^{M-2} G_{s,s}}{|\Sigma_M|} \Sigma_\tau \right] [(\sigma_u^2 + G_{s,s})\mathbf{I} + G_{s,s}\Sigma_\tau] \quad (\text{A.20}) \\ &= \frac{|\Sigma_{M-1}|(\sigma_u^2 + G_{s,s})}{|\Sigma_M|} \mathbf{I} + \frac{|\Sigma_{M-1}|G_{s,s}}{|\Sigma_M|} \Sigma_\tau \\ &\quad - \frac{(\sigma_u^2 + G_{s,s})(\sigma_u^2)^{M-2} G_{s,s}}{|\Sigma_M|} \Sigma_\tau - \frac{(\sigma_u^2)^{M-2} |G_{s,s}|^2}{|\Sigma_M|} [(M-1)\mathbf{I} + (M-2)\Sigma_\tau] \\ &= \left[\frac{|\Sigma_{M-1}|(\sigma_u^2 + G_{s,s}) - (M-1)(\sigma_u^2)^{M-2} |G_{s,s}|^2}{|\Sigma_M|} \right] \mathbf{I} \\ &\quad + \left[\frac{|\Sigma_{M-1}|G_{s,s} - (\sigma_u^2 + G_{s,s})(\sigma_u^2)^{M-2} G_{s,s} - (M-2)(\sigma_u^2)^{M-2} |G_{s,s}|^2}{|\Sigma_M|} \right] \Sigma_\tau \\ &\triangleq a\mathbf{I} + b\Sigma_\tau. \end{aligned}$$

It can be shown with straightforward algebra that $a = 1$ and $b = 0$ so that $\Sigma_M^{-1} \Sigma_M = \mathbf{I}$. ■

Now that it was shown that $|\mathbf{\Sigma}_M(\omega)|$ does not depend on the delay vector $\boldsymbol{\tau}$, it can safely be omitted from the optimization process. Therefore, the optimal ML delay vector $\hat{\boldsymbol{\tau}}_{ML}$ must maximize the following function:

$$\begin{aligned} f(\boldsymbol{\tau}) &= -\frac{1}{2} \int_{-\infty}^{\infty} \mathbf{X}^H(\omega) \mathbf{\Sigma}_M^{-1}(\omega, \boldsymbol{\tau}) \mathbf{X}(\omega) d\omega \\ &= -\frac{1}{2} \int_{-\infty}^{\infty} \frac{|\mathbf{\Sigma}_{M-1}(\omega)|}{|\mathbf{\Sigma}_M(\omega)|} \mathbf{X}^H(\omega) \mathbf{X}(\omega) - \frac{(\sigma_u^2)^{M-2} G_{s,s}(\omega)}{|\mathbf{\Sigma}_M(\omega)|} \mathbf{X}^H(\omega) \mathbf{\Sigma}_{\boldsymbol{\tau}}(\omega) \mathbf{X}(\omega) d\omega. \end{aligned} \quad (\text{A.21})$$

However, the first term in the integral and the constant factor $\frac{1}{2}$ can be ignored because they do not depend on $\boldsymbol{\tau}$. Therefore, the optimal $\hat{\boldsymbol{\tau}}_{ML}$ maximizes:

$$\begin{aligned} \phi(\boldsymbol{\tau}) &= \int_{-\infty}^{\infty} \frac{(\sigma_u^2)^{M-2} G_{s,s}(\omega)}{|\mathbf{\Sigma}_M(\omega)|} \mathbf{X}^H(\omega) \mathbf{\Sigma}_{\boldsymbol{\tau}} \mathbf{X}(\omega) d\omega \\ &= \int_{-\infty}^{\infty} \frac{(\sigma_u^2)^{M-2} G_{s,s}(\omega)}{|\mathbf{\Sigma}_M(\omega)|} \sum_{i=1}^M \sum_{j=1}^M X_i^*(\omega) [\mathbf{\Sigma}_{\boldsymbol{\tau}}]_{i,j} X_j(\omega) d\omega \\ &= \int_{-\infty}^{\infty} \frac{(\sigma_u^2)^{M-2} G_{s,s}(\omega)}{|\mathbf{\Sigma}_M(\omega)|} \sum_{i=1}^M \sum_{j=1, j \neq i}^M X_i(\omega) e^{j\omega(\tau_i - \tau_j)} X_j^*(\omega) d\omega \\ &= \sum_{i=1}^M \sum_{j=1, j \neq i}^M \left[\int_{-\infty}^{\infty} \frac{(\sigma_u^2)^{M-2} G_{s,s}(\omega)}{|\mathbf{\Sigma}_M(\omega)|} X_i(\omega) X_j^*(\omega) e^{j\omega(\tau_i - \tau_j)} d\omega \right]. \end{aligned} \quad (\text{A.22})$$

Therefore, the maximum likelihood time delay estimator for a microphone array with M sensors is:

$$\hat{\boldsymbol{\tau}}_{ML} = \underset{\boldsymbol{\tau} \in \mathbb{R}^M}{\operatorname{argmax}} \sum_{i=1}^M \sum_{j=1, j \neq i}^M \left[\int_{-\infty}^{\infty} \frac{(\sigma_u^2)^{M-2} G_{s,s}(\omega)}{|\mathbf{\Sigma}_M(\omega)|} X_i(\omega) X_j^*(\omega) e^{j\omega(\tau_i - \tau_j)} d\omega \right]. \quad (\text{A.23})$$

That is, the ML estimator is the sum of $\binom{M}{2}$ generalized cross correlation functions between all pairwise combinations of microphones, each using the same weighting function:

$$W(\omega) = \frac{(\sigma_u^2)^{M-2} G_{s,s}(\omega)}{|\mathbf{\Sigma}_M(\omega)|} = \frac{(\sigma_u^2)^{M-2} G_{s,s}(\omega)}{(\sigma_u^2)^M + M(\sigma_u^2)^{M-1} G_{s,s}(\omega)} = \frac{G_{s,s}(\omega)}{(\sigma_u^2)^2 + M\sigma_u^2 G_{s,s}(\omega)}. \quad (\text{A.24})$$

For the case where $M = 2$, the result of [Knapp and Carter (1976)] is obtained.

Bibliography

- [Abel and Smith (1989)] ABEL, J., and SMITH, J. (1989). Source range and depth estimation from multipath range difference measurements. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-37(8):1157–1165.
- [Adcock (2001)] ADCOCK, J. (2001). *Optimal Filtering and Speech Recognition with Microphone Arrays*. Ph.D. thesis, Brown University, Providence, RI.
- [Affes and Grenier (1997)] AFFES, S., and GRENIER, Y. (1997). A signal subspace tracking algorithm for microphone array processing of speech. *IEEE Trans. Speech Audio Proc.*, 5(5):425–437.
- [Allen and Berkley (1979)] ALLEN, J. B., and BERKLEY, D. A. (1979). Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, 64(4):943–950.
- [Ananthapadmanabha and Yegnanarayana (1975)] ANANTHAPADMANABHA, T. V., and YEGNANARAYANA, B. (1975). Epoch extraction of voiced speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(6):562–570.
- [Atal and Rabiner (1976)] ATAL, B., and RABINER, L. (1976). A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-24:201–212.
- [Atal and Remde (1982)] ATAL, B. S., and REMDE, J. R. (1982). A new model of LPC excitation for producing natural-sounding speech at low bit rates. In *Proceedings of ICASSP82*, pages 614–617. IEEE.

- [Bees et al. (1991)] BEES, D., BLOSTEIN, M., and KABAL, P. (1991). Reverberant speech enhancement using cepstral processing. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 977–980.
- [Benesty (2000)] BENESTY, J. (2000). Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. *J. Acoust. Soc. Am.*, 107(1):384–391.
- [Bervman and Baras (1993)] BERVMAN, Z., and BARAS, J. S. (1993). Properties of the multiscale maxima and zero-crossings representations. *IEEE Transactions on Signal Processing*, 41(12):3216–3231.
- [Boll (1979)] BOLL, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27(2):113–120.
- [Brandstein (1997)] BRANDSTEIN, M. (1997). A pitch-based approach to time-delay estimation of reverberant speech. In *1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York.
- [Brandstein (1998)] BRANDSTEIN, M. (1998). On the use of explicit speech modeling in microphone array applications. In *ICASSP98*, pages 3613–3616, Seattle, WA. IEEE.
- [Brandstein (1999a)] BRANDSTEIN, M. (1999a). An event-based method for microphone array speech enhancement. In *ICASSP99*, pages 953–956, Phoenix, AZ. IEEE.
- [Brandstein (1999b)] BRANDSTEIN, M. (1999b). Time-delay estimation of reverberated speech exploiting harmonic structure. *J. Acoust. Soc. Am.*, 105(5):2914–2919.
- [Brandstein et al. (1996)] BRANDSTEIN, M., ADCOCK, J., and SILVERMAN, H. (1996). Microphone array localization error estimation with application to sensor placement. *J. Acoust. Soc. Am.*, Vol. 99(6):3807–3816.
- [Brandstein et al. (1997)] BRANDSTEIN, M., ADCOCK, J., and SILVERMAN, H. (1997). A closed-form location estimator for use with room environment microphone arrays. *IEEE Trans. Speech Audio Proc.*, 5(1):45–50.

- [Brandstein and Silverman (1997)] BRANDSTEIN, M., and SILVERMAN, H. (1997). A practical methodology for speech source localization with microphone arrays. *Computer, Speech, and Language*, 11(2):91–126.
- [Brandstein and Ward (2001)] BRANDSTEIN, M., and WARD, D. (2001). *Microphone Arrays: Techniques and Applications*. Springer-Verlag.
- [Carmona (1995)] CARMONA, R. A. (1995). Extrema reconstructions and spline smoothing: variations on an algorithm of Mallat and Zhong. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, pages 83–94. Springer-Verlag.
- [Carter (1981)] CARTER, G. C. (1981). Time delay estimation for passive sonar signal processing. *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-29(3):463–470.
- [Champagne et al. (1996)] CHAMPAGNE, B., BÉDARD, S., and STÉPHENNE, A. (1996). Performance of time-delay estimation in the presence of room reverberation. *IEEE Trans. on SAP*, 4(2):148–152.
- [Cheng and O’Shaughnessy (1989)] CHENG, Y. M., and O’SHAUGHNESSY, D. (1989). Automatic and reliable estimation of glottal closure instant and period. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 17(12):1805–1815.
- [Cohen (1995)] COHEN, L. (1995). *Time Frequency Analysis*. Prentice Hall.
- [Cole et al. (1997)] COLE, D., MOODY, M., and SRIDHARAN, S. (1997). Position-independent enhancement of reverberant speech. *Journal of the Audio Engineering Society*, 45(3):142–147.
- [Cvetković and Vetterli (1995)] CVETKOVIĆ, Z., and VETTERLI, M. (1995). Discrete-time wavelet extrema representation: design and consistent reconstruction. *IEEE Transactions on Signal Processing*, 43(3):681–693.
- [Dahl et al. (1997)] DAHL, M., CLAEISSON, I., and NORDEBO, S. (1997). Simultaneous echo cancellation and car noise suppression employing a microphone array. In *Proceedings of ICASSP97*, pages 239–242, Munich, Germany. IEEE.

- [Deller et al. (1987)] DELLER, J. R., PROAKIS, J. G., and HANSEN, J. H. L. (1987). *Discrete-Time Processing of Speech Signals*. Prentice Hall.
- [DeSouza (1983)] DESOUBA, P. (1983). A statistical approach to the design of an adaptive self-normalizing silence detector. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-31:678–684.
- [Di Claudio et al. (2000)] DI CLAUDIO, E., PARISI, R., and ORLANDI, G. (2000). A clustering approach to multi-source localization in reverberant rooms. In *Proceedings of the 2000 IEEE Sensor Array and Multichannel Signal Processing Workshop. SAM 2000*, pages 198–201. IEEE.
- [DiBiase (2000)] DIBIASE, J. (2000). *A High Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*. PhD thesis, Brown University, Providence, RI.
- [DiBiase et al. (2001)] DIBIASE, J., SILVERMAN, H., and BRANDSTEIN, M. (2001). Robust localization in reverberant rooms. In M. Brandstein and D. Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Verlag.
- [Dorbecker and Ernst (1996)] DORBECKER, M., and ERNST, S. (1996). Combination of two-channel spectral subtraction and adaptive wiener post-filtering for noise reduction and dereverberation. In *European Signal Processing Conference (EUSIPCO-96)*, Trieste, Italy.
- [Ephraim (1992)] EPHRAIM, Y. (1992). Statistical-model-based speech enhancement systems. *Proceedings of the IEEE*, 80(10):1526–1555.
- [Ephraim and Trees (1995)] EPHRAIM, Y., and TREES, H. V. (1995). A signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 3(4):251–266.
- [Etter and Stearns (1981)] ETTER, D., and STEARNS, S. (1981). Adaptive estimation of time delays in sampled data systems. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3.2):582–587.
- [Flanagan et al. (1993)] FLANAGAN, J., SURENDRAN, A., and JAN, E. (1993). Spatially selective sound capture for speech and audio processing. *Speech Communication*, 13(1-2):207–222.

- [Flanagan and Silverman (1992)] FLANAGAN, J. L., and SILVERMAN, H. F. (1992). Material for international workshop on microphone-array systems: Theory and practice. LEMS Technical Report 113, LEMS, Division of Engineering, Brown University, Providence, RI 02912.
- [Flanagan and Silverman (1994)] FLANAGAN, J. L., and SILVERMAN, H. F. (1994). Material for international workshop on microphone-array systems: Theory and practice. Technical report, CAIP, Rutgers University, Piscataway, NJ 08855-1390.
- [Floc'h (1992)] FLOC'H, A. L. (1992). Evaluation of linear and non-linear spectral subtraction methods for enhancing noisy speech. *ESCA*, pages 131–134.
- [Furui and Sondhi (1992)] S. Furui and M. Sondhi, editors (1992). *Advances in Speech Signal Processing*. Marcel Dekker, first edition.
- [Gierl (1990)] GIERL, S. (1990). Noise reduction for speech input systems using an adaptive microphone array. In *Proceedings 22nd ISATA*, pages 517–524.
- [Griebel and Brandstein (1999)] GRIEBEL, S., and BRANDSTEIN, M. (1999). Wavelet transform extrema clustering for multi-channel speech dereverberation. In *IEEE Workshop on Acoustic Echo and Noise Control*, pages 52–55, Pocono Manor, Pennsylvania.
- [Griebel and Brandstein (2001)] GRIEBEL, S., and BRANDSTEIN, M. (2001). Microphone array speech enhancement using coarse channel modeling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 201–204, Salt Lake City, Utah.
- [Haigh and Mason (1993)] HAIGH, J., and MASON, J. (1993). Robust voice activity detection using cepstral features. *Proceedings IEEE TENCON-93*, pages 1103–1106.
- [Hall (2001)] HALL, T. A. (2001). Objective speech quality measures for internet telephony in voice over IP (VOIP) technology. *Proceedings of SPIE*, 4522.
- [Hardwick (1992)] HARDWICK, J. (1992). *The Dual Excitation Speech Model*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

- [Harris and Nelson (1993)] HARRIS, J. D., and NELSON, D. (1993). Glottal pulse alignment in voiced speech for pitch determination. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 519–522.
- [Hess (1983)] HESS, W. (1983). *Pitch Determination of Speech Signals*. Springer-Verlag.
- [Hoshuyama and Sugiyama (2001)] HOSHUYAMA, O., and SUGIYAMA, A. (2001). Robust adaptive beamforming. In M. Brandstein and D. Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Verlag.
- [Huang et al. (2000)] HUANG, Y., BENESTY, J., and ELKO, G. (2000). Microphone arrays for video camera steering. In S. Gay and J. Benesty, editors, *Acoustic Signal Processing for Telecommunication*, pages 239–259. Kluwer.
- [Johnson and Dudgeon (1993)] JOHNSON, D., and DUDGEON, D. (1993). *Array Signal Processing-Concepts and Techniques*. Prentice Hall, first edition.
- [Kadambe (1992)] KADAMBE, S. (1992). On the choice of a wavelet, and the energy and the phase distributions of the wavelet transform. In *Time-Frequency and Time-Scale Analysis*, pages 379–382. IEEE Signal Processing Society.
- [Kadambe and Boudreaux-Bartels (1992)] KADAMBE, S. K., and BOUDREAUX-BARTELS, G. F. (1992). Application of the wavelet transform for pitch detection of speech signals. *IEEE Transactions on Information Theory*, 38(2):917–924.
- [Knapp and Carter (1976)] KNAPP, C. H., and CARTER, G. C. (1976). The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-24(4):320–327.
- [Kuttruff (1991)] KUTTRUFF, H. (1991). *Room Acoustics*. Elsevier, London, third edition.
- [Laroche et al. (1993)] LAROCHE, J., STYLIANOU, Y., and MOULINES, E. (1993). HNS: Speech modification based on a harmonic + noise model. In *Proceedings of ICASSP93*, pages II-550–II-553. IEEE.

- [Lee and Un (1986)] LEE, H., and UN, C. (1986). A study of on-off characteristics of conversational speech. *IEEE Transactions on Communications*, COM-34:630–637.
- [Lim (1983)] J. Lim, editor (1983). *Speech Enhancement*. Prentice Hall.
- [Lim and Oppenheim (1978)] LIM, J., and OPPENHEIM, A. (1978). All-pole modeling of degraded speech. *IEEE Trans. Acoust., Speech, Signal Processing*, 26:197–210.
- [Liu et al. (1995)] LIU, Q. G., CHAMPAGNE, B., and KABAL, P. (1995). Room speech dereverberation via minimum-phase and all-pass component processing of multi-microphone signals. In *IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing*, pages 571–574.
- [Liu et al. (1996)] LIU, Q. G., CHAMPAGNE, B., and KABAL, P. (1996). A microphone array processing technique for speech enhancement in a reverberant space. *Speech Communication*, 18:317–334.
- [Mahmoudi and Drygajlo (1998)] MAHMOUDI, D., and DRYGAJLO, A. (1998). Combined Wiener and coherence filtering in wavelet domain for microphone array speech enhancement. In *Proceedings of ICASSP98*, pages 385–388, Seattle, WA. IEEE.
- [Mallat (1998)] MALLAT, S. (1998). *A Wavelet Tour of Signal Processing*. Academic Press.
- [Mallat and Zhong (1992)] MALLAT, S. G., and ZHONG, S. (1992). Characterization of signals from multiscale edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(7):710–732.
- [Mansour and Jr. (1981)] MANSOUR, D., and JR., A. G. (1981). Unconstrained frequency-domain adaptive filter. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-29(3):561–571.
- [Markhoul (1993)] MARKHOUL, J. (1993). Linear prediction: A tutorial review. *Proc. IEEE*, 63(4):561–580.
- [Marro et al. (1998)] MARRO, C., MAHIEUX, Y., and SIMMER, K. (1998). Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Trans. Speech Audio Proc.*, 6(3):240–259.

- [Martin (1993)] MARTIN, R. (1993). An efficient algorithm to estimate the instantaneous SNR of speech signals. In *EUROSPEECH-93*, pages 1093–1096, Berlin.
- [Martin (1994)] MARTIN, R. (1994). Spectral subtraction based on minimum statistics. In *EUSIPCO-94*, pages 1182–1185, Edinburgh, Scotland.
- [Martin (2001)] MARTIN, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, 9 (5):504–512.
- [Martin and Vary (1994)] MARTIN, R., and VARY, P. (1994). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *Annals of Telecommunications*, 49(7-8):429–438.
- [McAulay and Malpass (1980)] MCAULAY, R., and MALPASS, M. (1980). Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28:137–145.
- [McAulay and Quatieri (1986)] MCAULAY, R., and QUATIERI, T. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34(4):744–754.
- [Meyer and Simmer (1997)] MEYER, J., and SIMMER, K. (1997). Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction. In *Proceedings of ICASSP97*, pages 1167–1170, Munich, Germany. IEEE.
- [Mourjopoulos (1985)] MOURJOPOULOS, J. (1985). On the variation and invertibility of room impulse response functions. *Journal of Sound and Vibration*, 102(2):217–228.
- [Neely and Allen (1979)] NEELY, S. T., and ALLEN, J. B. (1979). Invertibility of a room impulse response. *Journal of the Acoustical Society of America*, 66(1):165–169.
- [Nishiura et al. (2000)] NISHIURA, T., YAMADA, T., NAKAMURA, S., and SHIKANO, K. (2000). Localization of multiple sound sources based on a CSP analysis with a microphone array. In *Proceedings of ICASSP2000*, pages II-1053– II-1056. IEEE.

- [Omologo and Svaizer (1997)] OMOLOGO, M., and SVAIZER, P. (1997). Use of the cross-power spectrum phase in acoustic event localization. *IEEE Trans. Speech Audio Proc.*, 5(3):288–292.
- [Oppenheim and Shafer (1989)] OPPENHEIM, A. V., and SHAFER, R. W. (1989). *Discrete-Time Signal Processing*. Prentice Hall.
- [Peterson (1986)] PETERSON, P. (1986). Simulating the response of multiple microphones to a single acoustic source in a reverberant room. *J. Acoust. Soc. Am.*, 80(5):1527–1529.
- [Piersol (1981)] PIERSOL, A. (1981). Time delay estimation using phase data. *IEEE Trans. Acoust., Speech, Signal Processing*, 29(3, pt. 2):471–477.
- [Pintér (1996)] PINTÉR, I. (1996). Perceptual wavelet-representation of speech signals and its application to speech enhancement. *Computer Speech and Language*, 10(1):1–22.
- [Pohlmann (1995)] POHLMANN, K. (1995). *Principles of Digital Audio*. McGraw-Hill, New York, 3 edition.
- [Porat (1994)] PORAT, B. (1994). *Digital Processing of Random Signals: Theory and Methods*. Prentice Hall.
- [Quackenbush et al. (1988)] QUACKENBUSH, S., BARNWELL, T., and CLEMENTS, M. (1988). *Objective Measures of Speech Quality*. Prentice-Hall.
- [Quatieri (2002)] QUATIERI, T. (2002). *Discrete-Time Speech Signal Processing - Principles and Practice*. Prentice Hall.
- [Rabiner and Shafer (1978)] RABINER, L. R., and SHAFER, R. W. (1978). *Digital Processing of Speech Signals*. Prentice Hall.
- [Radlovic et al. (1999)] RADLOVIC, B., WILLIAMSON, R., and KENNEDY, R. (1999). On the poor robustness of sound equalization in reverberant environments. In *ICASSP99*, pages 881–884, Phoenix, AZ. IEEE.
- [Savoji (1989)] SAVOJI, M. (1989). A robust algorithm for accurate endpointing of speech signals. *Speech Communications*, 8:45–60.

- [Seok and Bae (1997)] SEOK, J. W., and BAE, K. S. (1997). Speech enhancement with reduction of noise components in the wavelet domain. In *Proceedings of ICASSP97*, pages 1323–1326, Munich, Germany. IEEE.
- [Simmer et al. (2001)] SIMMER, K., BITZER, J., and MARRO, C. (2001). Post-filtering techniques. In M. Brandstein and D. Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Verlag.
- [Simmer and Wasiljeff (1992)] SIMMER, K., and WASILJEFF, A. (1992). Adaptive microphone arrays for noise suppression in the frequency domain. In *Second Cost 229 Workshop on Adaptive Algorithms in Communications*, pages 185–194, Bordeaux, France.
- [Singhal and Atal (1984)] SINGHAL, S., and ATAL, B. S. (1984). Improving performance of multi-pulse LPC coders at low bit rates. In *Proceedings of ICASSP84*, pages I–131–I–134. IEEE.
- [Sohn and Sung (1999)] SOHN, J., and SUNG, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, pages 1–3.
- [Srinivasan and Jamieson (1996)] SRINIVASAN, P., and JAMIESON, L. H. (1996). Wavelets based on splines: an application. *SPIE*, 2825(4):456–464.
- [Strang (1994)] STRANG, G. (1994). Wavelets. *American Scientist*, 82:250–255.
- [Strube (1974)] STRUBE, H. W. (1974). Determination of the instant of glottal closure from the speech wave. *Journal of the Acoustical Society of America*, 56:1625–1629.
- [Subramaniam et al. (1996)] SUBRAMANIAM, S., PETROPULU, A. P., and WENDT, C. (1996). Cepstrum-based deconvolution for speech dereverberation. *IEEE Transactions on Speech and Audio Processing*, 4(5):392–396.
- [Svaizer et al. (1997)] SVAIZER, P., MATASSONI, M., and OMOLOGO, M. (1997). Acoustic source location in a three-dimensional space using cross-power spectrum phase. In *Proceedings of ICASSP97*, pages 231–234, Munich, Germany. IEEE.

- [Tohyama et al. (1993)] TOHYAMA, M., LYON, R. H., and KOIKE, T. (1993). Source waveform recovery in a reverberant space by cepstrum dereverberation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 157–159.
- [van Laarhoven and Aarts (1989)] VAN LAARHOVEN, P., and AARTS, E. (1989). *Simulated Annealing: Theory and Applications*. Kluwer Academic Publishers.
- [Wang et al. (2001)] WANG, C., GRIEBEL, S., BRANDSTEIN, M., and HSU, B. (2001). Real-time automated video and audio capture with multiple cameras and microphones. *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, 29(1/2):81–99.
- [Wang and Chu (1997)] WANG, H., and CHU, P. (1997). Voice source localization for automatic camera pointing system in videoconferencing. In *Proceedings of ICASSP97*, pages 187–190, Munich, Germany. IEEE.
- [Wang et al. (1992)] WANG, S., SEKEY, A., and GERSHO, A. (1992). An objective measure for predicting subjective quality of speech coders. *IEEE Journal Selected Areas in Communications*, 10(5):819–829.
- [Yang et al. (1993)] YANG, Z., SIMMER, K., and WASILJEFF, A. (1993). Improved performance of multi-microphone speech enhancement systems. In *Proceedings of the 14th GRETSI Symposium*, pages 479–482.
- [Yoo and Lim (1995)] YOO, C., and LIM, J. (1995). Speech enhancement based on the generalized dual excitation model with adaptive analysis window. In *Proceedings of ICASSP95*, pages 832–835, Detroit, Michigan. IEEE.
- [Youn et al. (1986)] YOUN, D., CHIOU, S., and MATHEWS, V. (1986). Adaptive phase transform processors for time delay estimation. *J. Acoust. Soc. Am.*, 80(1):188–194.
- [Zelinski (1988)] ZELINSKI, R. (1988). A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *Proc. IEEE ICASSP*, pages 2578–2580, Seattle, WA. IEEE.
- [Zwicker and Fastl (1990)] ZWICKER, E., and FASTL, H. (1990). *Psychoacoustics Facts and Models*. Springer-Verlag, Berlin, 2nd edition.