

Multi-Channel Wavelet Techniques for Reverberant Speech Analysis and Enhancement

Scott Matthew Griebel

Harvard University

Division of Engineering and Applied Sciences

Technical Report

February 15, 1999

Contents

1	Introduction	1
1.1	Speech Production Model	2
1.2	Reverberant Speech Processing	4
2	The Wavelet Transform	7
2.1	Introduction	7
2.2	The Short-Time Fourier Transform	7
2.3	The Wavelet Transform	8
2.3.1	Multiresolution Analysis	9
2.3.2	Filter Bank Implementation	10
2.3.3	Biorthogonal Wavelet Systems	12
2.3.4	Quadratic Spline Wavelets	13
3	Pitch Estimation	25
3.1	Autocorrelation Method	25
3.2	Cepstrum Method	28
3.3	Maximum Likelihood Pitch Estimation	30
3.4	Wavelet Pitch Estimation	34
3.5	Comparison of Pitch Estimators for Reverberant Speech	35
4	Delay Estimation	40
4.1	Time Delay Estimation	40
4.2	Generalized Cross Correlation Time Delay Estimation	41
4.3	Wavelet-based Time Delay Estimation	42

5	Speech Dereverberation	47
5.1	Motivation and Previous Work	47
5.2	Wavelet-based Speech Enhancement	48
6	Conclusions	53
6.1	Summary of Current Progress	53

List of Figures

1.1	Main components of speech production [15]	2
1.2	Block diagram of speech production system [15]	3
1.3	Glottal waveform example [36]	4
1.4	Main components of speech production [36]	5
1.5	Room setup for reverberant speech simulation [15]	6
2.1	Two stages of the wavelet transform analysis filter bank. The symbol $\downarrow$ refers to subsampling, or decimation, by a factor of 2.	11
2.2	Two stages of the wavelet transform synthesis filter bank. The symbol $\uparrow$ refers to interpolation by a factor of 2 by zero-insertion.	12
2.3	Filter bank structure for biorthogonal wavelet systems	13
2.4	Plot of the wavelet and scaling functions	15
2.5	Wavelet transform example: top plot is original signal. Remaining plots are wavelet coefficients at scales $2^{-1} \dots 2^{-4}$	16
2.6	Wavelet transform speech example	18
2.7	Wavelet transform of speech plus additive white noise	19
2.8	Clean (red) and reverberant (blue) speech signals and corresponding wavelet transforms	20
2.9	Alternating projection algorithm	21
2.10	Wavelet extrema reconstruction example	22
2.11	SNR of first six scales and reconstructed signal	23
2.12	Wavelet extrema reconstruction speech example	23
2.13	SNR of first six scales and reconstructed speech signal	24
3.1	Voiced /i/, unvoiced /s/, and corresponding autocorrelation functions	26
3.2	Voiced /i/, unvoiced /s/, and corresponding unbiased autocorrelation functions	27
3.3	Voiced /i/ after 30% center clipping and corresponding unbiased autocorrelation	28

3.4	Magnitude responses of speech production components	29
3.5	Separation of pitch and vocal tract in the cepstral domain	31
3.6	Maximum likelihood energy function	34
3.7	Wavelet peak tracing from coarse to fine scales	36
3.8	Wavelet-based clipped signal	37
3.9	Accuracy percentages for pitch estimation algorithms	39
3.10	Mean-squared errors for pitch estimation algorithms	39
4.1	Two-channel time delay estimation problem	41
4.2	Comparison of time delay estimators for a reverberation time of 200 msec	44
4.3	Wavelet coherence function for channels 1, 2, and 3	46
4.4	Comparison of the wavelet coherence method using $M = 2$ and $M = 3$	46
5.1	Plot of 8 reverberant channels and clustered peaks	49
5.2	Clustered reverberant peaks and scale 2^{-5} of nonreverberant speech	50
5.3	Short-term coherence function for scale 2^{-3}	51
5.4	Weighted peaks and scale 2^{-5} of nonreverberant speech	51
5.5	Single channel of clean speech, single channel of reverberant speech, beamformed result, and the wavelet extrema reconstruction result	52

List of Tables

2.1	Filter coefficients of analysis and synthesis filters	17
-----	---	----

Abstract

In many real-world situations, speech signals are degraded by the reverberant effects of the environment in which they are produced. Traditional speech processing problems are typically much more difficult to solve in a reverberant enclosure. In this paper, multi-channel wavelet techniques are developed for pitch estimation, time delay estimation, and the more general problem of reverberant speech enhancement. Wavelet extrema can be used to detect certain events in the speech production process - events which, over multiple channels, tend to be more coherent than the extrema due to reverberations. Therefore, the central theme of this work is to make use of the wavelet extrema for each of the tasks listed above.

Chapter 1

Introduction

In many real-world situations, speech signals are degraded by the reverberant effects of the environment in which they are produced. Traditional speech processing problems are typically much more difficult to solve in a reverberant setting. This paper addresses a variety of traditional speech processing problems where robust solutions for reverberant environments have yet to be found. In general, the approaches taken in this research involve a wavelet analysis of speech data acquired at multiple, spatially diverse microphones. An underlying theme of this work is the exploitation of wavelet extrema as a means of event detection in the reverberated speech signals. 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 in a slightly different way, as evidenced by variations in wavelet extrema from channel to channel. By deciding which signal portions are coherent over all microphones, we can attempt to capture the structure of the original speech.

The first speech processing problem to be addressed in this proposal is pitch estimation of reverberant speech. Accurate estimation of the pitch period (or frequency) of a speech signal is important for many applications such as speech coding, speech recognition, and speaker identification. Although there are many pitch estimation algorithms in existence, including a single-channel wavelet-based approach, their performance for reverberant speech has not been studied. The second problem to be studied is time delay estimation. For multi-channel signal acquisition in a nonreverberant environment, a transmitted speech waveform will arrive at each microphone at a slightly different time instant. This time difference is known as the relative time delay between the microphones. By calculating the relative time delay between two microphones, a direction of arrival known as the bearing angle can be estimated. By calculating the bearing angles for a set of microphones, the location of the sound source can be estimated. However, in reverberant environments, the additional echoes in the observed speech signals significantly reduce the

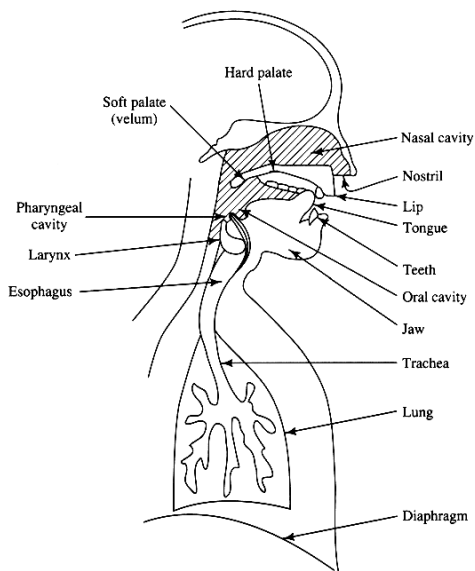


Figure 1.1: Main components of speech production [15]

performance of standard approaches. The final problem to be discussed is the enhancement of reverberant speech. Simple delay-and-sum beamforming approaches which have proven useful for additive noise applications do not perform adequately for reverberant noise, which is convolutional in nature. This paper presents a wavelet-based algorithm for each of these problems and compares their performance to traditional approaches. The remainder of this chapter will present the basic speech model that is typically used in speech processing applications as well as a brief description of the reverberant speech problem.

1.1 Speech Production Model

The human speech production system, shown in Figure 1.1, consists mainly of the vocal cords, the vocal tract, the nasal cavity, and the lips [15, 36]. 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. Speech sounds are produced by exciting this system by a voiced excitation, an unvoiced excitation, or a combination of both. A block diagram of the speech production system is given in Figure 1.2. 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. 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

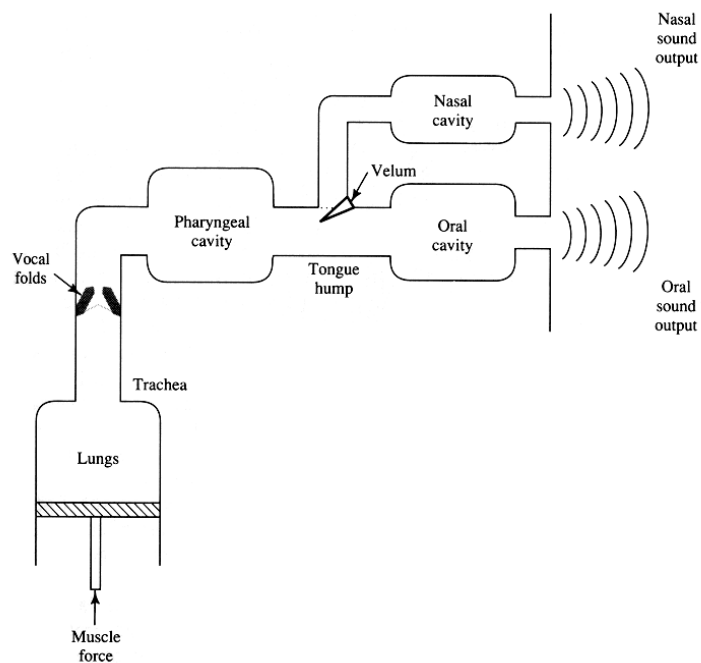


Figure 1.2: Block diagram of speech production system [15]

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). For 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 of the speech signal. The fundamental frequency is then defined as the reciprocal of the fundamental period. The discrete-time waveform of the glottal source is often approximated by Rosenberg's model [36]:

$$g[n] = \begin{cases} \frac{1}{2} \left[1 - \cos \left(\frac{\pi n}{P} \right) \right], & 0 \leq n \leq P \\ \frac{1}{2} \cos \left[\frac{\pi(n-P)}{2(K-P)} \right], & P \leq n \leq K \\ 0, & \text{otherwise} \end{cases} \quad (1.1)$$

where P and K specify the times at which the glottis is wide open and completely closed respectively. A glottal waveform example is plotted in Figure 1.3. The glottal input to the vocal tract is the convolution of a 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 is

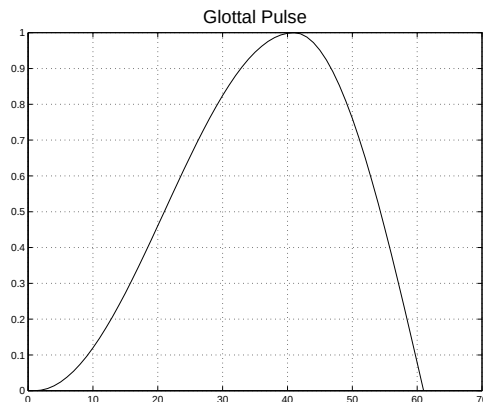


Figure 1.3: Glottal waveform example [36]

often modeled as an all-pole filter of the form:

$$H(z) = \frac{G}{1 - \sum_{k=1}^N a_k z^{-k}} \quad (1.2)$$

where $H(z)$ is the standard Z-transform of a discrete-time signal. The poles, or resonances, of the vocal tract are known as the *formant frequencies* due to the fact that they form the shape of the speech spectrum. Although the all-pole model is not accurate for all speech sounds, it is the model most often used in speech processing algorithms. Ignoring the production of nasal sounds, the final element of the speech model is the lip radiation transfer function. This function transfers the air volume velocity of the vocal tract to a sound pressure waveform - what is typically measured by a microphone. A typical choice for the lip radiation function is [15]:

$$R(z) = 1 - z_0 z^{-1}, \quad z_0 \approx 1, \quad z_0 < 1 \quad (1.3)$$

The discrete-time model of speech production based on the above discussion is shown in Figure 1.4 [36].

1.2 Reverberant Speech Processing

In a non-reverberant two-channel setting, a signal emanating from a distant source will typically arrive at each microphone at a slightly different time. This time delay is dependent on the source and microphone locations as well as the speed of sound ($c = 342$ m/s under standard conditions). Mathematically, this problem can be modeled as:

$$\begin{aligned} x_1(t) &= s(t) + n_1(t) \\ x_2(t) &= s(t - D) + n_2(t) \end{aligned} \quad (1.4)$$

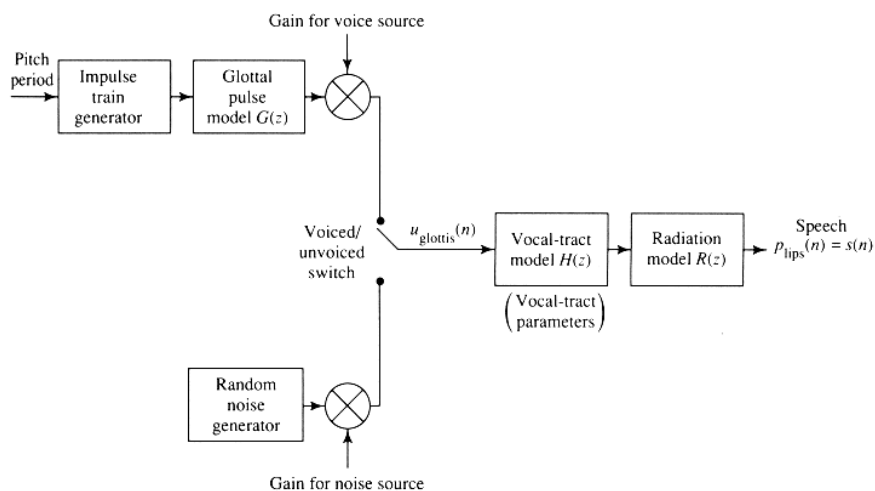


Figure 1.4: Main components of speech production [36]

where $s(t)$ is the source signal, $n_1(t)$ and $n_2(t)$ are additive noise signals, and D is the time delay. The noise signals are assumed to be uncorrelated with each other and with $s(t)$. In order to model multi-channel signal acquisition in a reverberant setting, Equation 1.4 is generalized to:

$$\begin{aligned} x_1(t) &= h_1(t) * s(t) + n_1(t) \\ x_2(t) &= h_2(t) * s(t) + n_2(t) \end{aligned} \tag{1.5}$$

where $h_1(t)$ and $h_2(t)$ are the channel responses between the source and the two microphones. Note that in the case of $h_1(t) = \delta(t)$ and $h_2(t) = \delta(t - D)$, Equation 1.5 reduces to the nonreverberant case of Equation 1.4.

In this paper, the reverberant channel responses were simulated using the image model method developed in [2]. The room size and setup used for these simulations are shown in Figure 1.5. The size of the room was 4x4x3 meters. The speech source was placed nearly 2.5 meters from the microphones which were located 0.5 meters from the opposite wall. When calculating the channel response for each microphone, the reflection coefficient β of the walls, floor, and ceiling must be specified, as well as the maximum reflection order to be considered. In this research, varying levels of reverberation (i.e. values of β) were used, calculating the channel responses up to 6th 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 to decay -40 dB from its initial energy level - a function of the room size. The

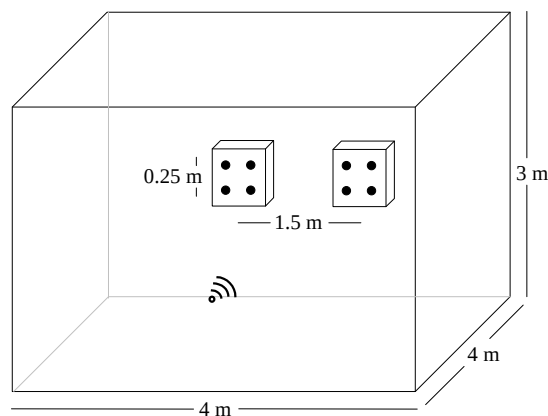


Figure 1.5: Room setup for reverberant speech simulation [15]

relationship between reverberation time, T , and the reflection coefficient is given by Eyring's formula:

$$\beta = \exp \left\{ -13.82 / [cT(L_x^{-1} + L_y^{-1} + L_z^{-1})] \right\} \quad (1.6)$$

where L_x , L_y , and L_z are the room dimensions. In this paper, reverberations times of 0, 100 msec, 200 msec, $\dots$ 500 msec are considered. These values correspond to β values of 0, 0.62, 0.78, 0.85, 0.89, and 0.91. To generate the different reverberant speech signals, a single-channel speech signal is filtered with each of the calculated channel responses.

This chapter provided a brief introduction to the basic speech model and the general reverberant speech problem. The remainder of this proposal is structured as follows: Chapter 2 presents a review of time-frequency and time-scale analysis tools, with an emphasis on the wavelet transform. Chapters 3 and 4 cover pitch estimation and time delay estimation of reverberant speech. Traditional approaches are presented, along with new wavelet-based techniques. Chapter 5 presents some preliminary results on reverberant speech enhancement. Finally, Chapter 6 makes some general conclusions and proposes a direction for future work.

Chapter 2

The Wavelet Transform

2.1 Introduction

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

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t} dt \quad (2.1)$$

where $i = \sqrt{-1}$. 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 dominant 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 [42]. The Fourier transform is equivalent to cataloging the numerous 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, or frequencies, progress with time.

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 [12]:

$$\text{STFT}(t, \omega) = \int_{-\infty}^{\infty} x(s)h^*(s-t)e^{-i\omega s} ds \quad (2.2)$$

where $*$ 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 we define the effective duration, τ_{eff} , and effective bandwidth, ω_{eff} , as [34]:

$$\tau_{eff} = \left[\frac{\int_{-\infty}^{\infty} s^2 |h(s)|^2 ds}{\int_{-\infty}^{\infty} |h(s)|^2 ds} \right]^{\frac{1}{2}}, \quad \omega_{eff} = \left[\frac{\int_{-\infty}^{\infty} \omega^2 |h(\omega)|^2 d\omega}{\int_{-\infty}^{\infty} |h(\omega)|^2 d\omega} \right]^{\frac{1}{2}} \quad (2.3)$$

then the uncertainty principle is given by:

$$\tau_{eff} \cdot \omega_{eff} \geq 0.5. \quad (2.4)$$

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, upon choosing the window function, the time-frequency resolution is fixed over the entire time-frequency plane.

2.3 The Wavelet Transform

The wavelet transform 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 signals having shorter transient phenomena as well as longer slowly varying phenomena, and is also similar to how the human ear analyzes sounds [33]. The continuous wavelet transform of a signal $f(t)$ at scale s is given by [25]:

$$Wf(u, s) = \int_{-\infty}^{\infty} f(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s} \right) dt = f \star \bar{\psi}_s(u) \quad (2.5)$$

where $\psi(t)$ is the “mother wavelet”, $\star$ is the convolution operator, and:

$$\bar{\psi}_s(t) = \frac{1}{\sqrt{s}} \psi^* \left(\frac{-t}{s} \right). \quad (2.6)$$

Note that the $\frac{1}{\sqrt{s}}$ term is for norm preservation of $\psi(t)$ at different scales. There are certain continuous wavelet admissibility conditions that any $\psi(t)$ should satisfy, but the current discussion does not benefit from their inclusion.

2.3.1 Multiresolution Analysis

Unless otherwise stated, the derivations and proofs contained in this section are based on the presentation in [9]. The multiresolution structure of the wavelet transform is typically built upon a scaling function, $\phi(t) \in L^2(\mathbf{R})$. The set of all integer translations of this scaling function is denoted by:

$$\phi_k(t) = \phi(t - k) \quad k \text{ integer.} \quad (2.7)$$

The space $\mathcal{V}_0$, a subset of $L^2(\mathbf{R})$, is defined as the span (technically the closure) of the translated scaling functions. The span is the set of all functions which can be expressed as a linear combination of integer translations of the scaling functions. That is, for any $f(t) \in \mathcal{V}_0$:

$$f(t) = \sum_k a_k \phi_k(t). \quad (2.8)$$

By also including time-scaled versions of the basic scaling function, a more expressive set of subspaces can be constructed. That is, by defining:

$$\phi_{j,k}(t) = 2^{j/2} \phi(2^j t - k) \quad j, k \text{ integer} \quad (2.9)$$

a new space is spanned, $\mathcal{V}_j$. Note that the $2^{j/2}$ term is for norm preservation only. For positive values of j , the basic scaling function is compressed, while for negative values of j , it is dilated, or expanded.

The multiresolution formulation requires $\mathcal{V}_j \subset \mathcal{V}_{j+1}$ for all integer j , with the limiting properties that $\mathcal{V}_{-\infty} = \{0\}$ and $\mathcal{V}_{\infty} = L^2(\mathbf{R})$. Therefore, we want $f(t) \in \mathcal{V}_j$ if and only if $f(2t) \in \mathcal{V}_{j+1}$. This nesting can be achieved through the dilation equation:

$$\phi(t) = \sum_n h[n] \sqrt{2} \phi(2t - n) \quad n \text{ integer.} \quad (2.10)$$

The existence and uniqueness of the scaling function coefficients $h[n]$ is not guaranteed.

The spaces $\mathcal{W}_j$ spanned by the wavelet functions are defined as the orthogonal complements of $\mathcal{V}_j$ so that $\mathcal{V}_{j+1} = \mathcal{V}_j \oplus \mathcal{W}_j$. Because $\mathcal{W}_j \subset \mathcal{V}_{j+1}$, the wavelet function satisfies the following two-scale difference equation:

$$\psi(t) = \sum_n h_1[n] \sqrt{2} \phi(2t - n) \quad n \text{ integer.} \quad (2.11)$$

This two scale difference equation generates a mother wavelet, $\psi(t)$, whose scalings and translations are given by:

$$\psi_{j,k}(t) = 2^{j/2} \psi(2^j t - k) \quad j, k \text{ integer.} \quad (2.12)$$

It can be shown that an orthogonal wavelet system requires the wavelet function coefficients to be related to the scaling function coefficients according to $h_1[n] = (-1)^n h[1-n]$. By including an infinite number of scales j in Equation 2.12, any function $f(t) \in L^2(\mathbf{R})$ can be represented by:

$$f(t) = \sum_{k=-\infty}^{\infty} c[k] \phi_k(t) + \sum_{j=0}^{\infty} \sum_{k=-\infty}^{\infty} d_j[k] \psi_{j,k}(t). \quad (2.13)$$

From this equation, we see that in this multiresolution framework, a function is represented by a low resolution approximation $c[k]$ and a series of increasingly higher resolution signals $d_j[k]$. These higher resolution, or detail, signals contain the information that is needed to reconstruct the original function from the coarse approximation. The meaning of these statements should become clearer as the relevant mathematics is presented. If the scaling and wavelet functions form an orthonormal basis or a tight frame, the coefficients can be found using the inner product:

$$c[k] = \langle f(t), \phi_k(t) \rangle = \int f(t) \phi_k(t) dt \quad (2.14)$$

and

$$d_j[k] = \langle f(t), \psi_{j,k}(t) \rangle = \int f(t) \psi_{j,k}(t) dt. \quad (2.15)$$

Another advantage of using scaling functions and wavelet functions which form an orthonormal basis or tight frame is Parseval's relation. Parseval's theorem refers to the ability of the wavelet transform to preserve the energy of the original signal:

$$\int |f(t)|^2 dt = \sum_{k=-\infty}^{\infty} |c[k]|^2 + \sum_{j=0}^{\infty} \sum_{k=-\infty}^{\infty} |d_j[k]|^2. \quad (2.16)$$

2.3.2 Filter Bank Implementation

In practice, the wavelet transform is implemented using filter banks. In this section, the multiresolution interpretation of the previous section is utilized to derive an appropriate filter bank structure [9]. Starting with Equation 2.10 and substituting $2^j t - k$ for t results in:

$$\phi(2^j t - k) = \sum_n h[n] \sqrt{2} \phi(2(2^j t - k) - n) = \sum_n h[n] \sqrt{2} \phi(2^{j+1} t - 2k - n) \quad (2.17)$$

which, through a change of variables, can be rewritten as:

$$\phi(2^j t - k) = \sum_m h[m - 2k] \sqrt{2} \phi(2^{j+1} t - m). \quad (2.18)$$

Denoting $\mathcal{V}_j = \text{span}_k \{2^{j/2} \phi(2^{j/2} t - k)\}$, any function $f(t) \in \mathcal{V}_{j+1}$ can be expressed as:

$$f(t) = \sum_k c_{j+1}[k] 2^{(j+1)/2} \phi(2^{j+1} t - k). \quad (2.19)$$

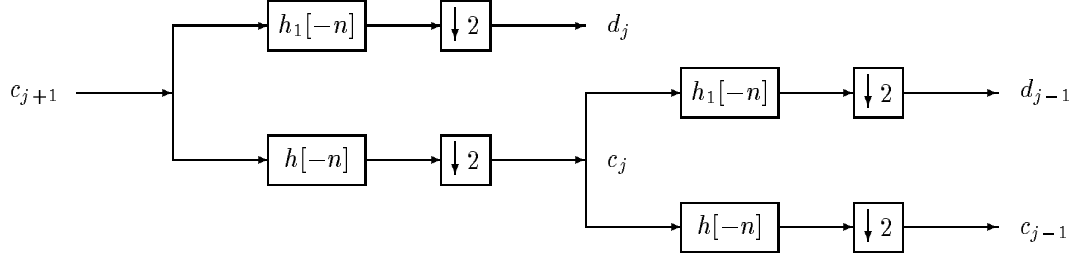


Figure 2.1: Two stages of the wavelet transform analysis filter bank. The symbol $\downarrow$ refers to subsampling, or decimation, by a factor of 2.

In order to express $f(t)$ at scale j , both the scaling function and the wavelet function are needed:

$$f(t) = \sum_k c_j[k] 2^{j/2} \phi(2^j t - k) + \sum_k d_j[k] 2^{j/2} \psi(2^j t - k) \quad (2.20)$$

The coefficients $c_j[k]$ are given by:

$$\begin{aligned} c_j[k] &= \langle f(t), \phi_{j,k}(t) \rangle = \int f(t) 2^{j/2} \phi(2^j t - k) dt \\ &= \sum_m h(m - 2k) \int f(t) 2^{(j+1)/2} \phi(2^{j+1} t - m) dt \\ &= \sum_m h(m - 2k) c_{j+1}[m] \end{aligned} \quad (2.21)$$

where we have used Equation 2.18 to go from the first line to the second. A similar result can be proven for the wavelet coefficients:

$$d_j[k] = \sum_m h_1[m - 2k] c_{j+1}[m]. \quad (2.22)$$

Therefore, the approximation and detail coefficients can be calculated for the next coarser resolution through iterated discrete convolutions. Two stages of the analysis filter bank corresponding to Equations 2.21 and 2.22 are shown in Figure 2.1.

A dual argument can be used to reconstruct a finer resolution's approximation from a coarser resolution approximation and the appropriate detail signal. Again, we make use of the fact that a function $f(t) \in \mathcal{V}_{j+1}$ can be expressed in terms of the scaling and wavelet functions of scale j according to 2.20. Making use of Equation 2.17 and its wavelet dual results in:

$$f(t) = \sum_k c_j[k] \sum_n h[n] 2^{(j+1)/2} \phi(2^{j+1} t - 2k - n) + \sum_k d_j[k] \sum_n h_1[n] 2^{(j+1)/2} \psi(2^{j+1} t - 2k - n). \quad (2.23)$$

Equating the righthand sides of Equations 2.19 and 2.23, multiplying by $\phi(2^{j+1} t - k)$, and integrating (again assuming the orthonormality of the scaling functions) gives:

$$c_{j+1}[k] = \sum_m c_j[m] h[k - 2m] + \sum_m d_j[m] h_1[k - 2m]. \quad (2.24)$$

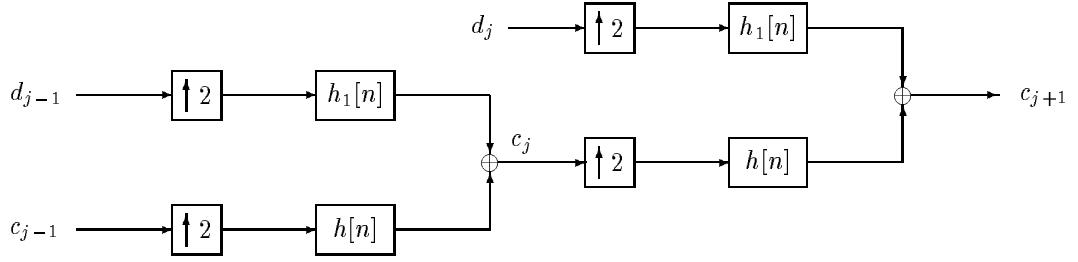


Figure 2.2: Two stages of the wavelet transform synthesis filter bank. The symbol $\uparrow$ refers to interpolation by a factor of 2 by zero-insertion.

Two stages of the synthesis filter bank corresponding to Equation 2.24 are shown in Figure 2.2. In order to use filter banks in a practical setting, the initial input $c_J[k]$ must be chosen, where J is the maximum obtainable resolution. We arbitrarily assign $J = 0$ to the sampled signal, with wavelet scales occurring at $2^{-1}, 2^{-2}, \dots$. It has been shown that for the wavelets used in this paper, any discrete sequence can be considered as the projection of a non-unique continuous-time function $f(t)$ onto the space V_0 [26]. Therefore, starting with a discrete signal, the filter banks can be used to compute exactly the wavelet coefficients for some continuous-time signal. Perhaps a more intuitive way to view this is that for very high resolutions (large j), the scaling function is very compressed and $\phi(2^j t - k)$ approaches $\delta(t - k)$ where k is an integer. Therefore, the highest scale wavelet coefficients of a continuous signal can be approximated by its discrete time samples. More formal arguments are given in [1, 37, 38, 39].

2.3.3 Biorthogonal Wavelet Systems

Although the wavelet structure of the above discussion has many nice properties, the orthogonality requirement limits the choice of $h[n]$ and $h_1[n]$. For example, linear phase systems cannot be obtained using orthogonal wavelets. Biorthogonal wavelet systems have more flexibility, and involve a separate set of analysis filters ($g[n]$ and $h[n]$) and synthesis filters ($\tilde{g}[n]$ and $\tilde{h}[n]$). Figure 2.3 shows one analysis and one synthesis stage of a biorthogonal wavelet filter bank. As in Equations 2.21 and 2.22, the coefficients at each stage of the analysis filter bank are found recursively from:

$$c_j[k] = \sum_n h[2k - n]c_{j+1}[n] \quad \text{and} \quad d_j[k] = \sum_n g[2k - n]c_{j+1}[n] \quad (2.25)$$

while the synthesis filter bank coefficients are found from:

$$\tilde{c}_{j+1}[m] = \sum_k \left[\tilde{h}[2k - m]c_j[k] + \tilde{g}[2k - m]d_j[k] \right]. \quad (2.26)$$

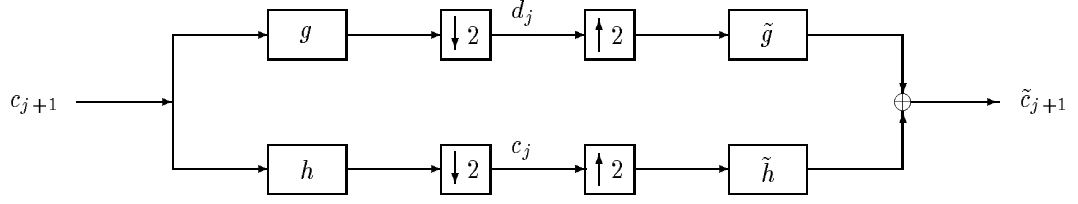


Figure 2.3: Filter bank structure for biorthogonal wavelet systems

Substituting Equation 2.25 into Equation 2.26 gives:

$$\tilde{c}_{j+1}[m] = \sum_n \sum_k \left[\tilde{h}[2k - m]h[2k - n] + \tilde{g}[2k - m]g[2k - n] \right] c_{j+1}[n]. \quad (2.27)$$

Therefore, the perfect reconstruction condition is:

$$\sum_k \left[\tilde{h}[2k - m]h[2k - n] + \tilde{g}[2k - m]g[2k - n] \right] = \delta[m - n]. \quad (2.28)$$

In the frequency domain, the following condition is required:

$$\tilde{H}(\omega)H^*(\omega) + \tilde{G}(\omega)G^*(\omega) = 2. \quad (2.29)$$

The multiresolution structure of a biorthogonal wavelet system is based on the following two-scale difference equations:

$$\begin{aligned} \phi(t) &= \sum_n h[n]\sqrt{2}\phi(2t - n) \\ \tilde{\phi}(t) &= \sum_n \tilde{h}[n]\sqrt{2}\tilde{\phi}(2t - n) \\ \psi(t) &= \sum_n g[n]\sqrt{2}\phi(2t - n) \\ \tilde{\psi}(t) &= \sum_n \tilde{g}[n]\sqrt{2}\tilde{\phi}(2t - n). \end{aligned} \quad (2.30)$$

In the Fourier domain, these equations can be rewritten as:

$$\begin{aligned} \Phi(\omega) &= \frac{1}{\sqrt{2}}H\left(\frac{\omega}{2}\right)\Phi\left(\frac{\omega}{2}\right) \\ \tilde{\Phi}(\omega) &= \frac{1}{\sqrt{2}}\tilde{H}\left(\frac{\omega}{2}\right)\tilde{\Phi}\left(\frac{\omega}{2}\right) \\ \Psi(\omega) &= \frac{1}{\sqrt{2}}G\left(\frac{\omega}{2}\right)\Phi\left(\frac{\omega}{2}\right) \\ \tilde{\Psi}(\omega) &= \frac{1}{\sqrt{2}}\tilde{G}\left(\frac{\omega}{2}\right)\tilde{\Phi}\left(\frac{\omega}{2}\right). \end{aligned} \quad (2.31)$$

2.3.4 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 paper. Suppose the wavelet function $\psi(t)$ is defined as:

$$\psi(t) = \frac{d\theta(t)}{dt} \quad (2.32)$$

where $\theta(t)$ is a smoothing (i.e. lowpass) function. As stated previously, the continuous wavelet transform of a signal $f(t)$ at scale s is given by [25]:

$$Wf(u, s) = \int_{-\infty}^{\infty} f(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s} \right) dt = f \star \bar{\psi}_s(u) \quad (2.33)$$

where:

$$\bar{\psi}_s(t) = \frac{1}{\sqrt{s}} \psi^* \left(\frac{-t}{s} \right). \quad (2.34)$$

Using Equation 2.32, the wavelet coefficients can be written as:

$$Wf(u, s) = f \star \left(s \frac{d\bar{\phi}_s}{du} \right) (u) = s \frac{d}{du} (f \star \bar{\phi}_s) (u). \quad (2.35)$$

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

We now derive a specific set of wavelet and scaling functions [25]. Defining the scaling function to be:

$$\Phi(\omega) = e^{-i\frac{\omega}{2}} \left(\frac{\sin \frac{\omega}{2}}{\frac{\omega}{2}} \right)^3 \quad (2.36)$$

results in the following lowpass filter function:

$$H(\omega) = \sqrt{2} \frac{\Phi(2\omega)}{\Phi(\omega)} = \sqrt{2} e^{-i\frac{\omega}{2}} \left(\cos \frac{\omega}{2} \right)^3 \quad (2.37)$$

where Equation 2.31 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:

$$G(\omega) = -i\sqrt{2} e^{-i\frac{\omega}{2}} \sin \frac{\omega}{2}. \quad (2.38)$$

Using Equation 2.31, the resulting wavelet is given by:

$$\Psi(\omega) = \frac{1}{\sqrt{2}} G\left(\frac{\omega}{2}\right) \Phi\left(\frac{\omega}{2}\right) = -\frac{i\omega}{4} e^{-i\frac{\omega}{2}} \left(\frac{\sin \frac{\omega}{4}}{\frac{\omega}{4}} \right)^4. \quad (2.39)$$

The wavelet and scaling function are shown in Figure 2.4.

By choosing $\tilde{h}(t) = h(t)$, the dual scaling and wavelet functions will also be splines. The perfect reconstruction condition for a biorthogonal filter bank is:

$$\tilde{H}(\omega) H^*(\omega) + \tilde{G}(\omega) G^*(\omega) = 2. \quad (2.40)$$

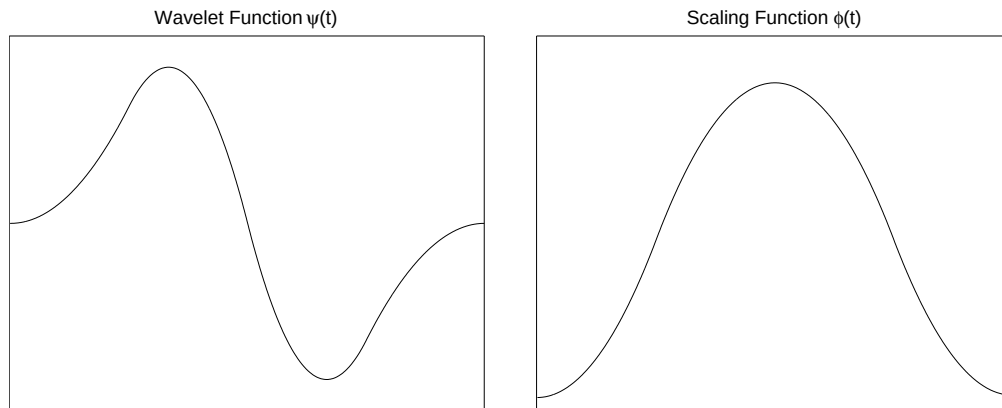


Figure 2.4: Plot of the wavelet and scaling functions

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

$$\tilde{G}(\omega) = \frac{2 - |H(\omega)|^2}{G^*(\omega)} = \frac{2 - 2 \left(\cos \frac{\omega}{2}\right)^6}{i\sqrt{2}e^{i\frac{\omega}{2}} \sin \frac{\omega}{2}}. \quad (2.41)$$

Using the fact that:

$$\sum_{n=0}^2 \left(\cos \frac{\omega}{2}\right)^{2n} = \frac{1 - \left(\cos \frac{\omega}{2}\right)^6}{\left(\sin \frac{\omega}{2}\right)^2} \quad (2.42)$$

leads to:

$$\tilde{G}(\omega) = -i\sqrt{2}e^{-i\frac{\omega}{2}} \sin \frac{\omega}{2} \sum_{n=0}^2 \left(\cos \frac{\omega}{2}\right)^{2n}. \quad (2.43)$$

According to Fourier theory, if a signal $x(t)$ has Fourier transform $F(\omega)$, then $\frac{dx(t)}{dt}$ has Fourier transform $i\omega F(\omega)$. Therefore, the wavelet function given in Equation 2.39 is the derivative of a smoothing function $\theta(t)$, whose Fourier transform is given by:

$$\Theta(\omega) = -\frac{1}{4}e^{-i\frac{\omega}{2}} \left(\frac{\sin \frac{\omega}{4}}{\frac{\omega}{4}}\right)^4. \quad (2.44)$$

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

Figure 2.5 shows an example using this derivative wavelet. Note that discontinuities appear as peaks in each scale. Wavelet local maxima 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 [41].

The Wavelet Transform of Speech Signals

Figure 2.6 shows an example of using the above wavelets for a speech signal. Recall from Chapter 1 that voiced speech is modeled as a vocal tract response excited by a glottal pulse. This pulse essentially

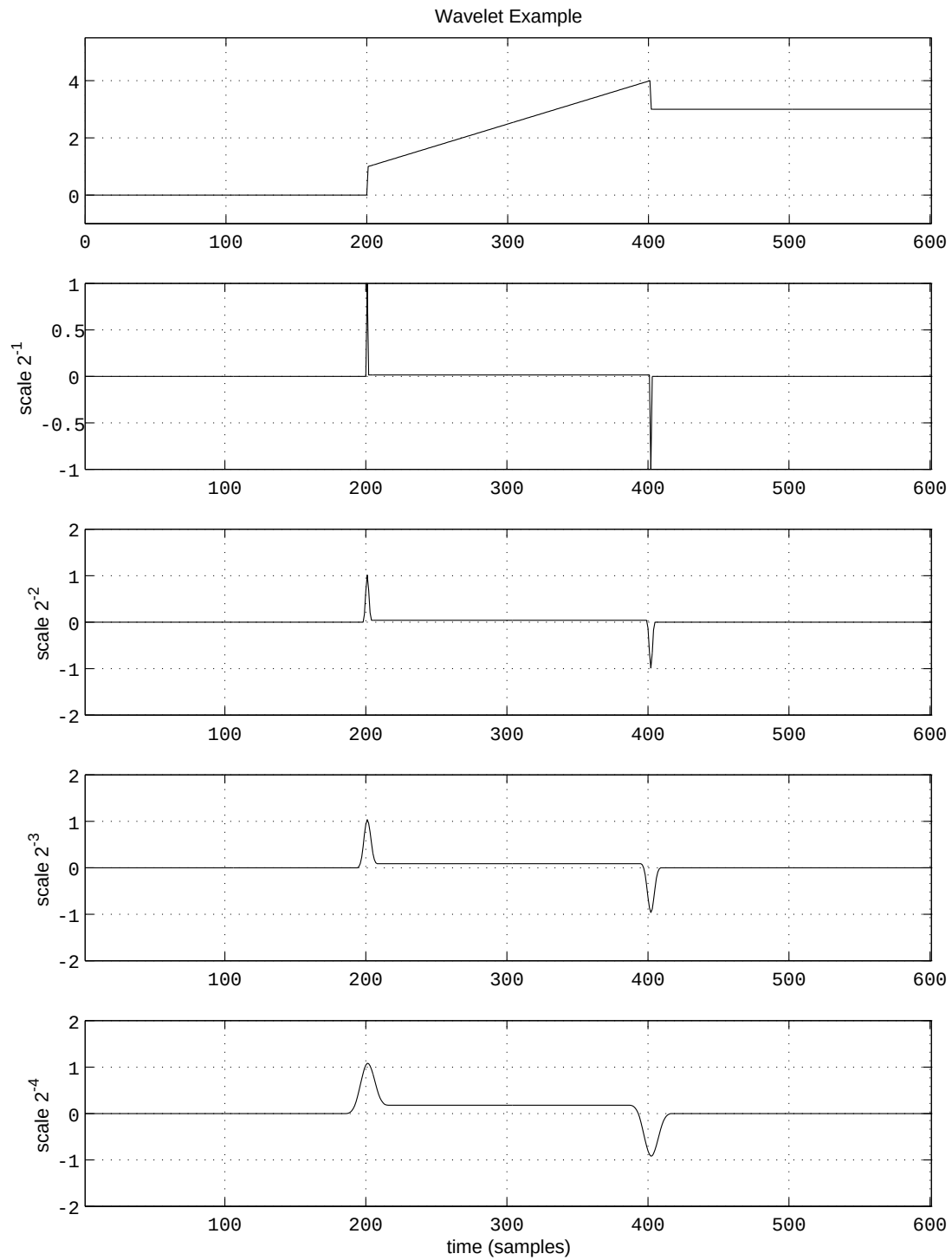


Figure 2.5: Wavelet transform example: top plot is original signal. Remaining plots are wavelet coefficients at scales $2^{-1} \dots 2^{-4}$

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	0.375		-0.21785
0	0.375	0.375	-0.5	-0.6875
1	0.375	0.375	0.5	0.6875
2	0.125	0.125		0.21785
3				0.03125

Table 2.1: Filter coefficients of analysis and synthesis filters

excites all of the formant frequencies simultaneously, resulting in a maximal derivative in the speech signal [19, 20, 21]. This results in wavelet extrema in each scale at the point of the discontinuity, a fact that was verified using simulated speech. Therefore, the wavelet transform can be used as an event detector, with the event being the instant of glottal closure, the estimation of which has been a heavy research area [3, 4, 11, 18, 43]. Figure 2.7 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 wavelet extrema still correspond to the points of glottal closure. Figure 2.8 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 wavelet extrema due to the points of glottal closure are more coherent than the extrema due to reverberations. This observation will be utilized throughout this paper.

Wavelet Extrema

The extrema of the wavelet coefficients at each scale have been used extensively for signal de-noising, detection, and coding [26, 41]. In fact, it has been shown that a signal $f(t)$ can be closely approximated by reconstruction from only the lowest resolution approximation and the largest wavelet extrema of each detail signal. Mallat proposed an algorithm in [26] that numerically reconstructs an estimate, $h(t)$, of the signal, $f(t)$, from the extrema of $Wf(u, 2^j)$ for integer j through the use of alternating projections. Define the space $\mathbf{V}$ as the space of all possible sets of wavelet coefficients generated by the wavelets derived previously. 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. The approximation $h(t)$ is then obtained through a series of alternating projections as shown in Figure 2.9. This algorithm should converge to an element of $\Lambda = \mathbf{V} \cap \Gamma$. 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 [25]. Many papers deal with the completeness and stability of extrema representations as well as more efficient and reliable algorithms than that proposed by Mallat [7, 10, 14].

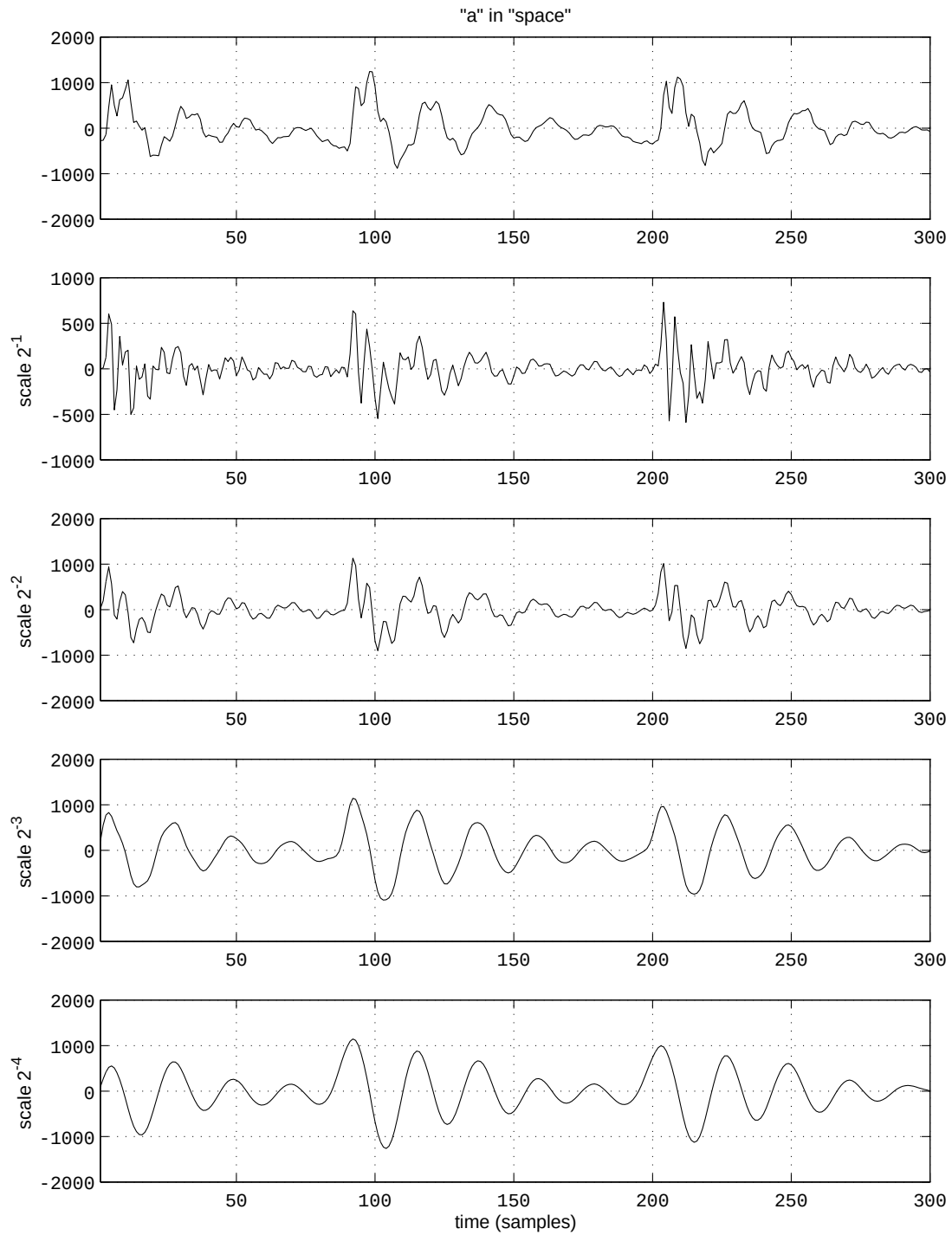


Figure 2.6: Wavelet transform speech example

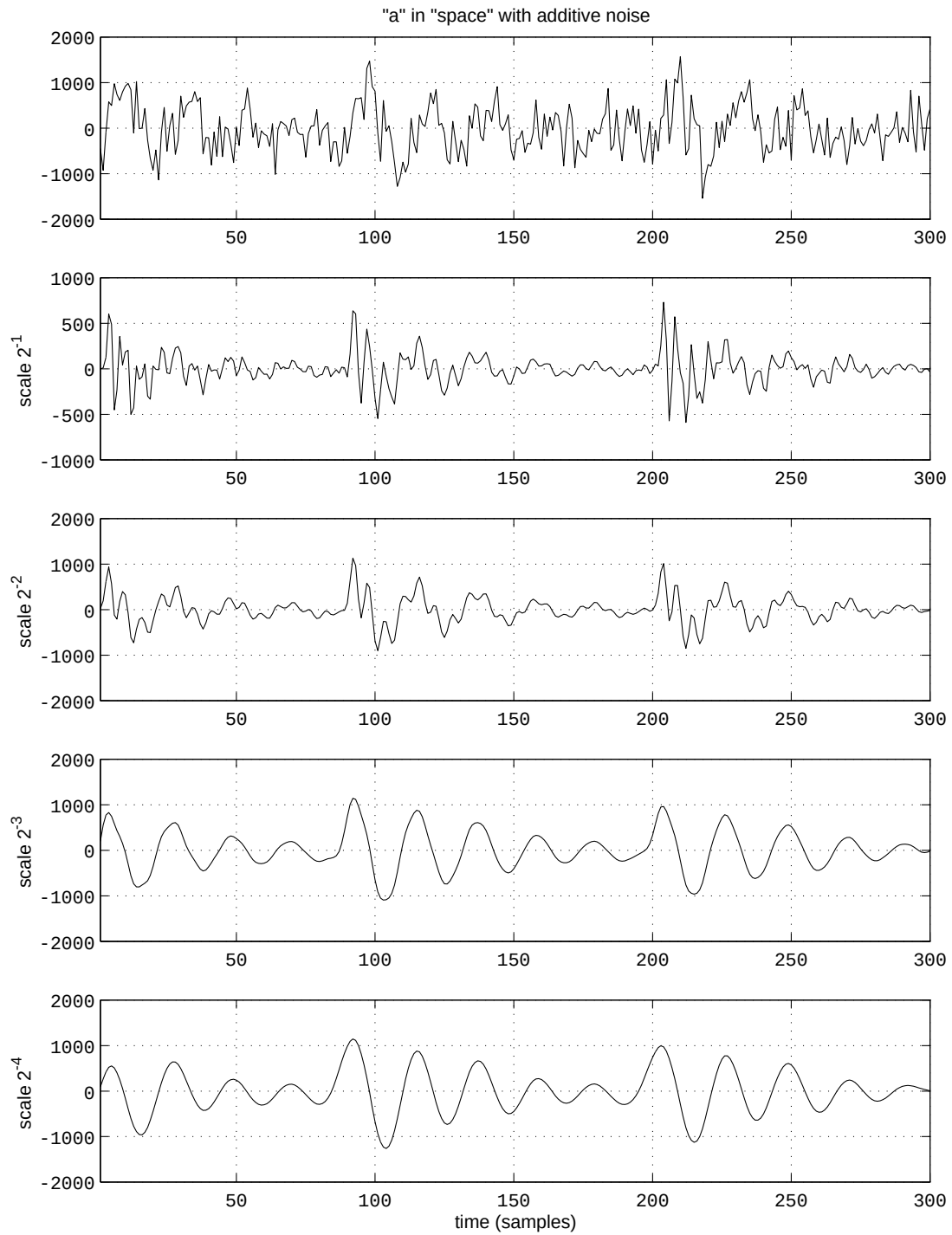


Figure 2.7: Wavelet transform of speech plus additive white noise

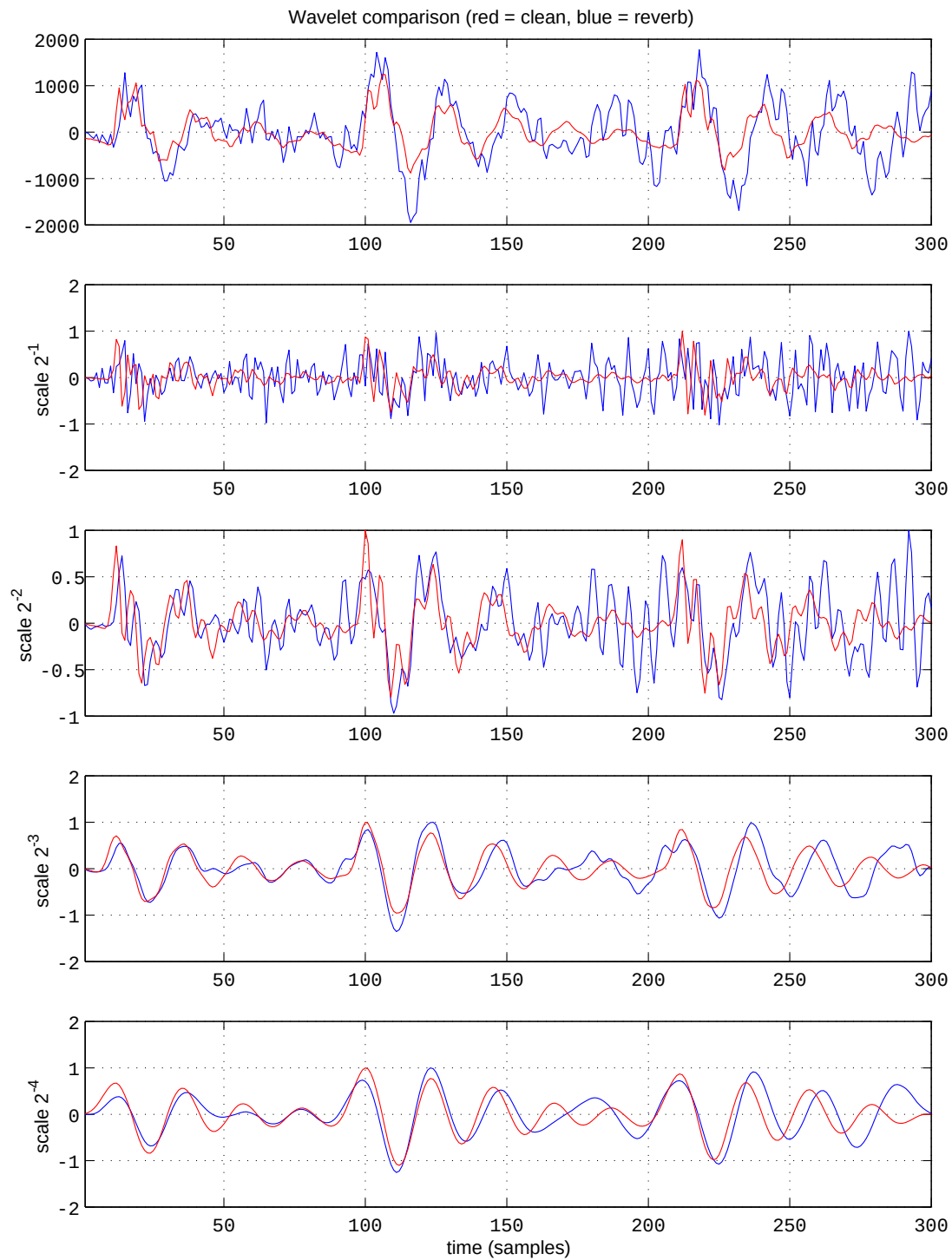


Figure 2.8: Clean (red) and reverberant (blue) speech signals and corresponding wavelet transforms

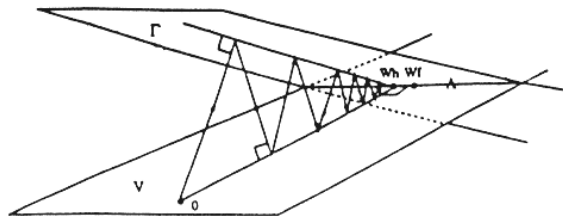


Figure 2.9: Alternating Projection Algorithm [26]

Figure 2.10 shows an example of wavelet 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, while the third plot shows the reconstruction after 25 iterations. Figure 2.11 shows the 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, as expected. Figure 2.12 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 2.13 shows the SNR of the individual wavelet scales as well as the reconstructed signal. This ability to reconstruct speech from its wavelet extrema will be used in Chapter 5 for reverberant speech enhancement.

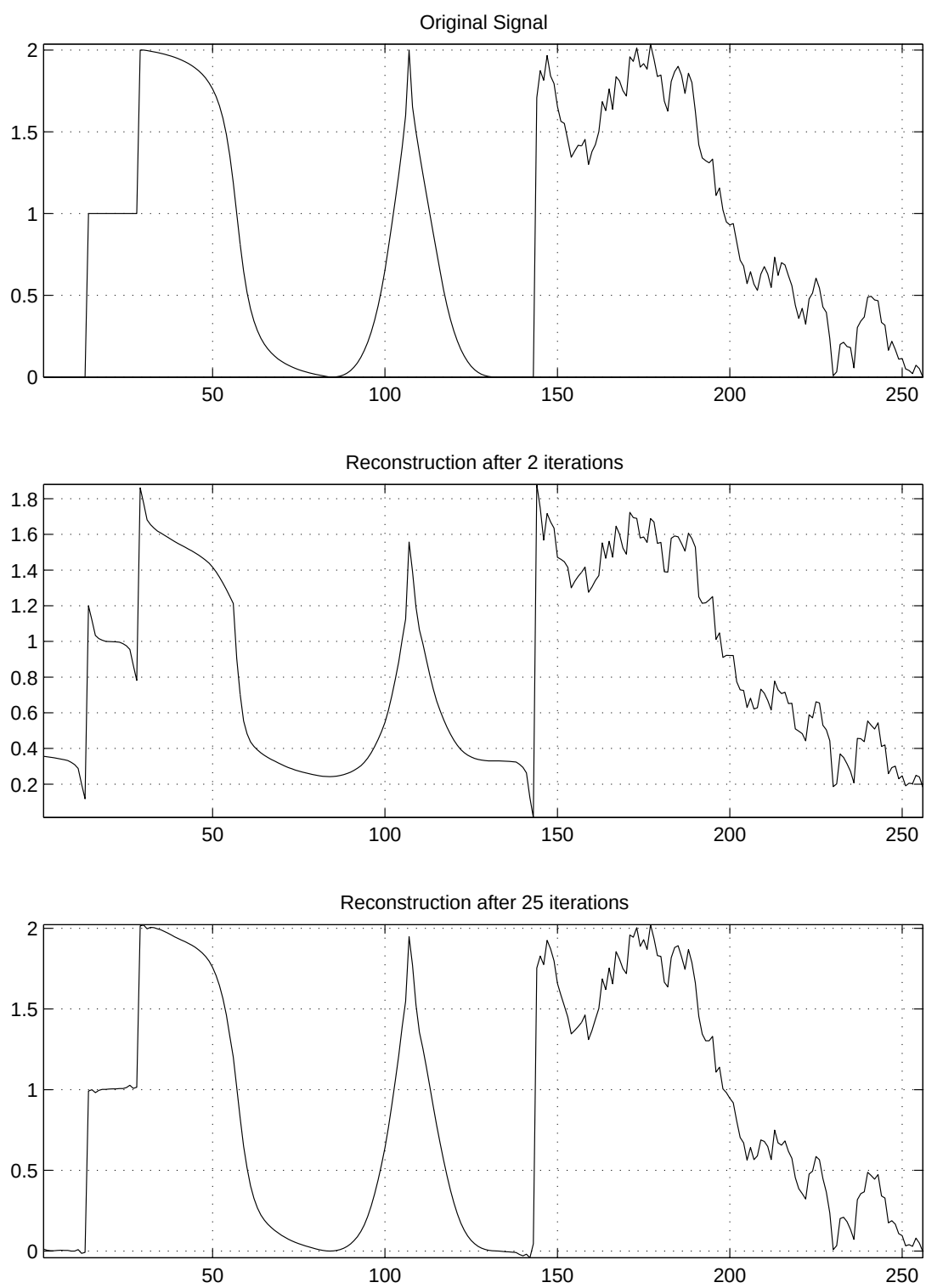


Figure 2.10: Wavelet extrema reconstruction example

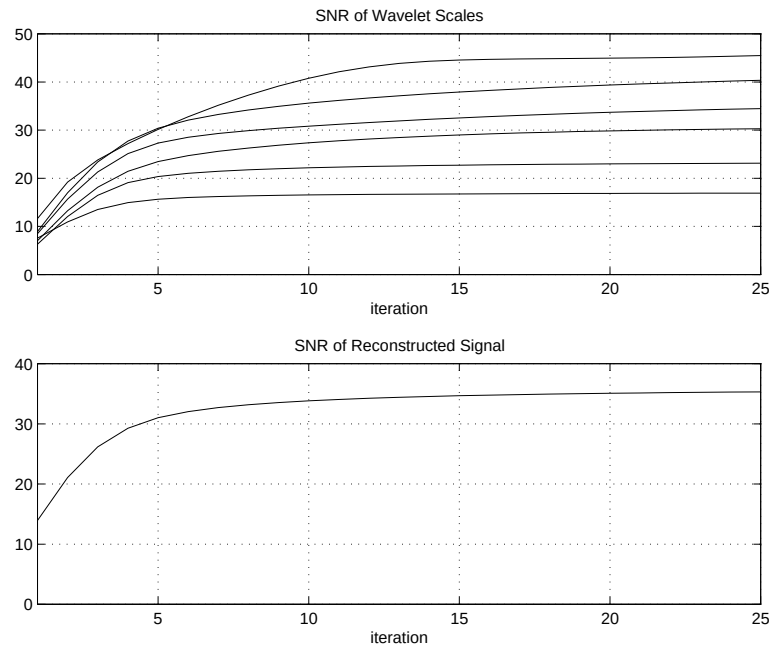


Figure 2.11: SNR of first six scales and reconstructed signal

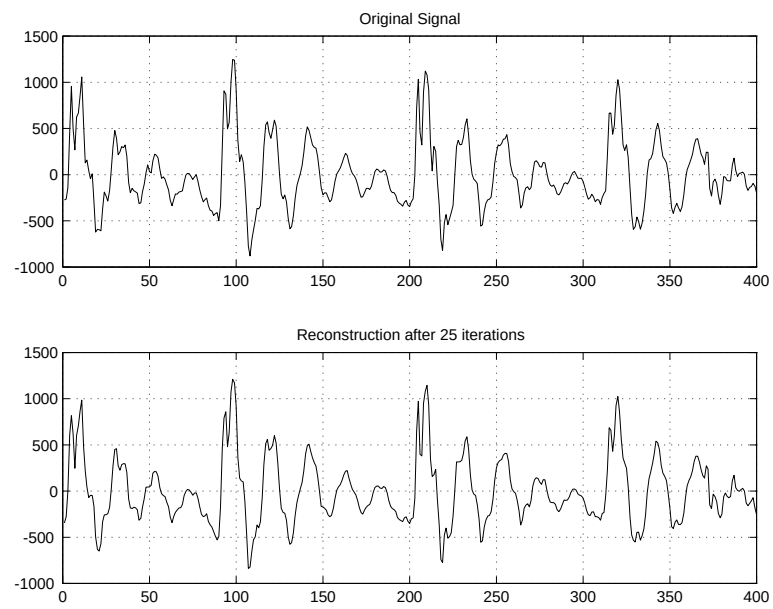


Figure 2.12: Wavelet extrema reconstruction speech example

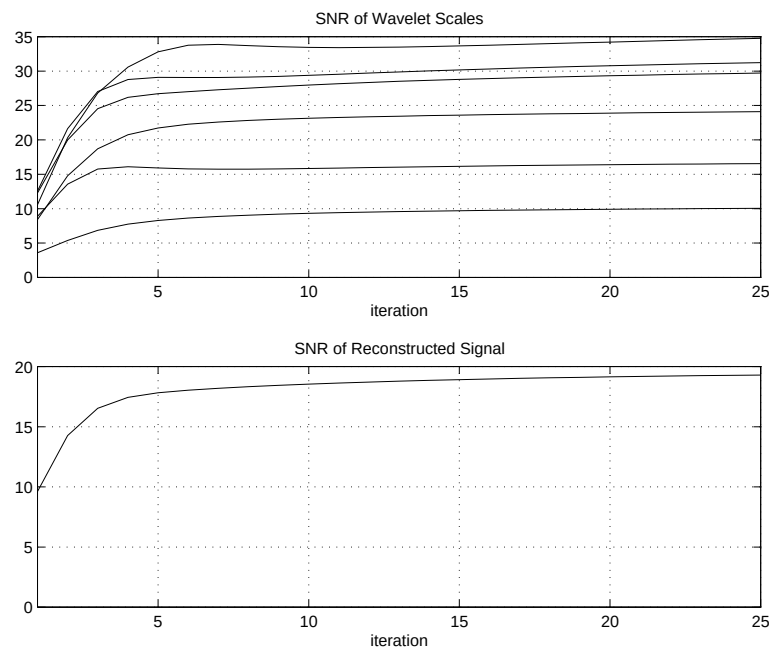


Figure 2.13: SNR of first six scales and reconstructed speech signal

Chapter 3

Pitch Estimation

Determining the pitch period or frequency present in a voiced speech signal is a fundamental part of many speech processing algorithms. Pitch estimation is important in applications of speech analysis, synthesis, coding, recognition, and speaker identification [15, 16, 27, 35, 36]. Other uses include linguistics, education (e.g. teaching correct pitch for hard of hearing or non-native speakers), and medicine (e.g. diagnosing voice diseases) [19]. Although considerable attention has been given to the problem of pitch estimation, the problem is far from solved due to the inherent complexities of the speech signal. In Chapter 1, a basic model of speech production was presented which allowed for a precise definition of the terms pitch period and pitch frequency. In this chapter, traditional single-channel pitch estimation techniques are presented as well as their extensions to multi-channel data. Wavelet-based pitch detection will then be explored, followed by a comparison of the efficacy of these methods under different levels of reverberation.

3.1 Autocorrelation Method

The autocorrelation function of a deterministic sequence is given by [36]:

$$\phi[k] = \sum_{m=-\infty}^{\infty} x[m]x[m+k]. \quad (3.1)$$

If the signal is periodic with period P , then $x[m] = x[m+P]$, and so the autocorrelation function is also periodic with period P :

$$\phi[k+P] = \sum_{m=-\infty}^{\infty} x[m]x[m+k+P] = \sum_{m=-\infty}^{\infty} x[m]x[m+k] = \phi[k]. \quad (3.2)$$

Also, $\phi[k]$ has a maximum at $k = 0$. This fact, together with the fact that the autocorrelation of a periodic signal is also periodic, suggests that the autocorrelation function has peaks at each integer multiple of

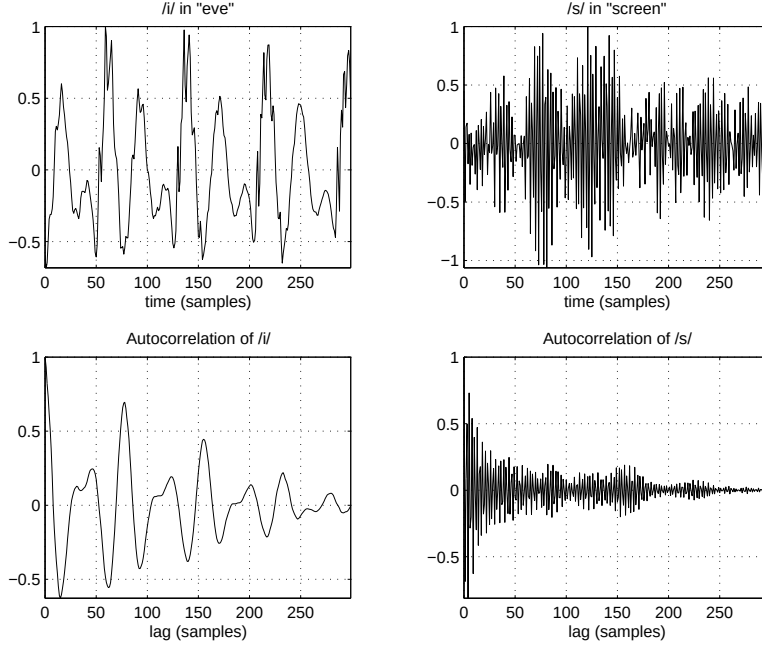


Figure 3.1: Voiced /i/, unvoiced /s/, and corresponding autocorrelation functions

the period P [36]. Therefore, any autocorrelation-based pitch estimator simply chooses the period as the lag (over all possible pitch periods) which maximizes the autocorrelation function. When dealing with speech signals, one typically calculates a short-term autocorrelation function (i.e. on a frame-by-frame basis) according to:

$$R_n[k] = \frac{1}{N} \sum_{m=-\infty}^{\infty} s[m]w[n-m]s[m+k]w[n-k-m] \quad (3.3)$$

where $w[n]$ is a window of length N . For a symmetric window centered around the origin, the short-term autocorrelation can be written as:

$$R_n[k] = \frac{1}{N} \sum_{m=0}^{N-1-k} s[n+m]w[m]s[n+m+k]w[k+m]. \quad (3.4)$$

When calculating the short-term autocorrelation of a speech signal, it is important to include at least two full pitch periods to allow for accurate estimation of the period. Figure 3.1 shows the vowel /i/ from the word “eve” and its autocorrelation function when using a rectangular window. Although pitch estimation only makes sense for voiced signals, Figure 3.1 also shows the autocorrelation function of an unvoiced sound - /s/ in “screen”, indicating a lack of periodicity.

Note that the peaks of the autocorrelation function of /i/ at integer multiples of the pitch period decay with increasing lag. The reasons for this are twofold [36]. First, the true pitch period tends to change from period to period. Due to the typically gradual nature of this change, the speech will most likely be

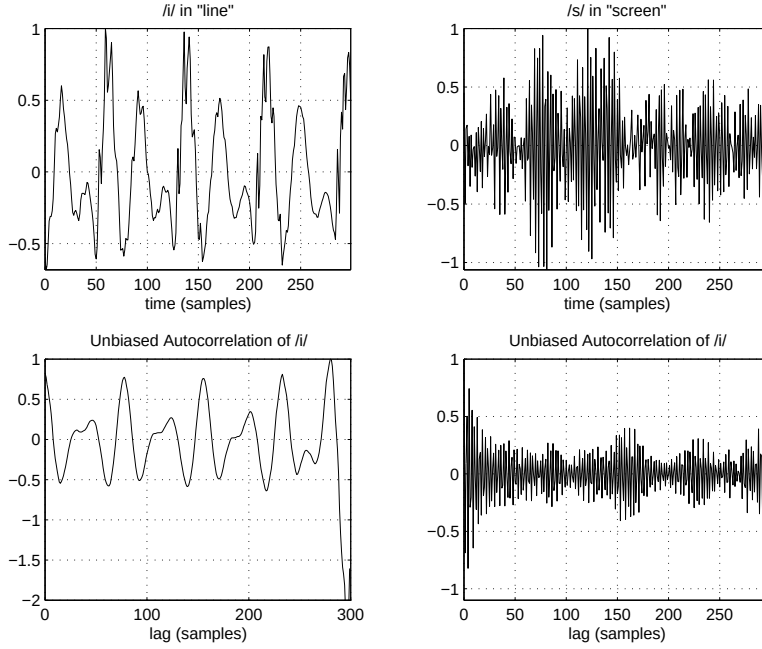


Figure 3.2: Voiced /i/, unvoiced /s/, and corresponding unbiased autocorrelation functions

more correlated with itself at a lag of one period than it would be for two or more periods. The second reason is that as the lag increases in Equation 3.4, there are fewer terms involved in the summation. The latter problem is often remedied through the use of the unbiased short-term autocorrelation function[36]:

$$R_n[k] = \frac{1}{N - |k|} \sum_{m=0}^{N-1-k} s[n+m]w[m]s[n+m+k]w[k+m]. \quad (3.5)$$

Figure 3.2 shows the unbiased autocorrelation functions for the sounds /i/ and /s/. Although the peaks at integer multiples of the pitch period exhibit less decay, there is a large amount of noise at high lags. This is due to the denominator in Equation 3.5. As $|k|$ approaches N , the autocorrelation estimate is less accurate.

The resonances of the vocal tract give rise to the presence of formant frequencies in the resulting speech signal. Often, the first formant has a frequency that is comparable to the pitch frequency. In order to dampen the effect of the first formant on the autocorrelation function, a center clipping operation is often used [19]. Center clipping is a non-linear operation described by [40]:

$$y[n] = \begin{cases} s[n] - C^+ & s[n] > C^+ \\ 0 & C^- \leq s[n] \leq C^+ \\ s[n] - C^- & s[n] < C^- \end{cases} \quad (3.6)$$

where $s[n]$ is the original speech signal, $y[n]$ is the clipped signal, and C^+ and C^- are clipping levels

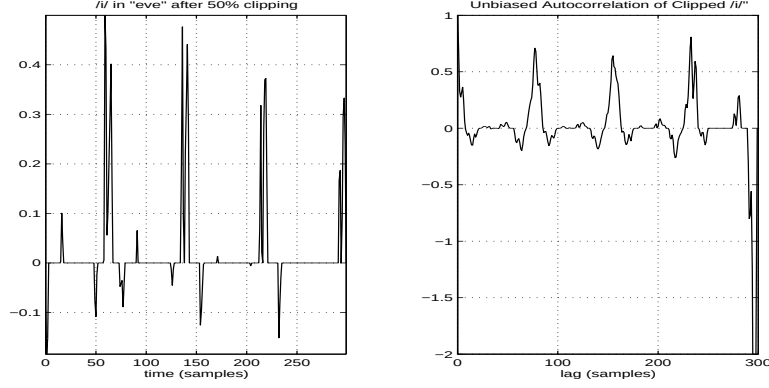


Figure 3.3: Voiced /i/ after 30% center clipping and corresponding unbiased autocorrelation

typically set to $\pm 30\%$ of the absolute maximum of $s[n]$. Figure 3.3 shows the vowel /i/ after performing center clipping, and the corresponding unbiased autocorrelation function.

The unbiased autocorrelation of the i^{th} of M channels is given by:

$$R_{i,n}[k] = \frac{1}{N - |k|} \sum_{m=0}^{N-1-k} s_i[n+m]w[m]s_i[n+m+k]w[k+m]. \quad (3.7)$$

where $s_i[n]$ is the unprocessed or center-clipped speech frame from the i^{th} microphone. We define the M -channel autocorrelation function as the average of the individual autocorrelation functions:

$$\overline{R}_n[k] = \frac{1}{M} \sum_{i=1}^M R_{i,n}[k]. \quad (3.8)$$

One advantage of this definition is that knowledge of the relative time-delays between microphones is not necessary because the autocorrelation function discards all phase information. The lag that maximizes $\overline{R}_n[k]$ over all possible pitch periods is chosen as the pitch estimate for the frame centered at time n .

3.2 Cepstrum Method

The speech production system for voiced sounds can be modeled as:

$$s[n] = e[n] \star g[n] \star v[n] \star r[n] \quad (3.9)$$

where $s[n]$ is the speech signal, $e[n]$ is the periodic excitation sequence, $g[n]$ is the glottal response, $v[n]$ is the vocal tract response, and $r[n]$ is the lip radiation response [15, 36]. Taking the Fourier transform of both sides of Equation 3.9, results in:

$$S(\omega) = E(\omega)G(\omega)V(\omega)R(\omega) = E(\omega)H(\omega) \quad (3.10)$$

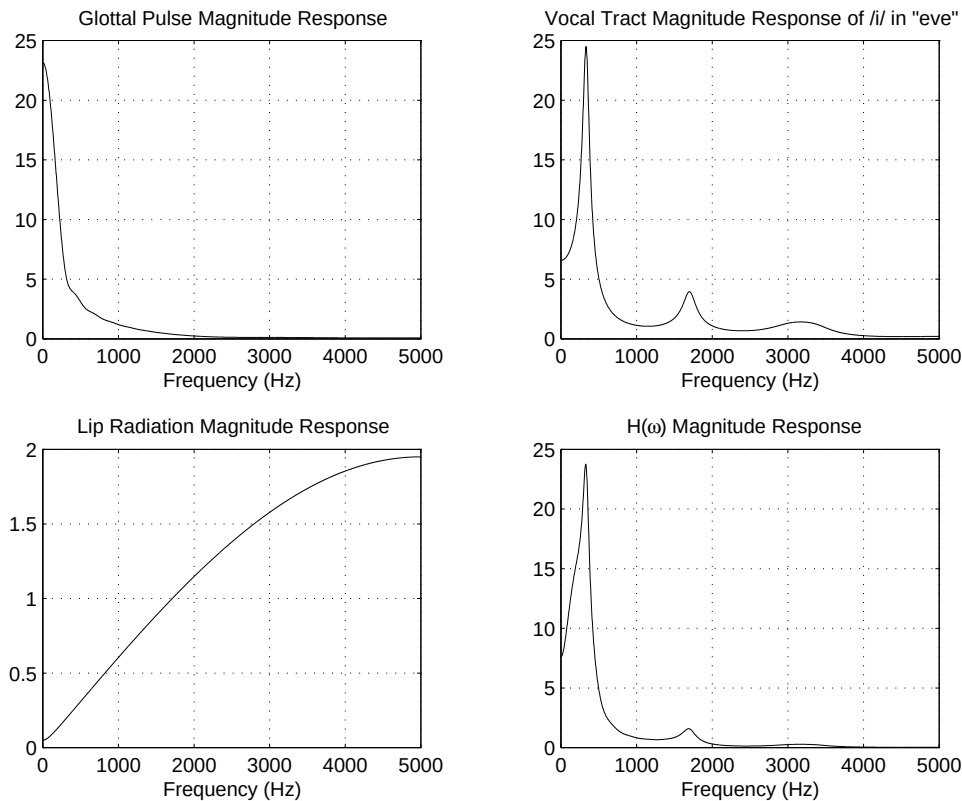


Figure 3.4: Magnitude responses of speech production components

where $H(\omega)$ summarizes the entire speech model driven by the excitation sequence. The cepstrum, $\hat{s}[n]$, of a speech signal, $s[n]$, is given by [15]:

$$\hat{s}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |S(\omega)| e^{i\omega n} d\omega. \quad (3.11)$$

That is, the cepstrum is the inverse Fourier transform of the log magnitude of the Fourier transform of $s[n]$. The motivation for using the cepstrum is that it transforms a convolution operation in the time domain to an addition operation in the cepstral domain. The usefulness of this homomorphic operation will be demonstrated in a series of plots. Figure 3.4 shows the magnitude response of the glottal pulse described in Chapter 1. Also plotted are the estimated magnitude response of the vocal tract and the lip radiation response. The product $H(\omega)$ is also shown. The excitation sequence (a periodic pulse train) serves as input into the speech system $H(\omega)$. The magnitude response of the excitation is shown in Figure 3.5 along with its log magnitude response. Also shown are $H(\omega)$ and its log magnitude spectrum. Taking logarithms of Equation 3.10 gives:

$$\log |S(\omega)| = \log |E(\omega)| + \log |H(\omega)|. \quad (3.12)$$

The inverse Fourier transform of Equation 3.12 is then calculated to form the cepstral relation:

$$\hat{s}[n] = \hat{e}[n] + \hat{h}[n]. \quad (3.13)$$

Now, since $\log |E(\omega)|$ has rapid oscillations relative to $\log |H(\omega)|$, $\hat{e}[n]$ will be located away from the origin (i.e. at “high quefrency”), while $\hat{h}[n]$ will be located near the origin. The bottom two plots of figure 3.5 show $\hat{e}[n]$ and $\hat{h}[n]$. The separation of $\hat{e}[n]$ and $\hat{h}[n]$ allows the pitch to be chosen as the quefrency where the cepstrum is a maximum over the desired range of pitch periods.

The cepstrum of the i^{th} of M channels is given by:

$$\hat{s}_i[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |S_i(\omega)| e^{i\omega n} d\omega. \quad (3.14)$$

where $S_i(\omega)$ is Fourier transform of the speech frame from the i^{th} microphone. We define the M -channel cepstrum as the average of the individual cepstra:

$$\bar{s}[n] = \frac{1}{M} \sum_{i=1}^M \hat{s}_i[n]. \quad (3.15)$$

3.3 Maximum Likelihood Pitch Estimation

Maximum likelihood pitch estimation is based on the assumption that the observed speech signal, $s[n]$, consists of a periodic component, $x[n]$, and a white, Gaussian noise sequence $u[n]$ as in [17, 19, 45]:

$$s[n] = x[n] + u[n] \quad n = 0 \dots N-1 \quad (3.16)$$

which, in vector notation, can be written as $S = X + U$. The signal $x[n]$ is assumed to be periodic with period P so that it can be written as $x[n] = q[n \bmod P]$ for some P length sequence $q[n]$, or Q in vector form. Under the assumption that the noise sequence components are independently and identically distributed Gaussian random variables with zero mean and variance σ^2 , the values of $q[n]$, P , and σ^2 can be estimated. These estimates are chosen to be most likely in a probabilistic sense.

For each value of n , $s[n]$ is a Gaussian random variable with mean $x[n]$ and variance σ^2 . Therefore, the conditional probability density function can be written as:

$$p(S|P, Q, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (s[n] - x[n])^2 \right\} \quad (3.17)$$

Using natural logs simplifies the ensuing optimization and results in:

$$\ln p(S|P, Q, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (s[n] - x[n])^2 \quad (3.18)$$

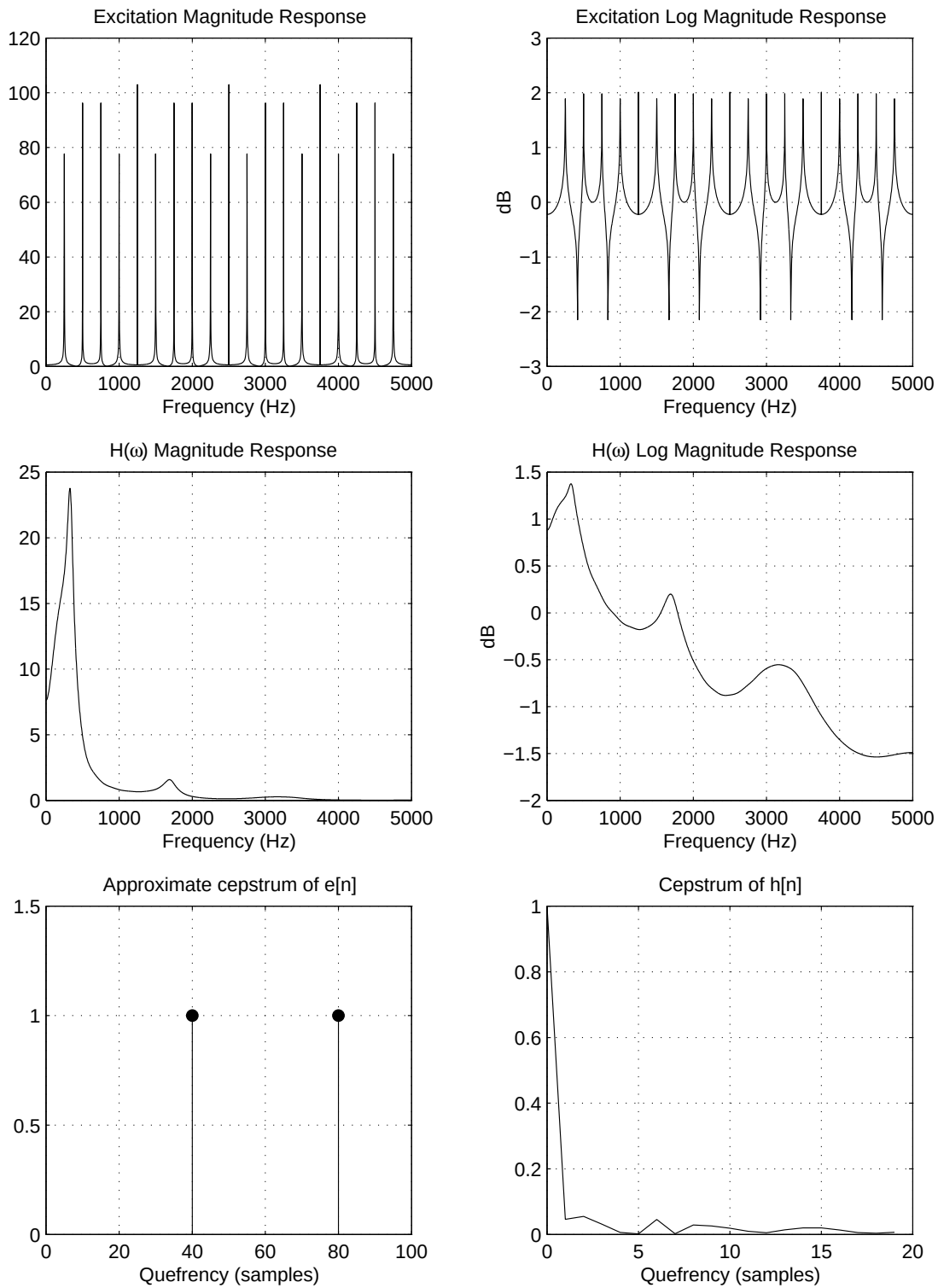


Figure 3.5: Separation of pitch and vocal tract in the cepstral domain

Now, due to the periodicity of $x[n]$, Equation 3.18 can be written as:

$$\mathcal{L} = \ln p(S|P, Q, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left[\sum_{l=0}^{L-1} \sum_{n=0}^{P-1} (s[n+lP] - q[n])^2 + \sum_{n=0}^{N-LP-1} (s[n+lP] - q[n])^2 \right] \quad (3.19)$$

where L is the number of complete periods of length P contained in the support of the speech signal (i.e. $L = \lfloor \frac{N}{P} \rfloor$). Taking the derivative with respect to $q[k]$ gives:

$$\frac{\partial \mathcal{L}}{\partial q[k]} = \begin{cases} \frac{1}{\sigma^2} \sum_{n=0}^P s[n+lP] - q[n] & n = 0 \dots N-LP-1 \\ \frac{1}{\sigma^2} \sum_{n=0}^{N-1} s[n+lP] - q[n] & n = N-LP \dots P-1 \end{cases} \quad (3.20)$$

which accounts for the fact that typically, the length of the observed sequence, $s[n]$, is not an integer multiple of the pitch period. Setting the derivative equal to zero gives the maximum likelihood estimate of the underlying periodic sequence:

$$\hat{q}[n, P] = \begin{cases} \frac{1}{L+1} \sum_{l=0}^P s[n+lP] & n = 0 \dots N-LP-1 \\ \frac{1}{L} \sum_{l=0}^{L-1} s[n+lP] & n = N-LP \dots P-1 \end{cases} \quad (3.21)$$

which is independent of σ^2 . Substituting these estimates, $\hat{q}[k]$, into Equation 3.18, and taking the derivative with respect to σ^2 results in a maximum likelihood estimate of:

$$\hat{\sigma}^2[P] = \frac{1}{N} \sum_{n=0}^{N-1} (s[n] - \hat{x}[n])^2 \quad (3.22)$$

where $\hat{x}[n] = \hat{q}[n \bmod P]$.

Finally, the pitch period P must be estimated. Substituting Equation 3.21 and Equation 3.22 into Equation 3.18 results in:

$$\ln p(S|P, Q, \sigma^2) = -\frac{N}{2} \ln(2\pi\hat{\sigma}^2[P]) - \frac{N}{2}. \quad (3.23)$$

Therefore, to maximize the log likelihood function with respect to P , the term $-\frac{N}{2} \ln(2\pi\hat{\sigma}^2[P])$ must be maximized. This translates to minimizing the estimated variance $\hat{\sigma}^2[P]$. It has been shown that the variance estimate can be rewritten as [45]:

$$\hat{\sigma}^2[P] = \frac{1}{N} \left[\sum_{n=0}^{N-1} s[n]^2 - \sum_{n=0}^{N-1} \hat{x}[n]^2 \right] \quad (3.24)$$

The first term in the above expression is the total energy in the observed speech signal and is independent of the choice of pitch period. Therefore, this variance is minimized by maximizing the energy of the estimated sequence $\hat{x}[n]$ which can be written as:

$$E[P] = \sum_{n=0}^{N-1} \hat{x}[n]^2 = (L+1) \sum_{n=0}^{N-LP-1} \hat{q}[n]^2 + L \sum_{n=N-LP}^{P-1} \hat{q}[n]^2 \quad (3.25)$$

In the event that $N = LP$, the energy function can be rewritten as:

$$E[P] = L \sum_{n=0}^{P-1} \hat{q}[n]^2 \quad (3.26)$$

which, after some manipulation, can be rewritten as:

$$E[P] = \frac{P}{N} \left[\phi_s[0] + 2 \sum_{l=1}^{L-1} \phi_s[lP] \right] \quad (3.27)$$

where $\phi_s[k] = \sum_{j=0}^{N-1-k} s(j)s(j+k)$ is the autocorrelation function of the observed speech signal. This has a very intuitive explanation. Instead of looking for an appropriate peak in the autocorrelation function, the maximum likelihood pitch estimator takes the pitch and all of its harmonics into account. Therefore, this estimator is very robust to missing harmonics or even the absence of the fundamental frequency itself. For example, with telephonic speech, the signal is often bandpass filtered so that the fundamental frequency is no longer present in the speech signal [19]. Another encountered problem is that the autocorrelation sometimes has isolated peaks due to noise present in the observed signal. If these peaks are larger than those due to the inherent periodicity of the speech, the autocorrelation method would provide an incorrect pitch estimate. In both of these examples, the maximum likelihood estimate should still perform quite well due to its inclusion of many harmonics when estimating the pitch period.

For quasi-periodic (i.e. realistic) speech signals, the energy function of Equation 3.27 always obtains its global maximum at $P = N$ [17, 45]. Also, when the noise signal is not Gaussian, the values of $E(lP)$ are often greater than $E(P)$ for integer $l > 0$. Therefore, a weighting function is typically applied to emphasize smaller values of P as in [17, 19, 45]:

$$\hat{E}[P] = E[P] - \frac{P}{N} \phi[0] \quad (3.28)$$

Figure 3.6 shows an example of the weighted maximum likelihood energy function for the vowel /i/ analyzed previously, indicating a clear peak at the pitch period. Although this estimation procedure makes the assumption of white i.i.d. noise samples, this technique is quite effective for practical signals where such assumptions are not necessarily justified. Note also that the clipped speech signal can be used as before to reduce the effect of the formant frequencies, although the performance improvement is less dramatic [19].

We define the maximum likelihood pitch estimator for the i^{th} channel as:

$$E_i[P] = \frac{P}{N} \left[\phi_{si}[0] + 2 \sum_{l=1}^{L-1} \phi_{si}[lP] \right] \quad (3.29)$$

where $\phi_{si}[k]$ is the autocorrelation of the speech frame at the i^{th} microphone. The multi-channel energy

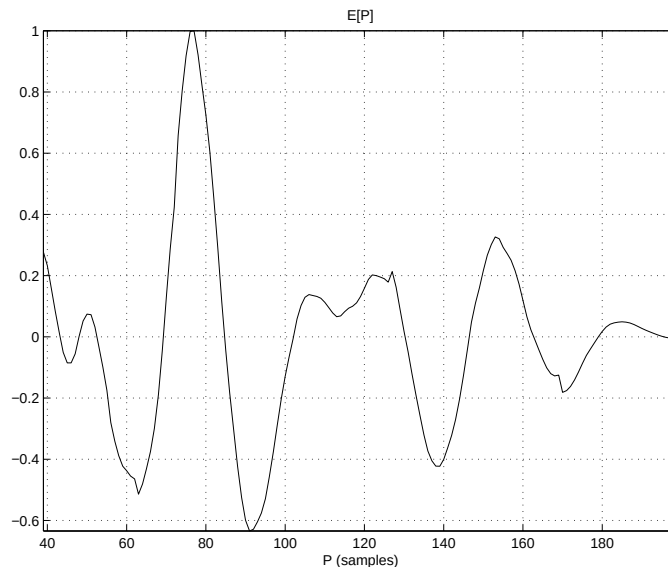


Figure 3.6: Maximum likelihood energy function

function is then defined as:

$$\overline{E}[P] = \frac{1}{M} \sum_{i=1}^M E_i[P] \quad (3.30)$$

3.4 Wavelet Pitch Estimation

As noted in Chapter 2, the largest wavelet coefficients correspond to the points of excitation driving the voiced speech production model. It was also seen that these segments are the most coherent over multiple channels. Wavelets have been used for pitch estimation in the past for speech signals degraded by additive noise rather than convolutional noise (i.e. reverberations) [21, 20]. These methods detected the points of excitations by locating large wavelet coefficients in scales where the pitch information was expected to occur. The distance between consecutive excitations was then used as the pitch estimate. Although this allowed for period-to-period pitch variations, reverberated speech often produces extra large peaks in the wavelet coefficients, causing such methods to break down quite rapidly. We devise a new scheme which combines these wavelet ideas with the robustness of the maximum likelihood pitch estimator. The main steps of this algorithm for each voiced frame of speech are as follows:

1. **Wavelet Calculation:** For each channel, calculate the first few scales of the wavelet transform. The number of scales depends on the sampling rate, as well as the expected pitch frequency of the speech signal.

2. **Event Detection:** At the coarsest scale, locate the largest wavelet coefficients in each channel. These are estimates of the points of voiced excitation.
3. **Extrema Refinement:** For each channel, trace the coarse resolution extrema up to a finer scale for better temporal resolution. Figure 3.7 shows an example of tracing wavelet maxima (indicated by circles) up to finer scales.
4. **Event-based Clipping:** At each scale, keep only the wavelet coefficients within a few (e.g. 2) samples of the fine resolution peaks. An example of this event-based clipped signal for one channel is shown in Figure 3.8.
5. **Maximum Likelihood Pitch Estimation:** The clipped signals (one for each channel) from the previous step are used as input to the multi-channel maximum likelihood pitch estimator described previously.

Two questions naturally arise from this wavelet pitch estimator:

1. Why should wavelet-based clipping perform better than simple center-clipping? Recall that with center clipping, the clipping is based on large amplitudes of the speech signal. Here, the clipping is based on large derivatives of the speech signal - points which indicate a possible voiced excitation and which are also less affected by reverberation.
2. Why is the maximum likelihood autocorrelation method used if the large wavelet coefficients indicate the start of each period? Although for many examples the largest wavelet coefficients indicate points of excitation, in a reverberant setting, false peaks are sometimes detected. This is a breakdown in the theory behind the event detecting wavelet, but such reverberant anomalies also tend to be periodic. Therefore, using the maximum likelihood method proves to be more robust than methods defining pitch as the distance between consecutive significant wavelet peaks.

3.5 Comparison of Pitch Estimators for Reverberant Speech

In this section, the center clipped autocorrelation, cepstrum, center clipped maximum likelihood, and wavelet pitch estimators are compared. A male speech signal sampled at 10 kHz was used for this experiment. Voiced segments of this signal were determined using standard short-term energy and zero crossing techniques. For each of the 85 resulting voiced frames, the “true” pitch was calculated manually. The single channel of speech was then convolved with simulated channel responses to obtain 8 channels of reverberated speech. The specifics of the room setup can be found in Chapter 1.

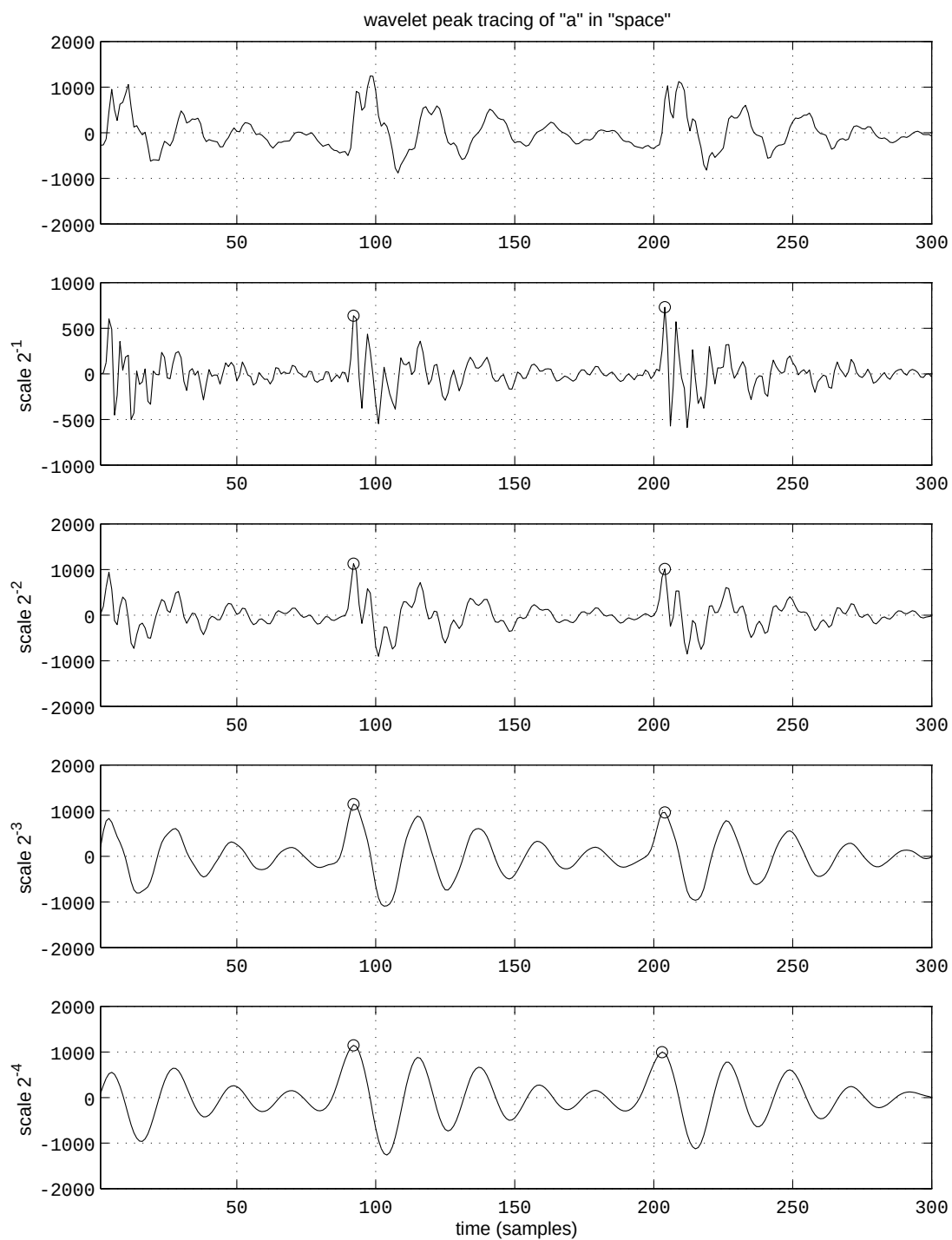


Figure 3.7: Wavelet peak tracing from coarse to fine scales

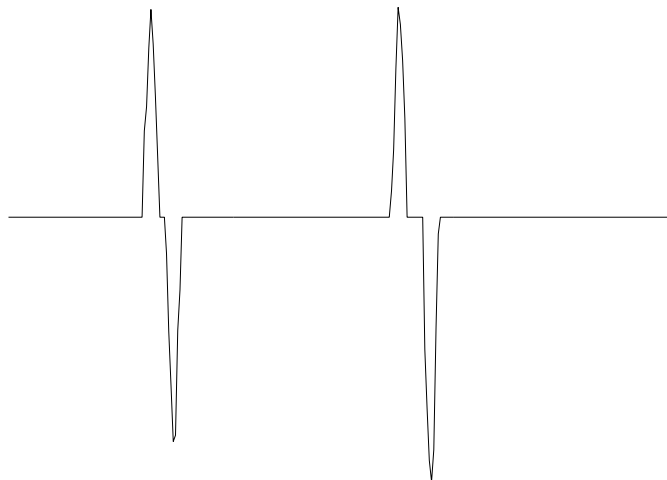


Figure 3.8: Wavelet-based clipped signal

In order to compare the performance of the different pitch estimators, two error criteria were considered. A pitch estimate was considered “correct” if it was within 3% of the true pitch:

$$\frac{|P_i - \hat{P}_i|}{P_i} \leq .03 \quad (3.31)$$

where P_i is the “true” pitch of the i^{th} voiced frame and $\hat{P}_i$ is the corresponding pitch estimate. Accuracy percentage is defined as the percentage of frames where a pitch algorithm chose the “correct” pitch. Figure 3.9 shows the accuracy percentages for the center-clipped autocorrelation (AC), cepstrum, center-clipped maximum likelihood (ML), and wavelet pitch estimators. The wavelet results are based on locating extrema at scale 2^{-5} and tracing up to scale 2^{-2} in the above algorithm. The wavelet pitch estimator performed the best for reverberant speech signals, though quite comparable to the maximum likelihood method using center clipping. This is to be expected since both methods maximize the same energy function with slightly different clipped inputs. The center-clipped autocorrelation method appears to be the least robust to reverberations.

The second error criterion considered was the mean squared error (MSE), defined as:

$$MSE = \frac{1}{F} \sum_{i=1}^F (P_i - \hat{P}_i)^2 \quad (3.32)$$

where F is the number of voiced frames used in the experiment (85 in this case). The MSE values for each algorithm are plotted in Figure 3.10. Note that although the accuracy percentages were fairly similar for the four methods, the wavelet and ML methods have considerably lower MSE errors. This suggests that these two techniques have a lower number of gross errors. From the results of this section, it is clear that pitch estimation of reverberant speech is a relatively solvable problem. This is due to the fact

that reverberations, though distracting to a listener, are also quasi-periodic, with the same period as the underlying speech signal.

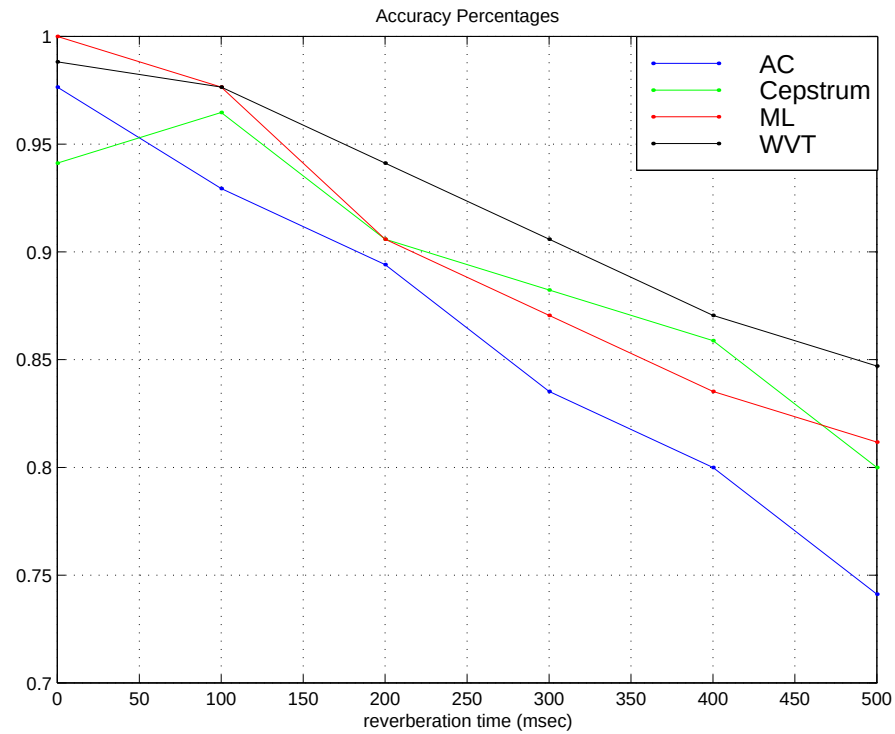


Figure 3.9: Accuracy percentages for pitch estimation algorithms

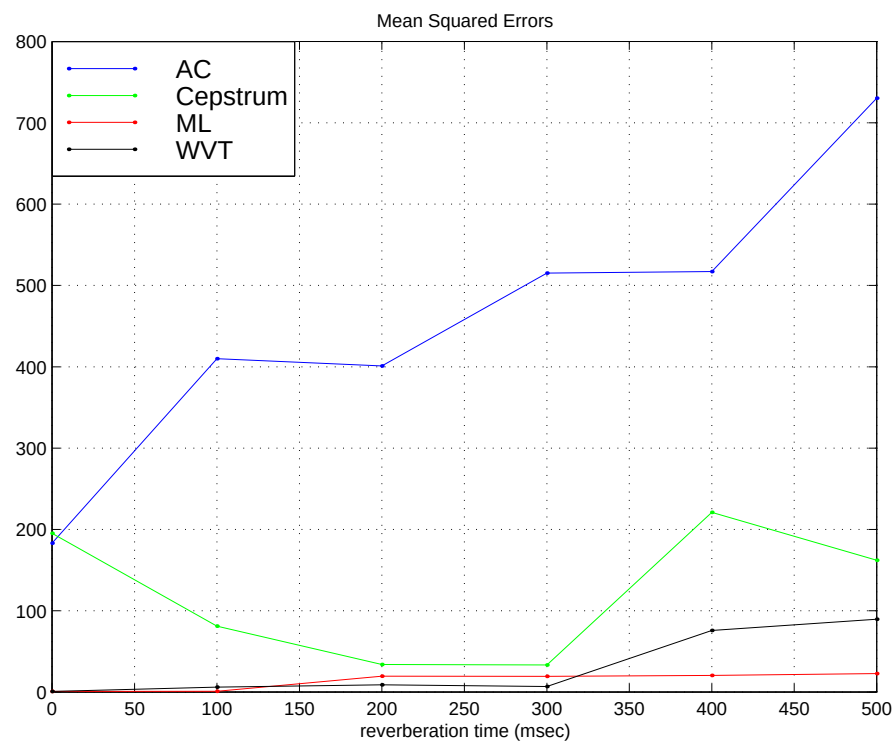


Figure 3.10: Mean-squared errors for pitch estimation algorithms

Chapter 4

Delay Estimation

4.1 Time Delay Estimation

In a two-channel setting, a signal emanating from a distant source will typically arrive at one microphone at a slightly different time than at the second microphone. This time delay is dependent on the source and microphone locations as well as the signal propagation speed in the environment. This can be modeled as:

$$\begin{aligned}x_1(t) &= s(t) + n_1(t) \\x_2(t) &= s(t - D) + n_2(t)\end{aligned}\tag{4.1}$$

where $s(t)$ is the source signal, $n_1(t)$ and $n_2(t)$ are noise signals, and D is the time delay. The noise signals are assumed to be uncorrelated with each other and with $s(t)$. Figure 4.1 shows the setup for a typical time delay estimation (TDE) problem. If we assume that the source is in the far field, the propagation can be modeled as a plane wave. Therefore, the bearing angle, θ_B , can be estimated from:

$$\theta_B = \cos^{-1} \left(\frac{cD}{d} \right)\tag{4.2}$$

where c is the speed of sound and d is the distance between the microphones. When using 3 or more microphones, a collection of bearing angles are obtained which can then be used to estimate the source location using triangulation methods. The accuracy of these bearing and location estimates are heavily dependent on the precision of the individual time delays. Equation 4.1 is easily extended to multiple channels, with all delays D_i specified relative to the reference signal, $x_1(t)$.

In order to model the TDE problem in a reverberant environment, Equation 4.1 is generalized to:

$$\begin{aligned}x_1(t) &= h_1(t) * s(t) + n_1(t) \\x_2(t) &= h_2(t) * s(t) + n_2(t)\end{aligned}\tag{4.3}$$

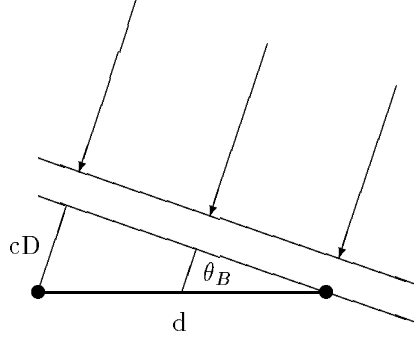


Figure 4.1: Two-channel time delay estimation problem

where $h_1(t)$ and $h_2(t)$ are the channel responses between the source and the two microphones. Note that in the case of $h_1(t) = \delta(t)$ and $h_2(t) = \delta(t - D)$, Equation 4.3 reduces to Equation 4.1. This chapter presents the generalized correlation TDE method and a new method which again exploits the wavelet extrema.

4.2 Generalized Cross Correlation Time Delay Estimation

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

$$R_{x_1x_2}(\tau) = E[x_1(t)x_2(t - \tau)] \quad (4.4)$$

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

$$G_{x_1x_2}(\omega) = \int_{-\infty}^{\infty} R_{x_1x_2}(\tau)e^{-i\omega\tau}d\tau \quad (4.5)$$

The generalized cross correlation (GCC) function is derived from a frequency-weighted version of the cross power spectrum:

$$\hat{R}_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} \psi(\omega)G_{x_1x_2}(\omega)e^{i\omega\tau}d\omega \quad (4.6)$$

where $\psi(\omega)$ is a frequency-dependent weighting function. The time delay estimate is then selected as the value of τ which maximizes $\hat{R}_{x_1x_2}(\tau)$. For the case where the additive noise is Gaussian and uncorrelated across scales, an optimal maximum likelihood weighting function derived from a frequency-dependent coherence measure is available. However, this ML delay estimator falls apart dramatically in the presence of even simple reverberations. An alternative weighting, the phase transform, has proven to be particularly

robust to reverberant conditions [8]. The phase transform method uses the weighting function [22, 32]:

$$\psi_{PHAT}(\omega) = \frac{1}{|X_1(\omega)X_2(\omega)|}. \quad (4.7)$$

Referring to the non-reverberant model of Equation 4.1, 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:

$$G_{x_1x_2}(\omega) = G_{s_1s_2}(\omega)e^{-i\omega D} + G_{n_1n_2}(\omega). \quad (4.8)$$

In the case of uncorrelated noise signals, $G_{n_1n_2}(\omega) = 0$. The cross power spectrum can be rewritten as:

$$G_{x_1x_2}(\omega) = |G_{x_1x_2}(\omega)|e^{-i\theta(\omega)} \quad (4.9)$$

where $\theta(\omega)$ is a phase function. By inspection, the magnitude is equal to $|G_{x_1x_2}(\omega)| = G_{s_1s_1}(\omega)$ and $\theta(\omega) = \omega D$. Therefore, the PHAT weighting function results in a generalized cross power spectrum of:

$$\frac{G_{x_1x_2}(\omega)}{|G_{x_1x_2}(\omega)|} = e^{-i\omega D} \quad (4.10)$$

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

$$\hat{R}_{x_1x_2}^{PHAT}(\tau) = \int_{-\infty}^{\infty} e^{-i\omega D} e^{i\omega\tau} d\omega = \delta(\tau - D). \quad (4.11)$$

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 4.11. Another pitfall of the phase transform is that it tends to amplify errors at frequencies where SNR is poor. However, despite these shortcomings for the additive-noise case, the phase transform has proven quite robust to room reverberations.

4.3 Wavelet-based Time Delay Estimation

In order to develop a wavelet-based time delay estimator, we first define the coherence between two signals, $x_1[n]$ and $x_2[n]$:

$$C(x_1, x_2) = \frac{2 \sum_{n=0}^{N-1} x_1[n]x_2[n]}{\sum_{n=0}^{N-1} x_1[n]^2 + \sum_{n=0}^{N-1} x_2[n]^2} \quad (4.12)$$

where N is the length of the data sequences. Note that $-1 \leq C(x_1, x_2) \leq 1$, with $C(x_1, x_2) = 1$ if and only if $x_1[n] = x_2[n]$. By defining:

$$C(x_1, x_2, d) = \frac{2 \sum_{n=0}^{N-1} x_1[n]x_2[n+d]}{\sum_{n=0}^{N-1} x_1[n]^2 + \sum_{n=0}^{N-1} x_2[n+d]^2}, \quad (4.13)$$

a reasonable estimate of the time delay between $x_1[n]$ and $x_2[n]$ is:

$$\hat{D} = \arg \max_d C(x_1, x_2, d). \quad (4.14)$$

For $M \geq 2$ channels, the coherence function can be generalized to:

$$C(x_1, x_2, \dots, x_M, d_2, \dots, d_M) = \frac{2}{M-1} \frac{\sum_{i=1}^M \sum_{j=i+1}^M \sum_{n=0}^{N-1} x_i[n+d_i] x_j[n+d_j]}{\sum_{i=1}^M \sum_{n=0}^{N-1} x_i[n+d_i]^2} \quad (4.15)$$

which consists of all pairwise combinations of the delay-compensated 2-channel coherence function. Note that all delays are relative to a reference channel, in this case, channel 1 (i.e. $d_1 = 0$). Note that $-1 \leq C(x_1, x_2, \dots, x_M, d_2, \dots, d_M) \leq 1$. This allows for the simultaneous estimation of each channel's delay according to:

$$\{\hat{D}_2, \dots, \hat{D}_M\} = \arg \max_{d_2, \dots, d_M} C(x_1, x_2, \dots, x_M, d_2, \dots, d_M). \quad (4.16)$$

We are now in a position to develop a multi-channel wavelet TDE having the following steps:

1. **Wavelet Calculation:** For each channel, calculate the first few scales of the wavelet transform.
2. **Event Detection:** At a coarse scale, locate the largest wavelet coefficients in each channel. These are estimates of the points of voiced excitation.
3. **Extrema Refinement:** For each channel, trace the coarse resolution peaks up to a finer scale for better temporal resolution.
4. **Event-based Clipping:** At each scale, keep only the wavelet coefficients within a few (e.g. 2) samples of the fine resolution peaks.
5. **Delay Estimation:** The clipped signals (one for each channel) are used as the $x_i[n]$ in the coherence calculations of Equation 4.15. The values of $\hat{D}_2, \dots, \hat{D}_M$ which maximize the coherence according to Equation 4.16 are chosen as the time delay estimates.

The advantage of this TDE over traditional methods is that it takes all channels into account when calculating the relative time delays. Although a multi-channel version of the PHAT method may be produced, the above technique was developed in order to illustrate the use of wavelet extrema for TDE.

Figure 4.2 shows a comparison of TDE methods for $M = 2$ and a reverberation time of 200 msec. The top plot involves channels 1 and 2, which have a relative delay of 4 samples. The bottom plot involves channels 1 and 3, which have a relative delay of 7 samples. For both examples, the cross-correlation of the two speech channels (CC), the PHAT autocorrelation function of Equation 4.11, and the two-channel wavelet coherence function described above (WVT 2) are plotted. The wavelet results are

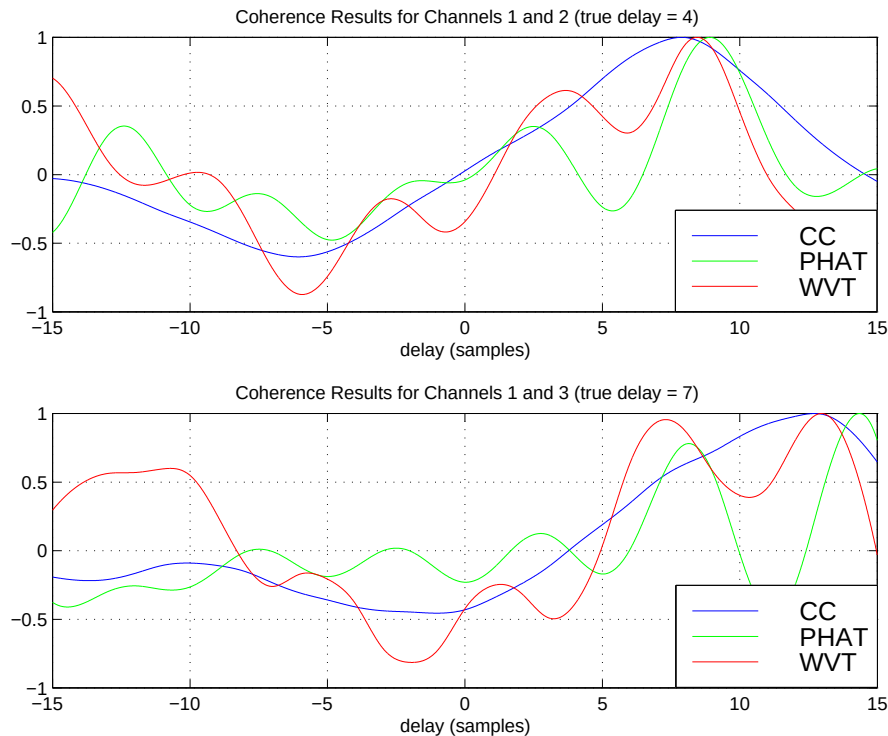


Figure 4.2: Comparison of time delay estimators for a reverberation time of 200 msec

based on locating extrema at scale 2^{-5} and tracing up to scale 2^{-2} in the above algorithm. The simple autocorrelation function tends to be heavily affected by the reverberations. The PHAT and wavelet based techniques both have peaks at or near the true delays, but these peaks are not global maxima and therefore will not be chosen as the delay estimates.

By calculating the coherence of all 3 channels ($M = 3$), the delays D_2 and D_3 of channels 2 and 3 are simultaneously estimated by maximizing $C(x_1, x_2, x_3, d_2, d_3)$. Figure 4.3 plots the coherence function vs. d_2 and d_3 . Note that by including 3 channels in the coherence function, a global peak is now located at desired location of $d_2 = 4$ and $d_3 = 7$. Figure 4.4 plots $C(x_1, x_2, x_3, d_2, D_3)$ vs. d_2 and $C(x_1, x_2, x_3, D_2, d_3)$ vs. d_3 (labeled WVT 3). These plots correspond to the two slices of the coherence function which intersect at the global maximum. The wavelet coherence functions for $M = 2$ from Figure 4.2 are also plotted to show how the simultaneous delay estimation of three channels improves the results for any 2 channels.

As M increases, the coherence function tends to have an increasingly distinct peak at the true delays $D_2, \dots, D_M$. Although not yet implemented, this peak will be selected using nonlinear optimization techniques. However, there are typically spurious local maxima in the coherence function which would be problematic for this optimization. The phase transform could be used to initialize the algorithm to

decrease the probability of converging to a non-global maximum in the coherence function. However, as observed above, this method is not completely robust to room reverberations. For speech following silence intervals in the recorded signal, the initial portions of the utterance are not affected by reverberations - there is no previous sound energy to reverberate and degrade the true speech. Therefore, in a real-time system, delay can be accurately estimated from these silence-speech transitions, the value of which can be used to re-initialize the search algorithm for the coherence maximum.

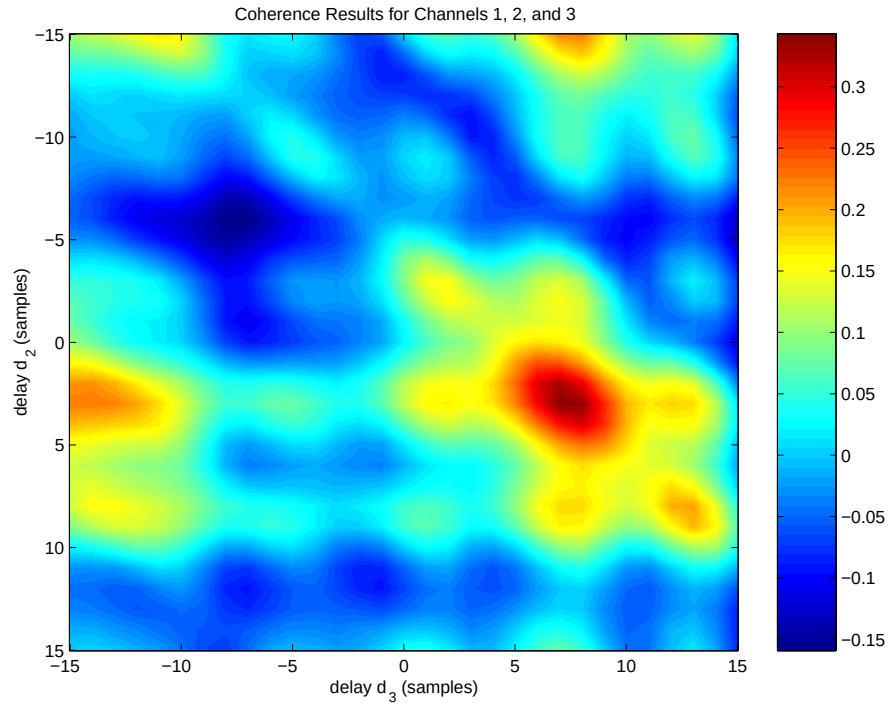
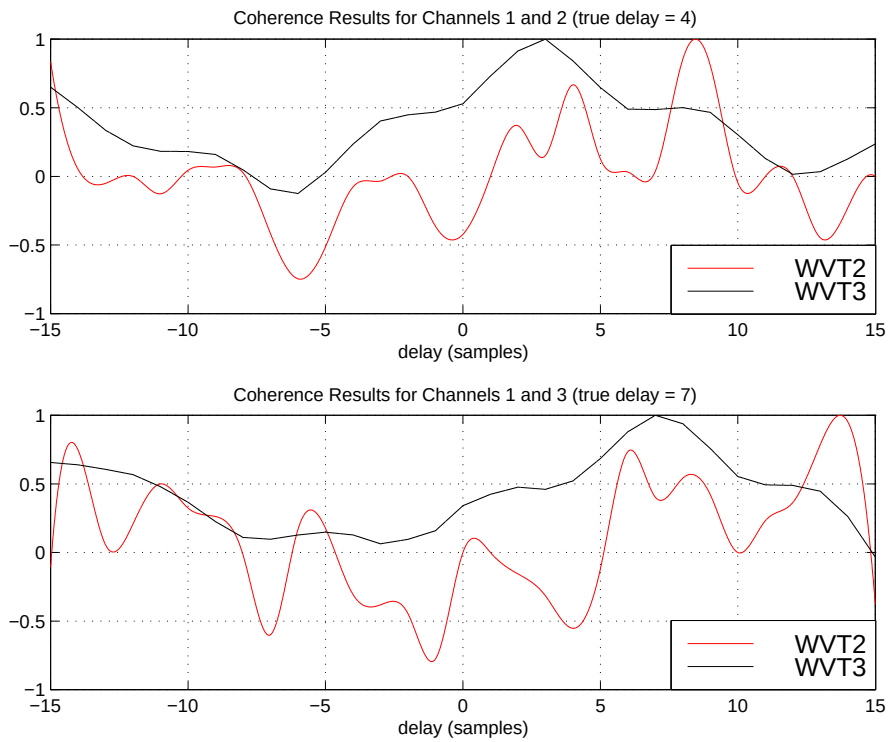


Figure 4.3: Wavelet coherence function for channels 1, 2, and 3

Figure 4.4: Comparison of the wavelet coherence method using $M = 2$ and $M = 3$

Chapter 5

Speech Dereverberation

5.1 Motivation and Previous Work

In a reverberant enclosure, certain signal parameters (e.g. pitch and time-delay) are more difficult to estimate. Furthermore, the intelligibility of the speech waveform is decreased. These degradations, both quantitative and qualitative in nature, provide the motivation for research in the area of reverberant speech enhancement. Recall that an enclosure's reverberant effects can be modeled as:

$$x_i[n] = h_i[n] * s[n] \quad (5.1)$$

where $h_i[n]$ is the channel response between the source and the i^{th} microphone. Past research has focused on two main approaches: inverse filtering of the channel responses and cepstral averaging.

Many approaches have attempted to invert the channel function $h_i[n]$ [29, 44]. Alternatively, this may be performed in the cepstral domain, where Equation 5.1 can be written as:

$$\hat{x}_i(n) = \hat{h}_i[n] + \hat{s}[n] \quad (5.2)$$

and the inverse filtering operation becomes a subtraction operation. In practice, the channel function is not completely minimum phase, but only the minimum phase portion can be inverted.

Although the inverse filtering approach is appealing in theory, the room impulse responses $h_i[n]$ or $\hat{h}_i[n]$ 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 [23, 31]. The inversion technique can actually further degrade the reverberated speech if the room response is not accurately estimated [28].

The second approach to speech dereverberation makes use of the fact that the speech spectrum is mainly located in the lower frequencies of the cepstrum [15, 24, 30, 36]. Also, the reverberation effects tend

to cause 'rapid ripples' in the log spectrum and therefore appear far away from the origin in the cepstrum [6]. By windowing the cepstrum, low-frequency 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 spectrum are often not well separated, and the optimal length of window used on the cepstrum is difficult to estimate [13, 23, 24]. Therefore, this technique is generally useful only for the simplest of echoes.

5.2 Wavelet-based Speech Enhancement

As discussed in Chapter 2, it is possible to approximately reconstruct signals or images from wavelet extrema [26]. This technique will now be used to enhance reverberated speech. In the following discussion, it is assumed that the different channels have been appropriately time-aligned (i.e. time delay compensated). Each step of this multi-channel method for a given speech frame is described below:

1. **Extrema Representation:** The first step of this method is to construct a single wavelet extrema representation from the multi-channel data, which will then be used to reconstruct a single speech signal. The wavelet transform of each channel is calculated first. This results in a set of extrema for each scale and each channel. Because reverberations affect each channel in a slightly different way, the local extrema of a given scale will not have the same time indices over all channels. In order to develop an extrema representation for each scale, a k-means clustering technique is used [5]. The following discussion will include only local maxima for simplicity. The process is then repeated for local minima. This technique will look for local maxima occurring in all channels at similar times, clustering them into one representative peak. Define $x_{i,j}$ as the time index of the j^{th} local maximum of the i^{th} channel at a certain scale. The clustering technique is described below:
 - (a) **Initialization:** Count the local maxima in one channel. Use this number as the number of clusters, N_c , expected over the length of the data. Set the initial cluster center, or seed points, s_j equal to maxima locations of an arbitrary channel (e.g. $s_j = x_{1,j}$). Specify the minimum number of channels, M_c , which must contribute to each cluster.
 - (b) **Cluster Assignment:** Assign each $x_{i,j}$ to the nearest cluster center, forming N_c clusters.
 - (c) **Cluster Center Update:** Re-specify the cluster centers as the average element of each cluster.
 - (d) **Convergence Check:** Repeat the previous two steps until the process converges (i.e. when no points are assigned to new clusters in the previous iteration).

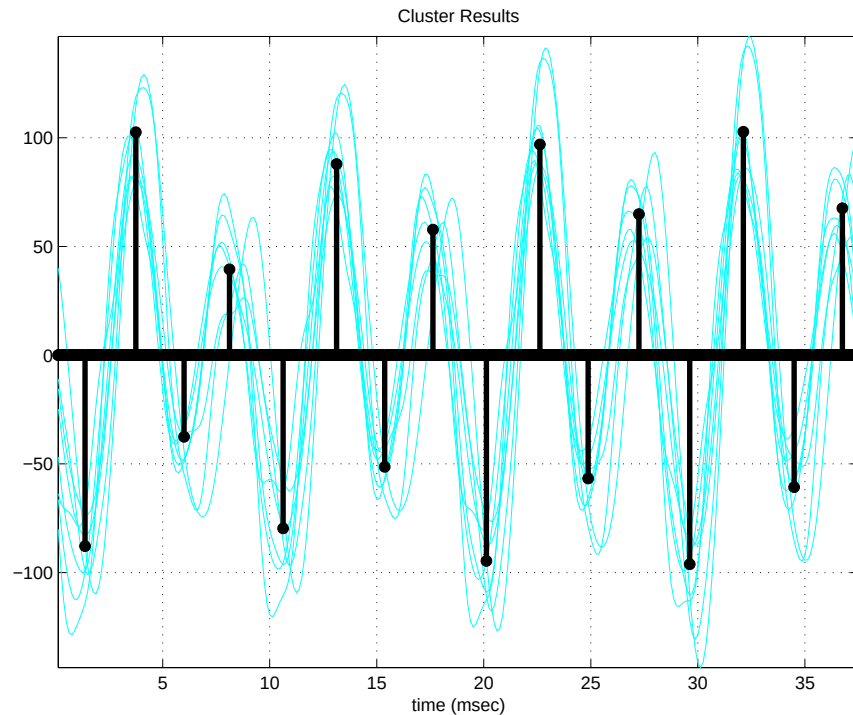


Figure 5.1: Plot of 8 reverberant channels and clustered peaks

- (e) **Prune:** Remove any clusters whose membership is smaller than M_c .
- (f) **Amplitude Assignment:** Now that maxima *locations* have been calculated, amplitudes are assigned as the average amplitude of the peaks used to generate each cluster.

This clustering process is then repeated for the local minima of each scale. Figure 5.1 shows the results of clustering 8 channels of the wavelet scale 2^{-5} . Figure 5.2 shows the clustered extrema of scale 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 energy.

2. **De-emphasis of Reverberant Extrema:** This step calculates a weighting function for the clustered peaks in order to de-emphasize those that are less coherent across all channels - the motivation being that channels have more coherent wavelet coefficients in areas of low reverberation and less coherent coefficients in areas of heavy reverberation. The coherence of each scale's wavelet coefficients is calculated using the multi-channel coherence function defined in Chapter 4. This coherence can be calculated on a short-term basis (i.e. every 15-20 msec) over the length of the speech frame. The normalized coherence function for scale 2^{-3} is shown in Figure 5.3. For coarser scales, the

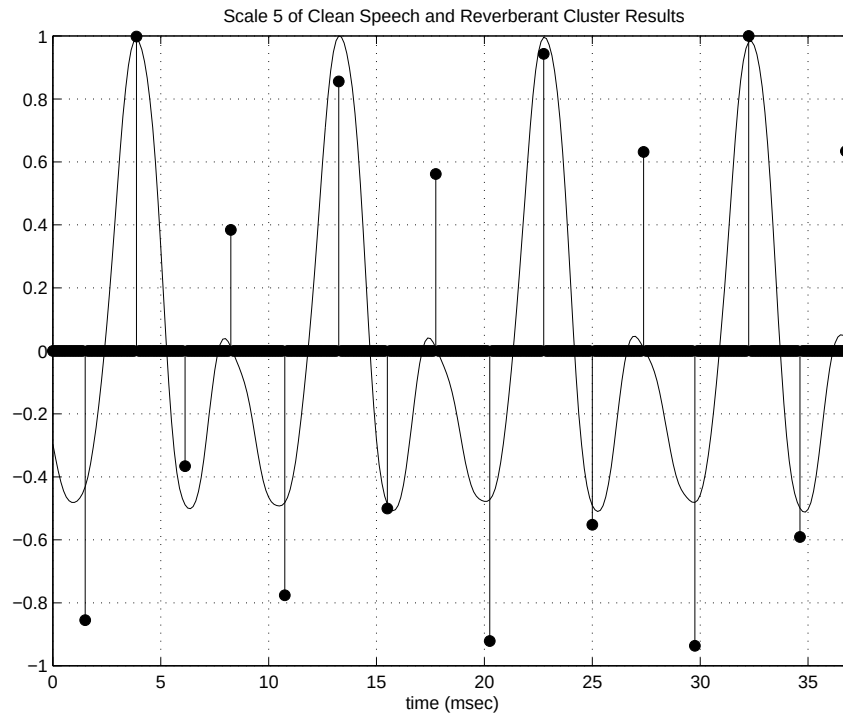


Figure 5.2: Clustered reverberant peaks and scale 2^{-5} of nonreverberant speech

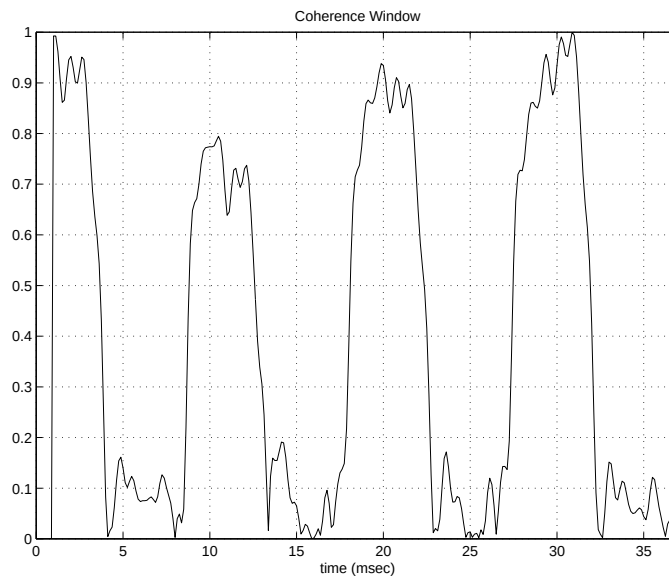
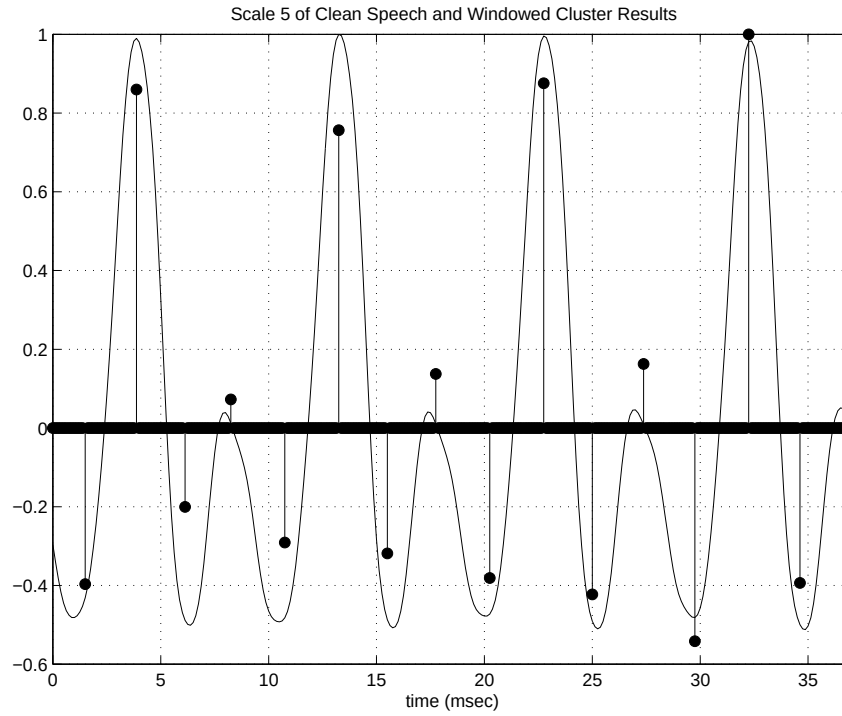
coherence is closer to 1 throughout the frame. This coherence function can be smoothed and/or thresholded to obtain a weighting function for the clustered peaks from the previous step. The results of weighting the clustered peaks by the coherence function are shown in Figure 5.4.

3. **Reconstruction from Wavelet Extrema:** Repeating the above process for each scale gives a single multi-scale extrema representation, de-emphasizing areas of low coherence. The alternating projection algorithm of Mallat can now be used to reconstruct the enhanced speech signal.

Figure 5.5 plots one channel of clean speech, one channel of reverberant speech (reverberation time = 200 msec), the results of a delay-and-sum beamformer, and wavelet extrema reconstruction. The delay-and-sum beamformed signal is calculated from:

$$s[n] = \frac{1}{M} \sum_{i=1}^M s_i[n - D_i] \quad (5.3)$$

where D_i is the delay of each channel relative to a reference channel. The beamformed approach tends to work well at the beginning of each pitch period, but is less effective late in the period due to the higher relative reverberation energy. The wavelet extrema approach was more successful in these portions of the signal.

Figure 5.3: Short-term coherence function for scale 2^{-3} Figure 5.4: Weighted peaks and scale 2^{-5} of nonreverberant speech

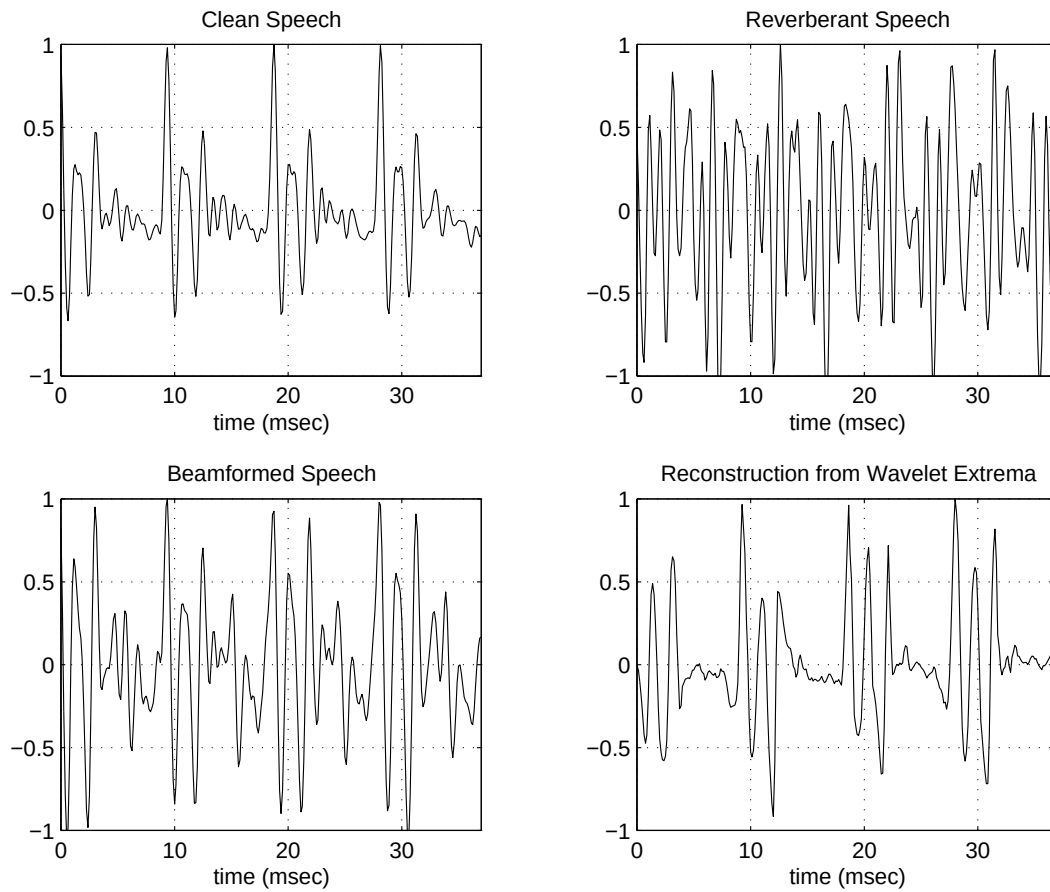


Figure 5.5: Single channel of clean speech, single channel of reverberant speech, beamformed result, and the wavelet extrema reconstruction result

Chapter 6

Conclusions

6.1 Summary of Current Progress

In this paper, a series of multi-channel wavelet techniques were presented for pitch estimation, time-delay estimation, and enhancement of reverberant speech. When using a derivative wavelet, the largest wavelet coefficients serve as event detectors, the event being the instant of glottal closure. It was observed that over multiple channels, the wavelet extrema corresponding to the points of glottal closure were more coherent than the extrema due to reverberations. Preliminary results indicate the following:

1. Due to the periodic nature of the reverberant components of the acquired speech signals, traditional pitch estimation methods were quite successful. A new wavelet-based technique was designed that kept high resolution wavelet coefficients only in the neighborhood of glottal closure instants. These coefficients were then fed into the maximum likelihood pitch estimator. This wavelet pitch estimator offered a moderate improvement over the maximum likelihood pitch estimator using center-clipped speech as input.
2. Time-delay estimation of reverberated speech proved to be a very challenging problem. A preliminary multi-channel coherence technique was presented that showed promise when using 3 channels simultaneously.
3. An enhancement technique was presented which reconstructed a dereverberated speech signal from wavelet extrema which were de-emphasized in areas of low coherence - an indication of heavy reverberation. This method showed promising results for the example in this paper.

Bibliography

- [1] P. Abry and P. Flandrin. On the initialization of the discrete wavelet transform algorithm. *IEEE Signal Processing Letters*, 1(2):32–34, February 1994.
- [2] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, 64(4):943–950, April 1979.
- [3] T. V. Ananthapadmanabha and B. Yegnanarayana. Epoch extraction of voiced speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(6):562–570, December 1975.
- [4] T. V. Ananthapadmanabha and B. Yegnanarayana. Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(4):309–319, August 1979.
- [5] M. R. Andenberg. *Cluster Analysis for Applications*. Prentice Hall, 1998.
- [6] D. Bees, M. Blostein, and P. Kabal. Reverberant speech enhancement using cepstral processing. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 977–980, 1991.
- [7] Z. Bervman and J. S. Baras. Properties of the multiscale maxima and zero-crossings representations. *IEEE Transactions on Signal Processing*, 41(12):3216–3231, December 1993.
- [8] M. Brandstein. 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, October 19–22 1997.
- [9] C. S. Burrus, R. A. Gopinath, and H. Guo. *Introduction to Wavelets and Wavelet Transforms: A Primer*. Prentice Hall, 1998.

- [10] R. A. Carmona. 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, 1995.
- [11] Y. M. Cheng and D. O’Shaughnessy. Automatic and reliable estimation of glottal closure instant and period. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 17(12):1805–1815, December 1989.
- [12] L. Cohen. *Time Frequency Analysis*. Prentice Hall, 1995.
- [13] D. Cole, M. Moody, and S. Sridharan. Position-independent enhancement of reverberant speech. *Journal of the Audio Engineering Society*, 45(3):142–147, March 1997.
- [14] Z. Cvetković and M. Vetterli. Discrete-time wavelet extrema representation: design and consistent reconstruction. *IEEE Transactions on Signal Processing*, 43(3):681–693, March 1995.
- [15] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete-Time Processing of Speech Signals*. Prentice Hall, 1987.
- [16] J. L. Flanagan. *Speech Analysis, Synthesis, and Perception*. Springer-Verlag, 1972.
- [17] D. H. Friedman. Psuedo-maximum-likelihood speech pitch extraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3):213–221, June 1997.
- [18] J. D. Harris and D. Nelson. Glottal pulse alignment in voiced speech for pitch determination. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 519–522, 1993.
- [19] W. Hess. *Pitch Determination of Speech Signals*. Springer-Verlag, 1983.
- [20] S. Kadambe. 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, 1992.
- [21] S. K. Kadambe and G. F. Boudreaux-Bartels. Application of the wavelet transform for pitch detection of speech signals. *IEEE Transactions on Information Theory*, 38(2):917–924, March 1992.
- [22] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, August 1976.

- [23] Q. G. Liu, B. Champagne, and P. Kabal. 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, May 1995.
- [24] Q. G. Liu, B. Champagne, and P. Kabal. A microphone array processing technique for speech enhancement in a reverberant space. *Speech Communication*, 18:317–334, 1996.
- [25] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.
- [26] S. G. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(7):710–732, July 1992.
- [27] J. D. Markel and A. H. Gray Jr. *Linear Prediction of Speech*. Springer-Verlag, 1976.
- [28] J. Mourjopoulos. On the variation and invertibility of room impulse response functions. *Journal of Sound and Vibration*, 102(2):217–228, 1985.
- [29] S. T. Neely and J. B. Allen. Invertibility of a room impulse response. *Journal of the Acoustical Society of America*, 66(1):165–169, July 1979.
- [30] A. V. Oppenheim and R. W. Shafer. *Discrete-Time Signal Processing*. Prentice Hall, 1989.
- [31] A. P. Petropulu and S. Subramaniam. Cepstrum based deconvolution for speech dereverberation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 9–12, 1994.
- [32] A. G. Piersol. Time delay estimation using phase data. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):471–477, June 1981.
- [33] I. Pintér. Perceptual wavelet-representation of speech signals and its application to speech enhancement. *Computer Speech and Language*, 66(1):165–169, July 1979.
- [34] B. Porat. *Digital Processing of Random Signals: Theory and Methods*. Prentice Hall, 1994.
- [35] L. R. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.
- [36] L. R. Rabiner and R. W. Shafer. *Digital Processing of Speech Signals*. Prentice Hall, 1978.
- [37] O. Rioul. A discrete-time multiresolution theory. *IEEE Transactions on Signal Processing*, 41(8):2591–2606, August 1993.

- [38] O. Rioul and P. Duhamel. Fast algorithms for discrete and continuous wavelet transforms. *IEEE Transactions on Information Theory*, 38(2):569–586, March 1992.
- [39] M. J. Shensa. The discrete wavelet transform: wedding the à trous and mallat algorithms. *IEEE Transactions on Signal Processing*, 40(10):2464–2482, October 1992.
- [40] M. M. Sondhi. New methods of pitch extraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(1):24–33, February 1977.
- [41] P. Srinivasan and L. H. Jamieson. Wavelets based on splines: an applications. *SPIE*, 2825(4):456–464, 1996.
- [42] G. Strang. Wavelets. *American Scientist*, 82:250–255, May-June 1994.
- [43] H. W. Strube. Determination of the instant of glottal closure from the speech wave. *Journal of the Acoustical Society of America*, 56:1625–1629, November 1974.
- [44] M. Tohyama, R. H. Lyon, and T. Koike. 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, 1993.
- [45] J. D. Wise, J. R. Caprio, and T. W. Parks. Maximum likelihood pitch estimation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5):418–423, October 1976.