Sine-Wave Amplitude Coding using Wavelet Basis Functions

by

Pankaj Oberoi

S.B., Massachusetts Institute of Technology (1991)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Feburary 1996

Author .....
Department of Electrical Engineering and Computer Science
December 22, 1995

Certified by .....
Robert J McAulay
Senior Staff, MIT Lincoln Laboratory
Thesis Supervisor

Certified by .....
Thomas Quatieri
Senior Staff, MIT Lincoln Laboratory
Thesis Supervisor

Accepted by .....
Frederic R. Morgenthaler
Chairman, Departmental Committee on Graduate Students

Sine-Wave Amplitude Coding using Wavelet Basis Functions

by

Pankaj Oberoi

Submitted to the Department of Electrical Engineering and Computer Science
on December 22, 1995, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

This thesis presents an alternate method for low-rate speech coding using wavelet basis functions to decompose the spectral envelope. Low-rate speech coding of the spectral envelope using correlated channel gains was implemented in the sinusoidal transform coder (STC). The spectral envelope is like a $1/f$ signal, and based on Wornell's [37] work which showed that for $1/f$ and nearly $1/f$ signals the wavelet coefficients are uncorrelated, the wavelet coefficients of the spectral envelope should be uncorrelated and more robust for coding.

A discrete symmetric-periodic wavelet algorithm was implemented to avoid problems that arise with decomposition of finite duration signals. Initial tests were performed using two sets of wavelet basis functions: Daubechies' compactly supported wavelet basis functions and Mallat's spline wavelet basis. The spline basis functions were chosen over Daubechies' function for the wavelet-based sinusoidal transform coder (WSTC) because they require fewer coefficients to represent the spectral envelope. The first five wavelet scale coefficients did not contribute much information to the represntation and were discarded allowing the representation to be reduced from 512 to 33 coefficients. The coefficients at each wavelet scale had a Gaussian distribution, and their variances were similar to the variance progression seen by Wornell [37]. A quantization scheme based on the coefficient histograms was developed to reduce quantization errors between the original and reconstructed envelopes.

A WSTC system was implemented at 3700 bits per second and compared to a similar STC system that was implemented at 4800 bits per second. The two systems sounded comparable, but the STC system performed slightly better at synthesizing unvoiced speech. Further enhancement of the coding schemes to exploit the statistical properties of the wavelet coefficients could improve the quality of the synthesized speech or reduce the coding rate.

Thesis Supervisor: Robert J McAulay
Title: Senior Staff, MIT Lincoln Laboratory

Thesis Supervisor: Thomas Quatieri
Title: Senior Staff, MIT Lincoln Laboratory

Acknowledgements

I would like to first thank my advisors Robert McAulay and Thomas Quatieri without whom this thesis would not have been possible. They provided valuable insight which motivated me to continue in the field of speech and hearing. I am grateful to all the people in the wavelet journal club at Lincoln Laboratory without whom the wavelet theory used in this thesis would have taken much longer to understand. It was a pleasure working with the members of the speech group at Lincoln Laboratory.

I would like to thank Sridhar for his insightful comments on several of the drafts. Janet has been a great help by reading the final draft. Quentin was an O.K. lab partner, but his comments on my research made me do a lot of thinking and his comments on my first draft were very helpful.

I would like to acknowledge several people who have kept me sane throughout my graduate career. My roomates James Youngclaus and Jeff Moeller have provided me with the best years I have had at MIT and lasting friendships. Another special person, Rosalie, supported me and gave me hope and inspired me to continue when things were not going so well.

Finally I would like to thank my parents Raj and Baljeet Oberoi for providing me with lots of love and support throughout my life. Nothing can compare with the brotherly love and support Manish has given, I hope I can support him as much in his endeavors.

This work was sponsored by the Department of the Air Force.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Improved speech representations can lead to advances in the areas of speech recognition, synthesis, enhancement, and coding. Speech production and auditory models have been developed to understand which speech parameters are essential for retaining high quality speech while reducing the number of parameters. By coding only those parameters of the speech signal needed for intelligibility and high quality, the bandwidth required to transmit speech is reduced. In communications areas such as cellular, mobile, and military radio communications where the transmission bandwidth is limited, reduced representations can allow for an increase in the number of connections and a re-allocation of bits to error protection for more robust communication. Current low-rate coders are able to reasonably represent speech at rates from 4800 to 1200 bits per second [23].

1.2 Description of a Canonical Vocoder

There is a tradeoff between the perceptual quality of coded speech and the amount of data allocated for coding. Even though speech has a bandwidth of 10 kHz and humans can resolve tones at over 16 bits of intensity variations, but only a small subset of a combination of these sounds is interpreted as speech. To fully represent the speech signal over 160,000 bits per second are needed, but for a perceptual represent of intelligible speech, many fewer bits are needed. The objective of a low-bit-rate vocoder is to produce good-sounding intelligible speech while compressing it to low data rates below 4800 bits per second. In addition to low

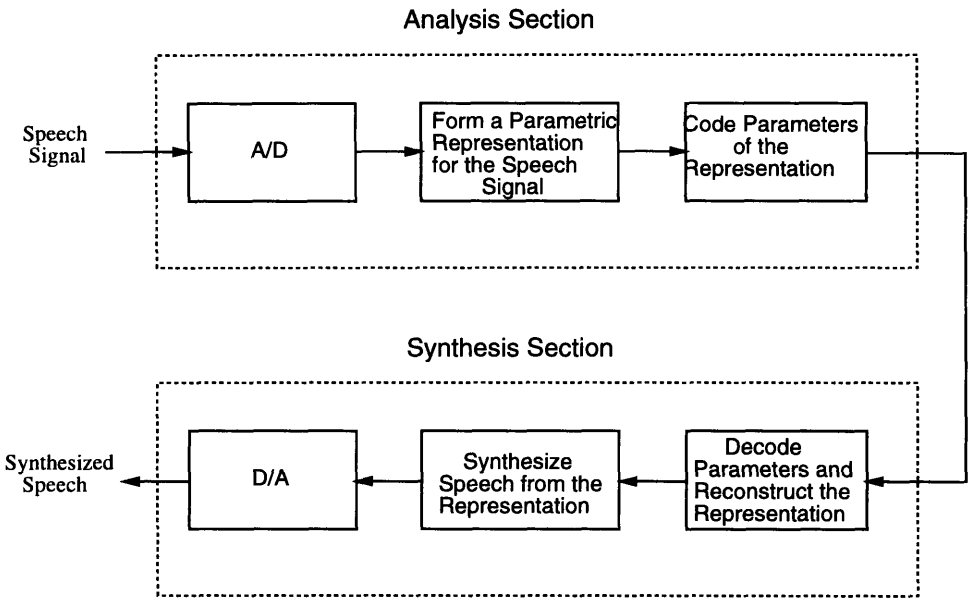


Figure 1-1: Block diagram of a generic vocoder.

rates, the vocoder must be robust to additive acoustical noise, and the coded signal needs to be resilient to transmission noise and errors.

A speech coding system consists of an analysis and a synthesis section as shown in Figure 1-1. In the analysis, the speech signal is sampled using an A/D converter and transformed into a minimal set of parameters which are then quantized for storage or transmission. In the synthesis section, the quantized parameters are first decoded to recover the speech parameters, from which a synthetic speech signal is generated and converted back to a continuous signal using a D/A converter. Compared to the original speech signal, the quality of the synthesized speech may be degraded due to the inadequacies of the parametric representation, and the algorithm used to quantize the parameters.

1.3 Sine-Wave Amplitude Representation

Several speech representations have been developed for coding speech, including subband coding, linear predictive coding, phase coding, formant coding, and cepstral coding [14, 30]. Another representation, developed at Lincoln Laboratory [23], uses the sinusoidal model (described in Chapter 2) to represent the speech signal. This thesis uses the speech representation developed for the sinusoidal transform coder (STC).

In the sine-wave representation, the speech signal is decomposed into a sum of sine-waves.

McAulay and Quatieri have shown that the amplitudes of these sine-waves form a sufficient set of speech parameters [23]. The sine-wave amplitudes are represented by the spectral envelope which is sampled and quantized in the form of channel gains in the STC system. The wavelet-based sinusoidal transform coder (WSTC), developed in this thesis, decomposes the spectral envelope into a set of wavelet coefficients.

1.4 Wavelet Introduction

Current applications of wavelet research have centered around removing redundancies found in images for the purpose of image coding. One goal of this thesis project is to determine if, by choosing a good wavelet basis, redundancies in the spectral content of speech can be removed.

Wavelet theory has been compared to subband and multirate systems which have been used to code speech [30] in which the speech waveform is separated into different spectral bands. The wavelet transform uses basis functions similar to the filters in the subband systems, but the basis functions are applied to the speech spectral envelope rather than to the speech waveform itself. This thesis will test the hypothesis that basis functions with shapes similar to the formant structure in the envelope should reduce the number of wavelet coefficients.

Recent work done by Wornell [37] has shown that orthonormal wavelets basis expansions act as a Karhunen-Loève-like representation for $1/f$ processes and "appear to be robust, nearly optimal representations for all $1/f$ processes, including nearly $1/f$ processes." [36] Wornell's work has shown that the wavelet coefficients obey a variance progression which can be useful in parameter estimation and coding. The wavelet coefficients for a $1/f$ or nearly $1/f$ process are uncorrelated along the wavelet scale and across scales. The complex cepstrum of the speech spectrum decays at least as fast as $1/n$ [26], therefore, the log magnitude of the spectral envelope will yield a $/(1/f)^n$ process. Based on Wornell's results, a wavelet decomposition of the speech spectral envelope should produce a set of uncorrelated coefficients.

1.5 Objective of Thesis

The overall goal of this thesis is to obtain a wavelet representation of the speech spectrum for low-bit-rate coding at rates near 4800 bps while maintaining high quality speech. The focus is on representing and quantizing the speech spectral envelope. In order to maintain good quality synthesized speech, the difference between the original and reconstructed spectral envelopes must be small. A complete wavelet representation should exactly reconstruct the envelope, thereby, producing high-quality speech; however, the complete wavelet representation contains too many coefficients for low-rate coding. An analysis of the wavelet coefficients should yield a subset that are necessary for reconstruction of the envelope.

Chapter 2

Speech Production Model and the Sine-Wave Representation

A good speech production model is important because unwanted sounds can be generated by a poor representation of speech. This chapter describes a typical speech production model and how speech can be represented using that model. The sinusoidal model, based on the the acoustic speech production model, produces parameters which can be quantized for storage or transmission. In this model, the speech signal is modeled as a sum of sine-wave amplitudes, which are the primary parameters coded. The sine-wave amplitudes can be gotten from sampling the spectral envelope.

2.1 Acoustic Speech Production Model

Speech production can be modeled as an excitation produced by the vocal cords which is passed through a linear time-varying system which represents the glottal system, the vocal tract, and the radiation at the lips [30]. The model, shown in Figure 2-1, is the underlying model for most speech analysis/synthesis systems [30], and can be further simplified as an excitation and a system function ($H_s(\omega, t)$).

If speech production is assumed to be linear system, the vocal tract can be modeled as a cascade of linear acoustic tubes. The transfer function of the vocal tract is simplified to an all-pole system, making it a computationally efficient representation because the system can be estimated by linear prediction analysis using all-pole coefficients.

The transfer function of an all-pole system is characterized by peaks at the natural fre-

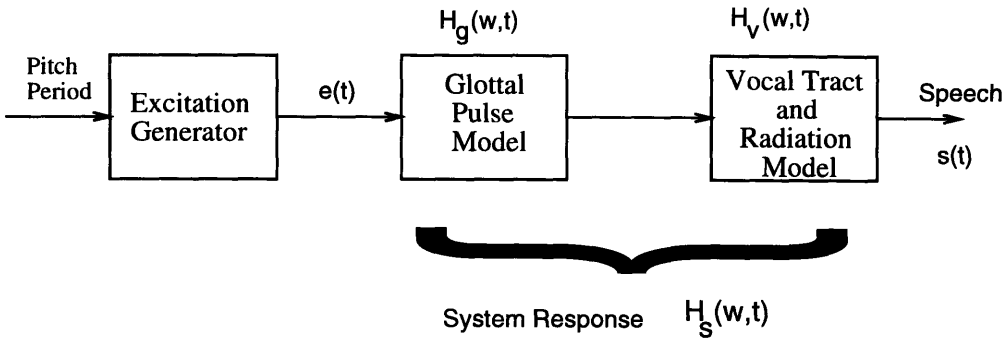


Figure 2-1: Block diagram of the acoustic speech production model. e(t) is a series of impulses separated by a pitch period. It is then modified by the glottal pulse to produce periodic glottal pulses. $H_g(\omega, t)$ is low-pass in nature. The vocal tract and radiation are represented as linear filters which modify the glottal pulses.

quencies of the vocal tract, called formant frequencies or poles. In the ideal lossless case, the transfer function has infinite gain at the formant frequencies, but losses such as heat, viscosity, and compliance of the vocal tract, cause the formants to have finite amplitudes and bandwidths. Because the vocal tract is real and dissipative, the system response can only contain real or complex-conjugate-pair poles. In practice, the transfer function only contains complex-conjugate-pairs and may be written as:

$$T(s) = K \frac{s_1 s_1^*}{(s - s_1)(s - s_1^*)} \frac{s_2 s_2 *}{(s - s_2)(s - s_2^*)} \cdots \tag{2.1}$$

where $s_n = \sigma_n + j\omega_n$ are the complex frequencies of the poles. The equation shows that for frequencies greather than the pole, the system function magnitude falls off as $1/f$ and can be seen in the Figure 2-2. This $1/f$ rolloff reduces the peak magnitude of higher-frequency formants.

The all-pole model requires that the parameters vary slowly and are approximately constant over 10 to 20 ms [30]. This assumes that the vocal tract is unchanging over a short-time period, so during speech transitions, where the vocal tract changes rapidly, the all-pole model does not perform well. However, the all-pole model is a good representation for vowels, where the vocal tract is the same for a short duration.

The all-pole model also breaks down when the speech is nasal and or is a fricative. During nasalized speech another cavity is opened in the vocal tract, so an acoustic zero is introduced by the resonance of the the second airway. During fricatives, a turbulent source at the glottis can produce acoustic zeros in the system function.

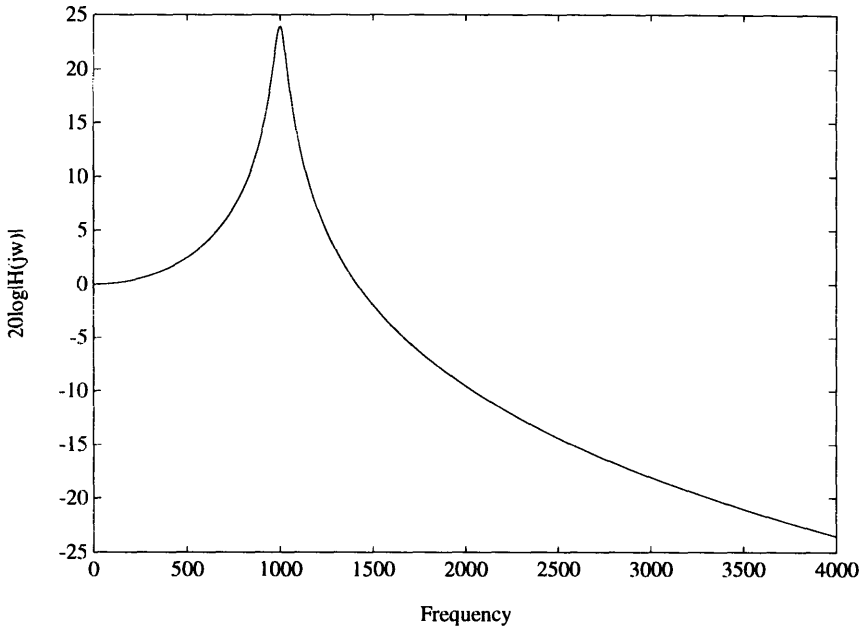


Figure 2-2: Single pole transfer function with a finite resonance bandwidth. Pole at 1000 Hz with bandwidth of 200 Hz.

These acoustic zeros can affect the shape of the transfer function in many ways. A zero can generate a formant-like structure close to the formant frequency, as shown in Figure 2-3, or the formant bandwidth can be decreased by a narrowing effect due to the zero. A zero occuring at a frequency almost identical to the formant frequency can either cancel the peak in the transfer function or decrease the amplitude of the peak.

The shape of the system function is important for speech synthesis. Errors in reconstructing the system function could have the same affect as additional zeros and the synthesized speech might artifically sound nasalized or like a fricative. When the formant bandwidth is incorrectly reconstructed, the synthesized speech may sound muffled. Psychoacoustical tests have shown that variations in the amplitude and the bandwidth of the formant peak can significantly affect speech perception and quality [11]. Appendix A shows the preliminary results of a psychoacoustic test in which the bandwidths of a formant in the utterance /a/ /b/ /a/ were varier and the just-noticeable differences were recorded. This test showed that untrained listeners were able to detect 4dB changes in the formant bandwidth, so the system function must be well represented for speech synthesis.

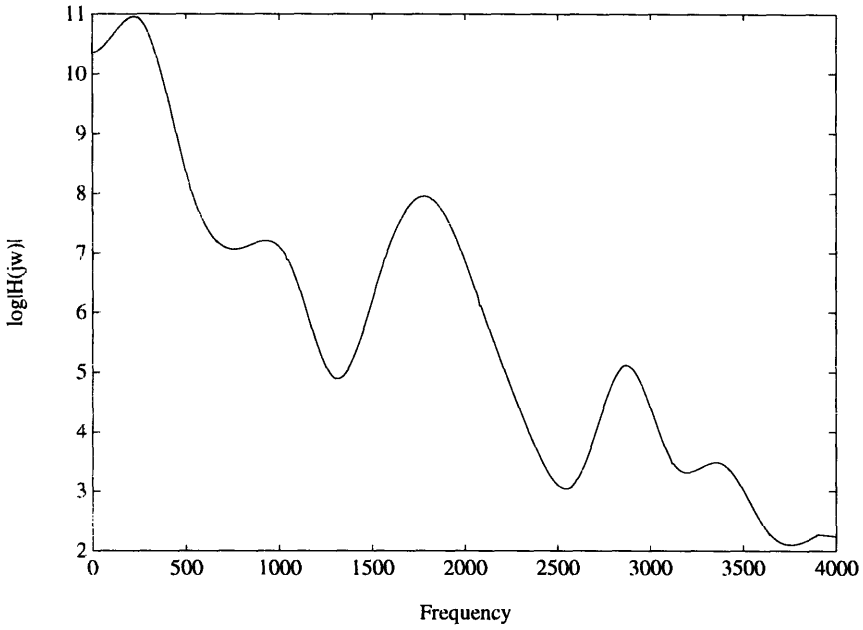


Figure 2-3: System transfer function of nasalized speech. The nasalization produces a second peak near 1000 Hz by a zero which occurs near 650 Hz.

2.2 Sinusoidal Model

The sinusoidal model is based on the speech production model shown in Figure 2-1 with the excitation being represented as a sum of sine-waves [22, 29] rather than pulses. During voiced speech, the excitation can be decomposed into harmonic sine-waves with predictable phase variations; during unvoiced speech the sine-wave components are aharmonic with random phases. When these sine-waves are filtered by the vocal cords and vocal tract, the speech, $s(n)$, is a linear combination of sine-waves with different amplitudes $\{A_l\}$ and phases $\{\phi_l\}$ [23]:

$$s(n) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \phi_l) \tag{2.2}$$

During speech analysis, the speech is broken up into frames by windowing the speech signal using a Hamming window. For each frame, the sine-wave peaks are obtained from the magnitude of the short-time Fourier transform (STFT), and the largest peaks are retained, but limited to a maximum of 80 peaks. The speech can be synthesized using the amplitudes and phases at the set of frequencies (ω_l).

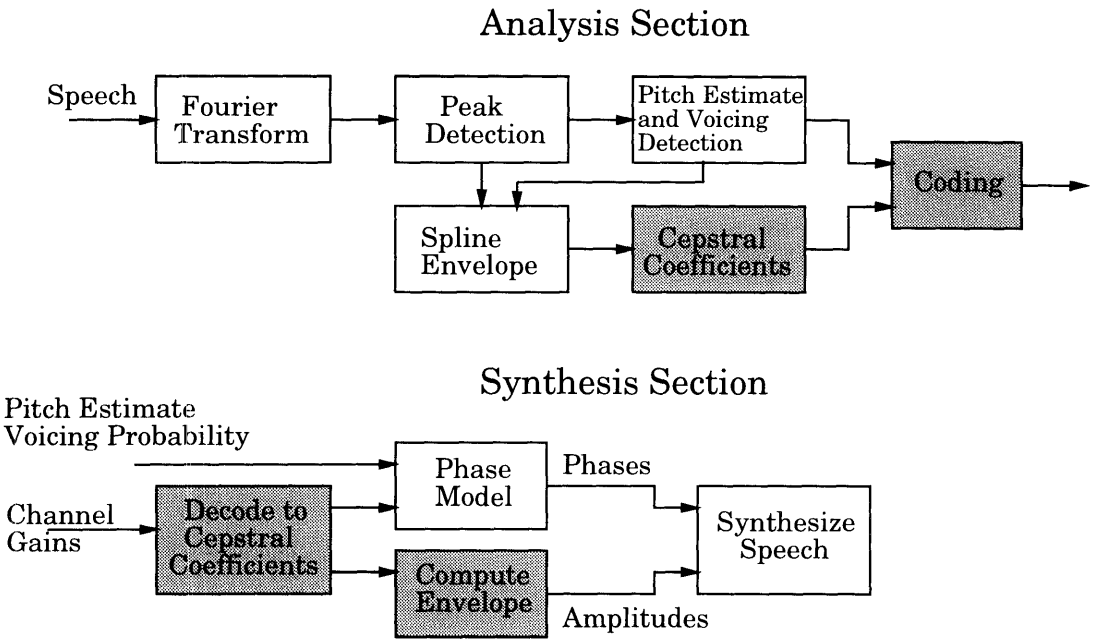


Figure 2-4: Block diagram of the major portions of the sinusoidal transform coder. After a STFT is obtained, a peak detection algorithm finds the peaks in the spectrum for each frame of speech. From the peaks, a smooth spectral envelope, which represents the vocal tract transfer function; pitch estimator; and voicing probability are obtained. The synthesis section decodes the coded signal and computes the spectral envelope, and reconstructs the speech signal. The shaded components are the ones that will be altered in this thesis.

2.2.1 Sinusoidal Transform Coder

Each speech frame requires a large number of parameters for the complete sine-wave representation of the transfer function and excitation. The number of sine-waves in each frame may vary as well, so this parameter set is not well suited for low-rate coding. The sinusoidal transform coder (STC) simplifies the sine-wave representation to a small set of amplitudes, a pitch and a voicing probability [23]. The STC system, shown in Figure 2-4, is briefly described, with emphasis placed on portions of the coder relevant to this thesis.

During voiced speech, the sine-waves frequencies will have harmonic spacing and the set of frequencies will be multiples of the first harmonic. A pitch estimator ($\hat{\omega_o}$) which estimates the spacing between the frequencies reduces the set of sine-wave frequencies to a single parameter for both voiced and unvoiced speech. Speech can be synthesized by the following equation where θ is a minimum phase function, and ϕ_k is the phase onset [23].

$$\hat{s}(n) = \sum_{k=1}^{k\hat{\omega_o}} A(k\hat{\omega_o}) \cos(nk\hat{\omega_o} + \phi_k + \theta(k\hat{\omega_o})) \tag{2.3}$$

For the purposes of this thesis it is not important how the onset is obtained, but that it is used to assist in aligning the speech frames when synthesizing speech. A residual phase function is computed as the difference between the phases of the synthesized speech using the estimate of the onset and the harmonic spacing. During voiced speech the residual is nearly zero, and it is random for unvoiced speech [23].

Cepstral coefficients are obtained by taking the inverse Fourier transform of the log magnitude of the Fourier transform. Since the speech signal is assumed to be minimum phase, the system function, amplitude envelope, and phase can be gotten directly from the cepstral coefficients.

$$\log A_s(\omega) = c_0 + 2 \sum_{m=1}^{\infty} c_m \cos(m\omega) \tag{2.4}$$

$$\Phi_s(\omega) = -2 \sum_{m=1}^{\infty} c_m \sin(m\omega) \tag{2.5}$$

The phase of the vocal tract, without the sign or temporal onset, can be recovered from the cepstral coefficients, which are derived from the log magnitude function.

The set of phases is reduced to a voicing probability which during synthesis forces the residual phase to zero, for voiced speech, or a random signal, for unvoiced speech. The voicing probability can be computed from the residual phase generated in the analysis section, or it can be determined from the mean squared error between the signal ($s(n)$) and the estimated signal ($\hat{s(n)}$) in the harmonic model [23] because the mean squared error is larger for unvoiced fames due to the aharmonic spacing of sine-waves.

Unvoiced speech is represented by a large number of sine-waves, but the STC limits the number of sine-waves to a harmonic set, so it is not represented well by the harmonic model. Estimates of the fundamental frequency larger than 150 Hz for unvoiced speech can produce perceptual artifacts due to too few sine-waves representing a noise-like signal [23]. For unvoiced speech with estimated fundamental frequencies larger than 100 Hz, the fundamental frequency is defaulted to 100 Hz, hence satisfying the Karhunen-Loève expansion.

The sine-wave amplitudes are converted into a continuous function using an algorithm from the Spectral Envelope Estimation Vocoder (SEEVOC) which finds the largest sine-wave amplitude within each pitch interval $[\frac{\omega}{2}, \frac{3\omega}{2}]$. A cubic spline interpolation between the largest peaks in each envelope form a smooth envelope which is similar to the system transfer function shown in Figure 2-3

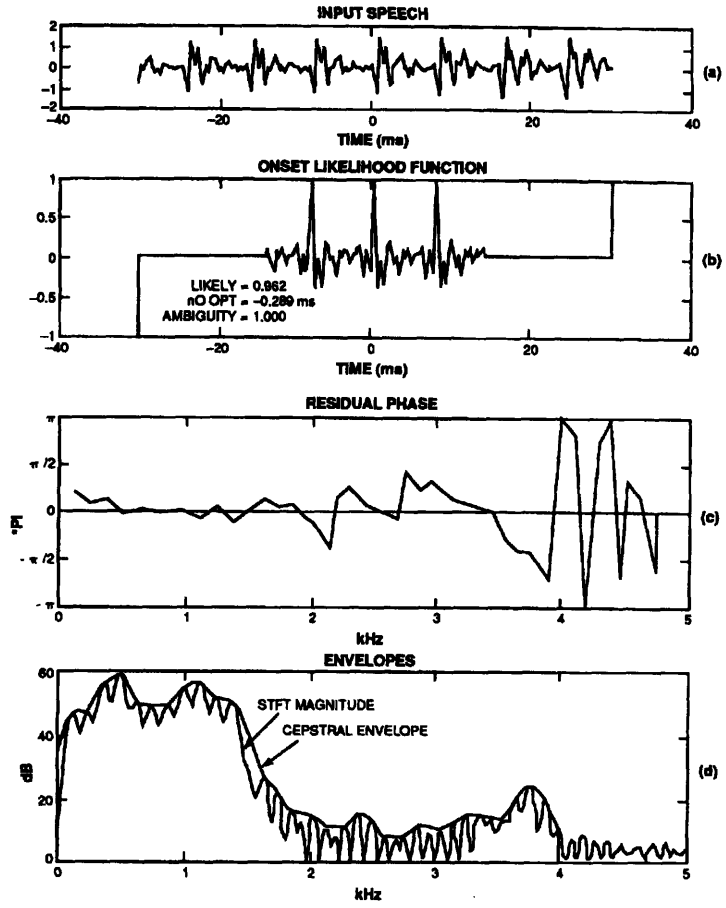


Figure 2-5: Onset estimation of a frame of speech. (a) a typical frame of voiced speech with the pitch pulse occuring near the center of the frame. (b) onset estimation. (c) Residual phase. (d) Spectral domain representations of the STFT and cepstral envelope [23].

The construction and representation of the spline envelope is crucial for synthesizing the coded speech because the amplitude and phase information are contained in the envelope. As mentioned in Section 2.1, alterations to this envelope can affect the quality of the speech since it represents the speech transfer function.

The envelope consist of 512 points, representing the speech spectrum up to 4kHz. The large number of points makes it a poor representation for low-rate coding, so a representation for the envelope is needed to reduce it to a smaller set of coefficients.

2.2.2 Coding the Spline Envelope

The spline envelope can be coded in many ways, but since the cepstral coefficients are needed for determining phase, the STC system uses a cepstral-based coding algorithm [23]. The

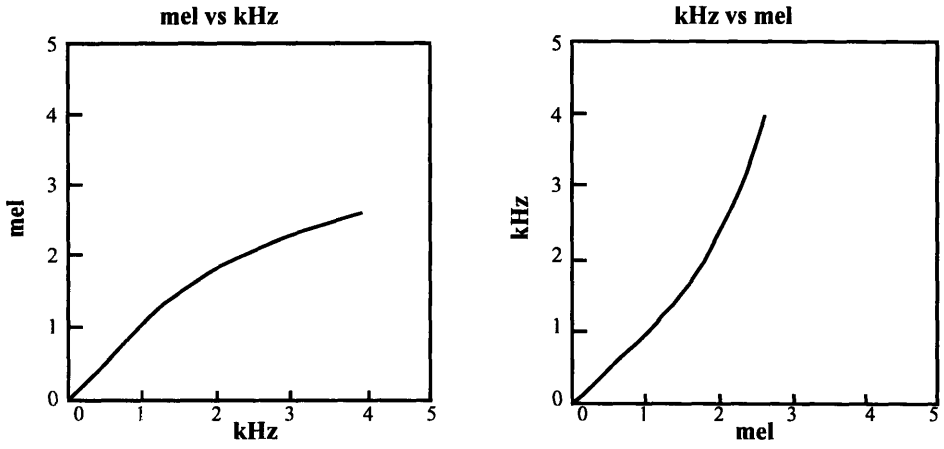


Figure 2-6: Mel warping function [23].

number of cepstral coefficients retained is a design parameter used to determine the rate of the vocoder. Experimentally it was shown that retaining more than 40 coefficients does not improve the quality of the speech, so the number of points for the representation is always less than 40 depending on the vocoder rate [23]. Truncating the cepstrum to M points ($M \leq 40$) has the effect of low-pass filtering the spline envelope.

Recordings from the auditory nerve as well as other information indicate that the cochlea is highly frequency-selective bank of filters which is not as frequency-selective at high frequencies. [28] The representation for the high-frequency bands can made coarser than the low-frequency bands by allocating fewer coding bits to the coefficients representing the high-frequency range. The frequency axis of the STFT is warped (shown in Figure 2-6) to exploit the properties of the auditory system.

The reduced number of cepstral coefficients are transformed back into the frequency domain because the large dynamic range of the cepstral coefficients makes them unfavorable for low-rate coding. A new smoother envelope is sampled at linearly spaced intervals known as channel gains. The mel-warping causes the gains to be non-linearly spaced on the original frequency axis as shown in Figure 2-7.

The channel gains are spectral amplitudes at a set of frequencies, and further reduction in coding can be achieved by quantizing them according to the perceptual properties of the ear [23]. More bits are given for the channels in the lower frequency range than in the higher frequency range. The channel gains are quantized and coded using delta pulse code modulation (DPCM) [17, 27].

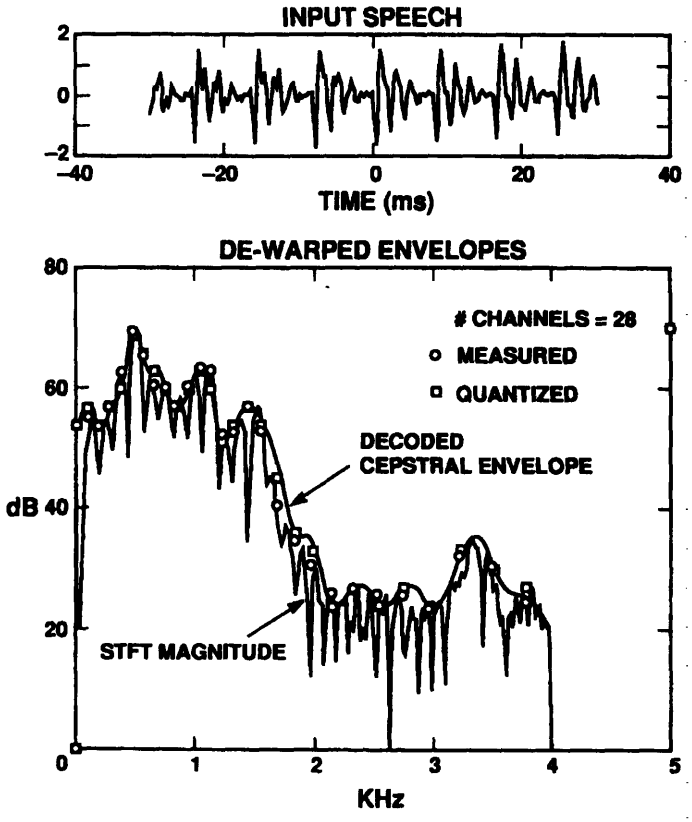


Figure 2-7: Channel gains computed from the cepstral envelope [23].

In order to decrease the data rate, a frame-fill interpolation method is used, in addition to the DPCM, so that channel gains are sent every other frame. The mid-frames are interpolated from the two adjacent frames. Two bits are used to tell the synthesis section whether the odd frame resembles the kth frame, the $k+2$nd frame, or a combination of the two frames.

2.2.3 STC Coding Rates

The STC system is able to code speech at data rates of 2400 and 4800 bits per second using just the spline envelope, pitch, and voicing probability, and frame interpolation bits. Since the pitch and voicing probability functions are inherent to the sine-wave system, they are not altered in the vocoder developed in this thesis.

Twelve bits per frame are used for coding non-envelope information. The pitch is coded using 7 bits, the voicing probability with 3 bits, and the frame-fill interpolation with 2 bits. If the frame interval is 20 ms long and frame-fill is used, 25 frames per second need to be coded. To achieve 4800 bits per second for 25 frames per second, 192 bits for frame must

code all information. This leaves 180 bits per frame for coding the envelope.

2.3 Wavelet STC System

This thesis describes the performance and development of the wavelet-based sinusoidal transform coder (WSTC). The front-end of the WSTC system is identical to the STC system. Speech is decomposed into the spline envelope, pitch estimate, and voicing probability, as shown in Figure 2-4. The main focus will be on the development of a wavelet representation for the sine-wave amplitudes.

Instead of using the cepstral coefficients to determine the channel gains, the WSTC performs a wavelet transform on the spline envelope and a set of the wavelet coefficients are quantized. Since the spline envelope is real and even, the log of the spline envelope can be shown to have a $1/f$ property (see Appendix B). $1/f$ processes are well represented by a wavelet decomposition [37] and the wavelet coefficients have the advantage of being uncorrelated (shown in the next chapter). An inverse wavelet transform of the quantized wavelet coefficients is performed at the synthesis portion to obtain the spline envelope. From the envelope, speech is synthesized, as it is in the STC system.

Chapter 3

Wavelet Theory

Development of the wavelet transform originated in seismic data analysis from the work of Morlet [25] in 1982. The development was motivated by the need for good low-frequency resolution and good time resolution of high-frequency bursts simultaneously. Original wavelet work decomposed signals into frequency bands to extract information from the different bands. Numerical analysis of wavelet coefficients led to signal decomposition using wavelet basis function which are best suited for decomposing the signal so that the resulting coefficients are easily compressed [7]. The number of coefficients needed to reconstruct the signal is minimized by using a set of carefully chosen basis functions.

Motivation for using the wavelet transform to represent the spline envelope came from work done by Wornell which suggests that the wavelet transform is supposed to be optimal and robust for representing $1/f$ signals, such as the spline envelope. The spline envelope is a smooth version of the log magnitude of the speech spectrum, and therefore its spectrum contains mostly low frequencies and is $1/f$. Wornell showed that the variance of the wavelet coefficients will be related to the order of $1/f$ or nearly $1/f$ signals and that as the scale increases, the variance of the coefficients increases [37]. Therefore, WSTC coefficients should be characterized by a set of variance parameters leading to the extraction of information about the speech signal.

While no new theory is developed in this chapter, basic wavelet theory and subband coding theory, upon which many of the results are based, is reviewed. Wavelet notation used throughout the thesis is introduced. A more rigorous mathematical treatment of wavelet theory can be found in Chui [5] and Daubechies [9]. Other tutorials can be found in Mallat [19]

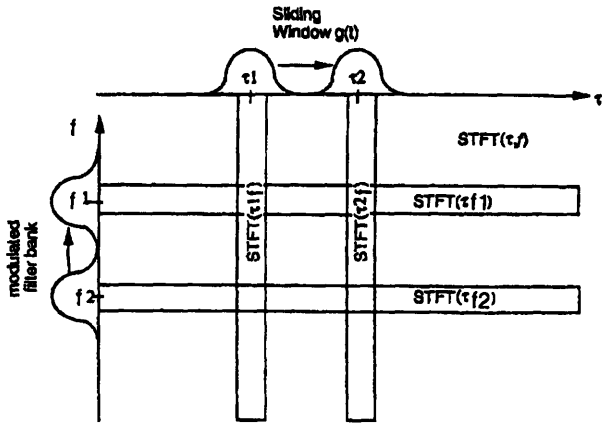


Figure 3-1: Time-Frequency plane for the STFT. Can be viewed as either a FT at time τ, or a set of modulated filters. [31]

and Vetterli [31]. A practical algorithm for the wavelet transform is presented which is similar to subband and multirate systems used to code images and speech [30, 33, 3]. The properties of the wavelet basis functions, and the wavelet transform are explained to provide the background for the algorithms used in the coder developed in this thesis. Finally, the wavelet representation for the $1/f$ signal shown by Wornell and Oppenheim will be reviewed.

3.1 Continuous Wavelet Transform

The Short Time Fourier Transform (STFT) is commonly used to analyze non-stationary signals. The STFT operation can be viewed as a two-step process: first the signal is windowed in time to produce a temporal localization, and then the Fourier transform is computed.

The STFT is given by,

$$STFT(\tau, f) = \int x(t)g^*(t-\tau)e^{-2j\pi ft}dt \tag{3.1}$$

The signal $x(t)$ is windowed by a function $g(t)$ at some time τ, and the Fourier transform of the windowed signal represents the STFT.

Alternatively, the STFT can be viewed as a bank of modulated filters, in which the signal is convolved with the modulated window $g^*(t-\tau)e^{-2j\pi ft}$. Figure 3-1 shows the two interpretations of the STFT. The STFT maps the signal into a two-dimensional time-frequency space [31].

The width of the window determines both the time and frequency resolutions. The time-

frequency trade-off for the STFT is governed by the uncertainty principle, which states that the time and frequency localizations of the decomposition are related by

$$\tau_o f_o \geq \frac{1}{4\pi} \tag{3.2}$$

where f_o and τ_o are the frequency and time resolutions. Time resolution is sacrificed for frequency localization. Shorter STFT windows produce better time resolution, but the frequency localization is impaired. For signals which contain both low-frequency component and high-frequency bursts, it is desirable to analyze the signal with both short and long windows.

The wavelet transform (WT) uses scaled basis functions to produce good temporal and spectral localizations, but not both simultaneously, by changing the STFT constant bandwidth filters to *constant* Q bandpass filters with $\Delta f/f$ as a constant (frequency band over the center frequency). The uncertainty principle (Equation 3.2) always holds, so at high frequencies where the window is short, the system produces better time localization and at the longer windows frequency resolution improves [31].

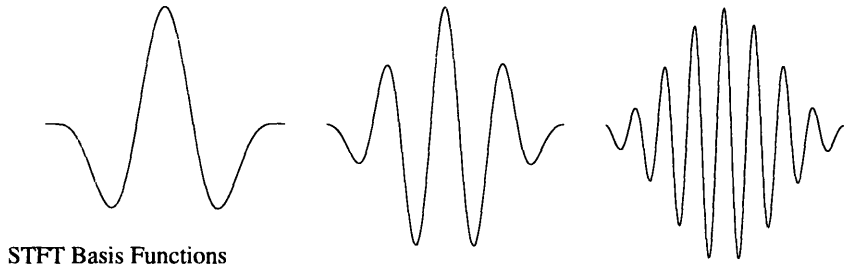
The continuous wavelet transform (CWT) uses frequency-scaled and time-shifted versions of a basic function called the *mother wavelet* $\psi(t)$. Unlike the basis functions of the STFT, which are modulated versions of the same windowing function, the basis functions of the continuous wavelet transform, shown in Figure 3-2, can be written as:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}) \tag{3.3}$$

$$CWT_x(a,b) = \mathcal{W}\{x(t)\} = \frac{1}{\sqrt{a}}\int x(t)\psi^*\left(\frac{t-b}{a}\right)dt \tag{3.4}$$

The CWT, Equation 3.4, is an inner product of the signal with shifts and dilation of a mother wavelet. Shifted versions of the wavelet basis needed to cover the time domain from $t = -\infty$ to $+\infty$ and the dilations allow for complete coverage of the frequency domain. Wavelets were explored primarily for their localization properties, so it is desirable, but not required, for the mother wavelet to have compact support, meaning that for $|x| > x_o$ the wavelet is zero.

$$\int_{-\infty}^{+\infty} |\psi(x)|dx < \infty \tag{3.5}$$



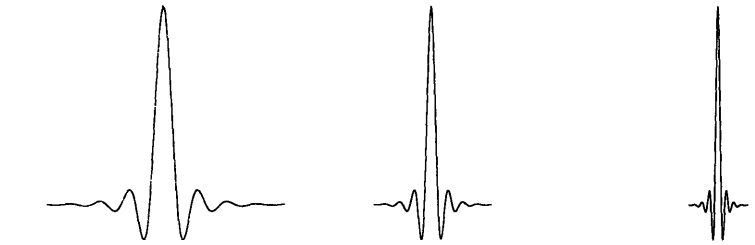


Figure 3-2: Modulated basis functions for the STFT are modulated exponentials. The wavelet basis functions are scaled versions of a mother wavelet. The mother wavelet shown here is a Battle-Lemairé spline wavelet.

An inverse synthesis formula must exist for this transform to be useful. The reconstruction formula imposes restrictions on ψ. A necessary and sufficient condition to reconstruct $x(t)$ from the wavelet transform (as shown by Grossman et. al. [15]) is that the wavelet $\psi(t)$ must satisfy the admissibility condition,

$$C_\psi = \int_{-\infty}^{+\infty} \frac{\left|\hat{\psi}(\omega)\right|^2}{|\omega|} d\omega < \infty \tag{3.6}$$

where $\hat{\psi}(\omega)$ is the Fourier transform of the mother wavelet.

Provided that $\hat{\psi}(\omega)$ is continuous, and $\psi(t)$ has reasonable decay (at least as fast as $|t|^{-1-\epsilon}$), then the admissibility requirement is equivalent to

$$\hat{\psi}(0) = \int_{-\infty}^{+\infty} \psi(t) dt = 0 \tag{3.7}$$

and for all practical purposes this is considered to be the admissibility condition [9].

The synthesis equation for the wavelet transform

$$x(t) = \mathcal{W}^{-1}\{CWT_x(a,b)\} = \frac{1}{C_\psi}\int_{-\infty}^{+\infty}\frac{da}{a^2}\int_{-\infty}^{+\infty} CWT_x(a,b)\psi_{a,b}db \tag{3.8}$$

holds for all admissible wavelets.

3.2 Discrete Wavelet Transform

The CWT is highly redundant and not practical to implement since the shifts and dilations are continuous functions. The shift-scale plane is discretized as

$$a_j = a_o^j; b_k = ka_o^j b_o \tag{3.9}$$

to minimize redundancy in the representation. This discretization plane is chosen because of the dilation property associated with the wavelet transform. The wavelet decomposition is simplified by using dyadic sampling with $a_o = 2$ and $b_o = 1$ and the wavelet basis functions and coefficients for this lattice are

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k); \tag{3.10}$$

$$x_{j,k} = \int_{-\infty}^{+\infty} x(t)\psi_{j,k}^*(t)dt. \tag{3.11}$$

Daubechies showed that the sampling lattice of the WT must be correctly chosen for a complete and stable reconstruction using the following synthesis equation [8, 9].

$$x(t) = \sum_{j,k} x_{j,k}\psi_{j,k}(t) \tag{3.12}$$

3.2.1 Multiresolution Analysis

Even though the scale and shift parameters are discrete, the basis functions and the signal are still continuous functions, making the DWT impractical since the speech signal and most other signals are discretized by an A/D system.

Mallat [19] developed the multiresolution signal analysis for wavelets which provides a practical way of applying the wavelet transform to discrete signals. The multiresolution decomposition projects the signal on to successive subspaces which are coarser representa-

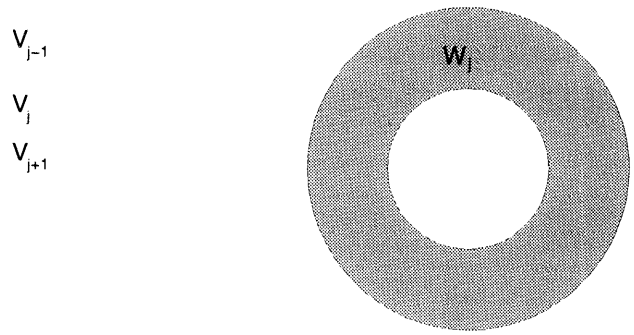


Figure 3-3: Graphical representation of the subspaces in the multiresolution analysis for the wavelet transform. The transform consists of nested approximation subspaces, and orthogonal complement subspace.

tions of the previous subspace. The projections are obtained by smoothing the signal, which produces a coarser temporal resolution. The multiresolution provides a method for looking at the wavelet transform as subspace and basis decomposition as opposed to frequency decomposition.

Multiresolution analysis leads to shifted and dilated functions that form a complete orthonormal basis for $L^2()$. The $L^2()$ space is composed of a sequence of successive approximation spaces V_j such that the subspaces are nested

$$\ldots V_{j+1} \subset V_j \subset V_{j-1} \subset \ldots \tag{3.13}$$

As j is increased, the resolution becomes coarser. [1] A coarser representation has signals which are smoother, or blurred from the finer representation. The subspaces must also satisfy

$$\overline{\bigcup_{j\in} V_j} = L^2() \tag{3.14}$$

$$\bigcap_{j\in} V_j = \{0\} \tag{3.15}$$

so that the subspaces define the entire range of $L^2()$. A graphical representation of the subspaces are shown in Figure 3-3.

Sets of non-orthogonal spaces can be generated from the previous constraints. To impose the orthonormality condition of the multiresolution analysis, scale-invariance and translation-

[1] The sign of j varies between papers. Some authors use a negative j corresponding to a coarser resolution.

invariance properties are needed. The scaling property (3.16) allows an approximation at a coarser resolution to be derived from an approximation at any other finer subspace. The translation constraint (3.17) guarantees that if the original signal is shifted by n, then the approximation to the signal is shifted by a proportional amount. These constraints (3.13 - 3.17) form orthonormal dyadic subspaces.

$$x(t) = V_j \Longleftrightarrow x(2t) = V_{j-1} \tag{3.16}$$

$$x(t) \in V_j \Longleftrightarrow x(t - 2^j n) \in V_j \tag{3.17}$$

A linear operator A_j, known as the approximation operator, defines the projection of the signal to the subspace at resolution at j. The approximation of the signal ($A_j x(t)$) approaches the signal at infinitely small resolutions, and converges to zero (since there is no DC $L^2()$) as the approximation gets coarser. This can be thought of looking at a signal through a pair of glasses. As $j \to \infty$, the glasses are very much out of focus, and everything is blurred. As $j \to -\infty$, the glasses come into focus with the infinite resolution.

$$\lim_{j \to -\infty} A_j x(t) = x(t) \tag{3.18}$$

$$\lim_{j \to \infty} A_j x(t) = 0 \tag{3.19}$$

Mallat [19] showed that there exists a *scaling function* $\phi(x) \in L^2()$ such that the dilated and shifted $(2^j \phi(2^j x - n)_{n\epsilon})$ functions form an orthonormal basis of V_j which projects the signal to the different subspaces.

$$A_j x(t) = \sum_n a_{j,n} \phi_{j,n}(t) \tag{3.20}$$

$$a_{j,n} = \int_{-\infty}^{\infty} x(t) \phi_{j,n}(t) dt \tag{3.21}$$

where the coefficients $a_{j,n}$ are the projections.

The scaling function $\phi(t)$ has a low-pass frequency response because it projects from a fine resolution to a coarser one [19]. Equation 3.21 can be represented as a filtering operation followed by downsampling

$$a_{j,n} = (x(t) * \phi_{j,0}) \mid_{t=2^j n} \tag{3.22}$$

The difference between two adjacent approximations, $a_{j,n}$ and $a_{j-1,n}$, is called the detail

signal, which can be downsampled as well if the detail basis functions exhibit the same dyadic constraints. The collection of approximation subspaces contain redundant information, but the detail signals contain different information at each scale or subspace and can be thought of as frequency bands. The detail signal resides in the orthogonal complement to V_j, W_j (see Figure 3-3), such that the following are satisfied

$$W_j \perp V_j \tag{3.23}$$

$$W_j \oplus V_j = V_{j-1} \tag{3.24}$$

$$W_j \perp W_{j'}, j \neq j'. \tag{3.25}$$

There exists a *wavelet* basis function $\psi(t) \in L^2()$ whose shifted versions form an orthonormal basis for W_j. The projection operator from V to W is

$$D_j x(t) = \sum_n x_{j,n} \psi_{j,n}(t) \tag{3.26}$$

where the coefficients $x_{j,n}$ are the projections

$$x_{j,n} = \int_{\infty}^{\infty} x(t) \psi_{j,n}(t) dt. \tag{3.27}$$

The wavelet basis function must have a bandpass spectrum because it contains the information found between two subspaces V_j and V_{j-1}. Any subspace V_j can be recursively decomposed into a sum of orthogonal subspaces [37]

$$V_J = W_{J+1} \oplus V_{J+1} = W_{J+1} \oplus (W_{J+2} \oplus V_{J+2}) = \bigoplus_{j>J} W_j \tag{3.28}$$

such that

$$V = \bigoplus_{j=\infty}^{-\infty} W_j.$$

Analog signals have infinite resolution, and are located in the space $j = -\infty$, but it is impossible to start a decomposition at $j = -\infty$, so the sampled signal after the A/D step is taken to be at the 0th subspace.

$$a_{0,n} = \int_{-\infty}^{\infty} x(t) \phi_{0,n}(t) dt$$

$$A_0 x(t) = \sum_n a_{0,n} \phi_{0,n}(t)$$

The projection step is low-pass in nature because of the scaling function. This is consistent with the need for an anti-aliasing low-pass filter within the A/D stage. This is an approximation that is necessary for implementing the multiresolution analysis.

3.2.2 Cascade Filter Implementation

The key feature of the multiresolution analysis is that any subspace can be contrived from a finer resolution subspace. V is included in V_{-1}, and $\phi_{-1,n}$ is an orthonormal basis in V_{-1}. ϕ, which is an orthonormal basis in V, is constructed as a linear combination of $\phi_{-1,n}$, with h_n as the Fourier coefficients between the two spaces.

$$\phi = \sum_n h_n \phi_{-1,n} \tag{3.29}$$

$$h_n = < \phi, \phi_{-1,n} > \tag{3.30}$$

The detail space W_j is derived from the approximation space V_{j-1}, so the wavelet basis function at the nth subspace is defined as:

$$\psi = \sum_n g_n \phi_{-1,n} \tag{3.31}$$

$$g_n = < \psi, \phi_{-1,n} > . \tag{3.32}$$

The approximation of a signal at a resolution is obtained by iteratively taking approximations from one space to its subspace (finer to coarser subspaces). The procedure is implemented as a filter-and-downsample (3.33).The detail coefficients are also obtained by a filter-and-sample procedure (3.34).

$$a_{j,n} = \sum_l h(l - 2n) a_{j-1,n} \tag{3.33}$$

$$x_{j,n} = \sum_l g(l - 2n) a_{j-1,l} \tag{3.34}$$

The combination of Equations 3.33 and 3.34 is known as the pyramidal algorithm. Figure 3-4 shows the analysis and synthesis steps required to go from one scale to an adjacent one. Inverse filters of h_n and g_n are used in the synthesis stages. Filters with finite length provide an efficient method for computing the wavelet transform.

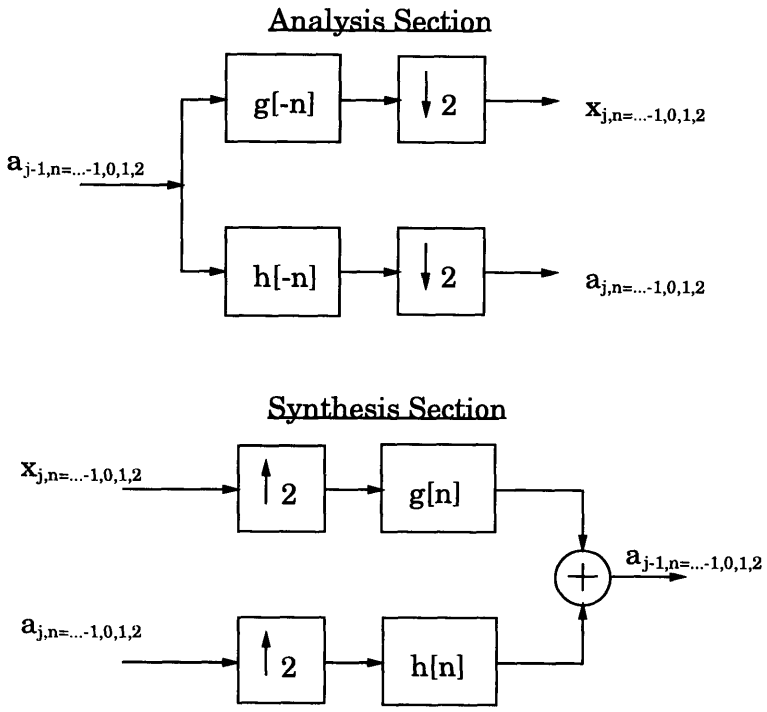


Figure 3-4: Single stage of the pyramidal analysis and synthesis stage of a dyadic discrete-wavelet transform. The analysis step consists of a filtering operation followed by a down-sampling. The synthesis is performed by upsampling the discrete representation and then filtering using the inverse filters of those used in the analysis section.

To satisfy the orthogonality constraint on the scaling and wavelet bases, Mallat [19] showed that the filters $h[n]$ and $g[n]$ must be be related by:

$$g[n] = (-1)^n h[1-n] \tag{3.35}$$

$$G(\omega) = e^{-jw} H^*(\omega + \pi) \tag{3.36}$$

A further constraint of orthonormality of the scaling functions $\phi_{j,n}$ requires

$$|H(0)|^2 = 1 \tag{3.37}$$

$$|H(\omega)|^2 + |H(\omega + \pi)|^2 = 1 \tag{3.38}$$

The filters $g[n]$ and $h[n]$ form quadrature mirror filters (QMF) pair, in which $g[n]$ is highpass and $h[n]$ is lowpass [19]. The wavelet transform is defined entirely through $h[n]$

since the scaling function is constructed by dilations of the $h[n]$ such that

$$\hat{\phi}(\omega) = \prod_{p=1}^{\infty} H(2^{-p}\omega) \tag{3.39}$$

and the detail coefficients are obtained using Equation 3.36 to get the corresponding highpass filter. The QMFs provides perfect reconstruction of the signal from their coefficients.

The wavelet transform is similar to using exact reconstruction QMFs in subband systems [35]. All orthonormal wavelet bases that are used in the multiresolution analysis produce QMFs, but not all QMFs satisfy the orthonormality condition [9]. The multiresolution analysis of the wavelet transform is therefore, a subset of the subband system. Wavelet research has produced methods for the construction of different types of wavelet bases and QMFs which are suited for certain signals [9].

3.2.3 Finite Length Signals

Each subspace is dyadically downsampled, so the maximum depth of decomposition depends on the number of points in the signal at the 0th scale. A signal which contains $N_o 2^M$ samples can be decomposed to x_n^j such that

$$j = 1, 2, \ldots, M$$

$$n = 0, 1, \ldots, N_o 2^{2j-1} - 1$$

with a maximum depth of decomposition of M and at scale M there are N_o wavelet coefficients. For example, if the signal is 256 points, then the maximum depth of decomposition is 8 at which scale there will be only one coefficient. A single coefficient at the maximum depth of decomposition scale usually corresponds to the DC value which should be zero for $L^2()$.

An implementation problem arises because the Fourier coefficients h_n used to transform between subspaces remains the same size at each subspace, but the size of the signal is being downsampled. The signal will be downsampled to the point where the number of approximation coefficients is smaller than the number of Fourier coefficient. When this happens, the wavelet coefficients are mapping energy outside the edges of the signal. When the signal is scaled such that it is smaller than the filter width, finite duration problems discussed in Section 4.1 are encountered.

3.3 Wavelet Functions

Wavelet bases have been constructed for different types of signals and applications. Choosing a basis function that better represents the signal reduces the number of non-zero coefficients, decreases numerical complexity, and reduced perceptual errors in the synthesized speech. Two types of wavelet basis function are examined in this thesis. Compactly supported wavelets reduce the amount of numerical computation in the wavelet algorithm by having a small and finite set of coefficients while symmetric filters improve signal reconstruction and reduce quantization noise at the boundaries of the signal.

3.3.1 Compactly Supported Wavelets

Daubechies [8] developed a method for constructing compactly supported orthonormal wavelet bases which have good frequency properties and R vanishing points for increasing regularity. A wavelet basis function with Rth-order regularity is constructed by choosing $H(\omega)$ to have R zeros or vanishing points at $\omega = \pi$ [37]. These wavelet bases and the corresponding scaling functions are shown in Figure 3-5. Daubechies' compactly supported orthonormal wavelet bases are not symmetric like the Meyer and Battle-Lemarié wavelets and it has been shown by Daubechies [9], and Smith and Barnwell [32] that exact reconstructing QMF filters cannot be symmetric about the origin if the synthesis filters are the inverse of the analysis filters.

The compact support allows for good time localizations and reduces additional coefficients produced to represent the edges when the wavelet algorithm is implemented through convolution. The convolution produces $n-1$ additional coefficients for wavelet functions with length n.

3.3.2 Symmetric Filters

Filters with symmetric coefficients are called linear phase if the phase of the function is a linear function of the frequency. The Harr wavelet, which is a piecewise spline,

$$\phi(x) = \begin{cases} 1 & 0 \leq x \leq 1, \\ 0 & \text{otherwise}, \end{cases}$$

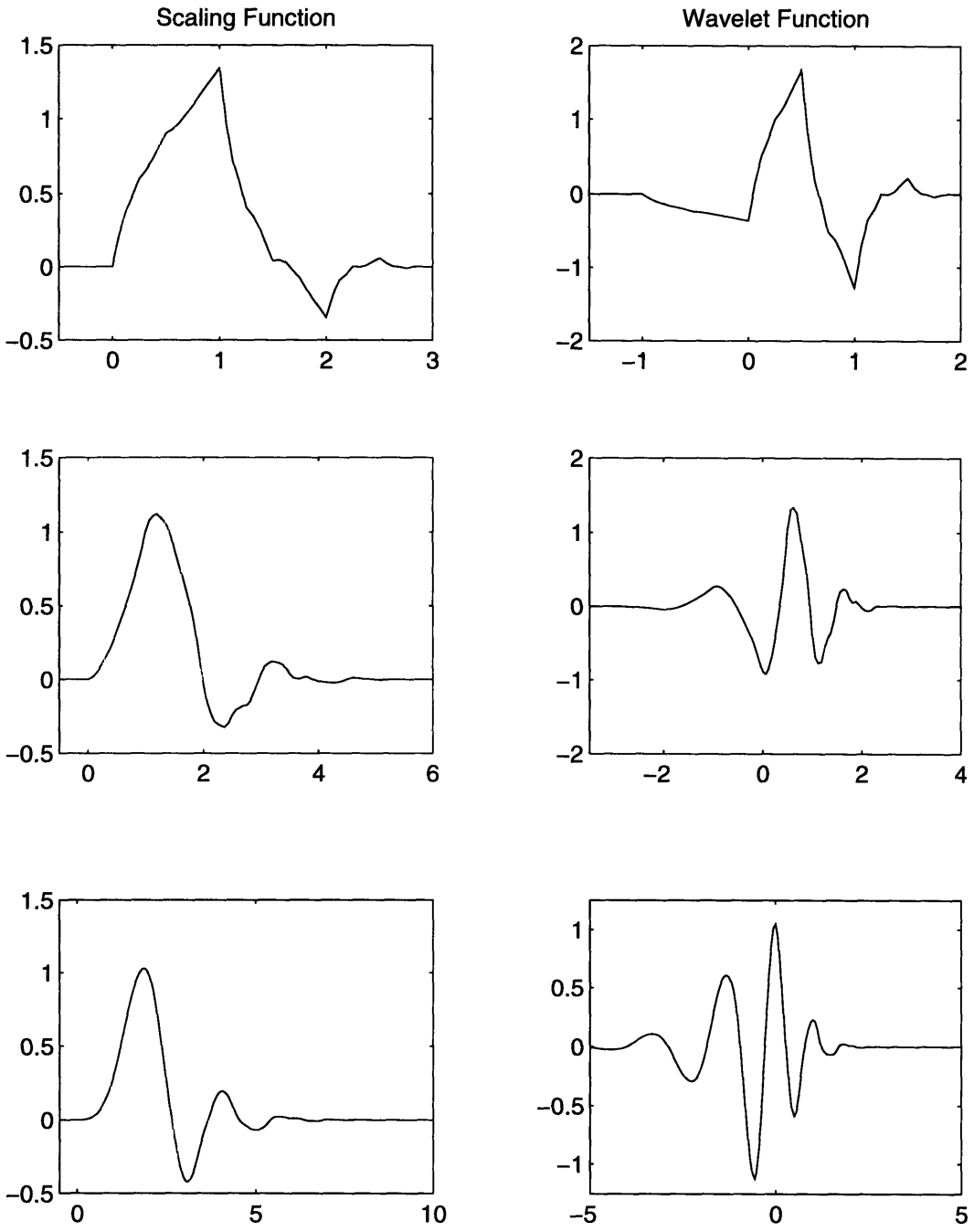


Figure 3-5: Daubechies' compactly supported scaling functions and wavelets with regularity R. (top) R=2, (middle) R=4, and (bottom) R=7.

$$\psi(x) = \begin{cases} 1 & 0 \leq x \leq \frac{1}{2}, \\ -1 & \frac{1}{2} \leq x \leq 1, \\ 0 & \text{otherwise}, \end{cases}$$

is not linear phase. Even though it is symmetric about $\frac{1}{2}$, the phase is discontinuous at π, so it is not linear phase under the current definition. The Harr filter is included as being linear phase if the definition is extended to include filters for which the phase is piecewise linear, with constant slope, and has discontinuities only where the function is 0 [9].

A class of spline scaling functions generated by Battle and Lemarié, were used in Mallat's development of the multiresolution analysis approach to wavelets [19]. Piecewise B-spline functions generate a multiresolution analysis, but they are not orthogonal to their translates [5] and can be orthogonalized by [9]

$$\hat{\phi}^{\#}(\omega) = \hat{\phi}(\omega)[\sum_{k=-\infty}^{\infty} |\hat{\phi}(\omega + 2\pi k)|^2]^{\frac{1}{2}} \tag{3.40}$$

Spline scaling functions are generated by recursive convolution of the Harr scaling function where the Nth-order scaling function is

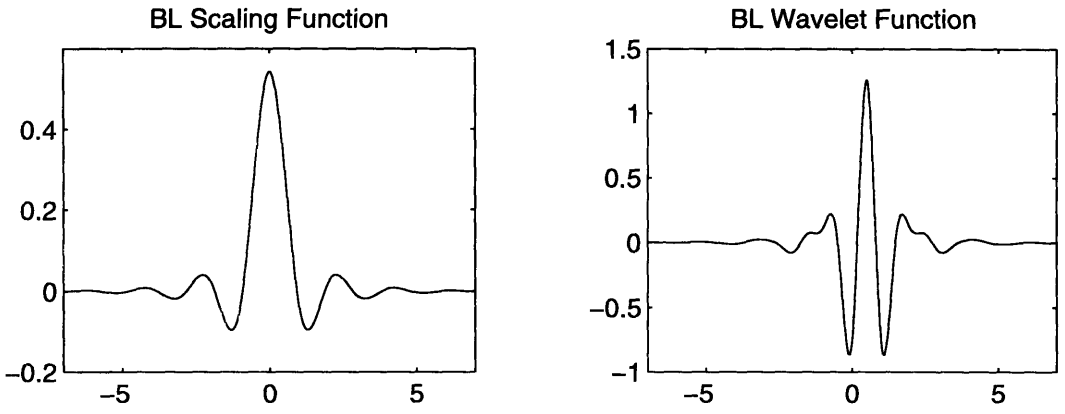
$$\hat{\phi}(\omega) = e^{-j\frac{\omega N}{2}} \left(\frac{\sin(\frac{\omega}{2})}{\frac{\omega}{2}} \right)^N \tag{3.41}$$

and the scaling function is then orthogonalized using Equation 3.40. The symmetry of the orthogonalized function ($\phi^{\#}$) is the same as the symmetry of the original spline function. For odd order splines, the functions are symmetric about $\frac{1}{2}$, and even order spline are symmetric about the origin. The orthogonalized spline functions no longer have finite duration, but their decay is exponential and is faster for larger N [9].

The filters $h[n]$ and $g[n]$ in Equations 3.33 and 3.34 can be obtained by using

$$\hat{\phi}(2\omega) = H(\omega)\hat{\phi}(\omega) \tag{3.42}$$

and Equation 3.36. The filters, $h[n]$ and $g[n]$, have infinite duration, which is not good for implementing the wavelet transform, but because of the exponential decay, the filter can be truncated, making them no longer perfect reconstruction filters. If the truncation is done properly, however, the errors are small compared to the computation required to reduce the



n	h[n]	n	h[n]	n	h[n]
0	0.5417	6	-0.0121	12	0.0015
1	0.3068	7	-0.0127	13	0.0013
2	-0.0355	8	0.0061	14	-0.0008
3	-0.0778	9	0.0058	15	-0.0007
4	0.0227	10	-0.0031	16	0.0004
5	0.0297	11	-0.0027	17	0.0003

Figure 3-6: (Top) Cubic Battle-Lemarié scaling and wavelet function. (Bottom) Coefficients for the filter used for the cubic spline. The filter coefficients are symmetric about 0.

errors.

Mallat [19] used a Battle-Lemairé wavelet, which is a cubic spline wavelet basis, when developing the multiresolution analysis approach to wavelets. The filter coefficients for $h[n]$ and the scaling function and wavelet bases that are generated are shown in Figure 3-6.

The spline envelope is made using a cubic spline interpolation between the peaks of the periodogram. The spline wavelet is also generated using splines, so it may be well suited to represent the spline envelope. It can be hypothesized that the spline wavelet will perform better at reducing the number of coefficient than the compactly supported wavelets. The next chapter will investigate which wavelet basis function is better suited for the spline envelope.

3.4 Wavelets and $1/f$ Processes

Wornell showed that for $1/f$ and nearly $1/f$ processes, the wavelet coefficients follow a variance progression [37]. Appendix B shows that the spline envelope is like a $1/f$ signal. This suggests that the wavelet transform is well suited to represent the spline envelope and

provides a Karhunen-Loève-like expansion. The wavelet coefficients for the spline envelope should also follow a similar variance progression.

Nearly $1/f$ processes are defined as being bound by [37]

$$\frac{\sigma_L^2}{|f|^\gamma} \leq S_x(f) \leq 2\frac{\sigma_U^2}{|f|^\gamma} \tag{3.43}$$

Wornell showed that the wavelet coefficients $x_{j,n}$ follow the variance progression of the form

$$\text{Var } x_{j,n} = \sigma^2 2^{-\gamma j} \tag{3.44}$$

and the magnitude of the correlation between coefficients across scales and within scales decays according to

$$|x_{j,n;j',n'}| = \mathcal{O}(|2^{-j}n - 2^{-j'}n'|^{|2R-\gamma|}) \tag{3.45}$$

as

$$|2^{-j}n - 2^{-j'}n'| \to \infty$$

where $x_{j,n;j',n'}$ is the autocorrelation, j is the scale, n is the shift, and R is the regularity of the wavelet [37]. Wornell verified both the variance progression and weak correlation using several $1/f$ processes.

The result that the coefficients are uncorrelated or weakly correlated is very exciting for coding the spectral envelope. This means that the wavelet coefficients of the spline envelope will each represent a different part of the speech signal. Analysis of the coefficients could lead to an understanding of which parameters or features are important for maintaining high quality speech. The variance progression might allow for statistical coding of $1/f$ processes and may assist in coding of the spline envelope at lower rates.

Chapter 4

Wavelet Algorithm Development

Early wavelet papers [8, 9, 19, 31] developed wavelet theory and mentioned applications in which wavelets might be useful; but explicit algorithms were not presented, even though implementation problems were noted. Taking a wavelet transform of a finite duration signal has many problems associated with it, and there have been proposed solutions by other researchers, but a rigorous analysis of the proposed solutions has not been performed. This chapter examines several methods for implementing the wavelet transform of the spline envelope.

Two wavelet bases are examined to determine which is better suited for representing the spline envelope. For the purposes of this thesis, the better basis function is the one which reduces the number of coefficients needed to reconstruct the envelope and minimizes the error between the original and reconstructed envelopes. The wavelet coefficients generated by both basis functions are analyzed to see which coefficients are needed to minimize reconstruction errors. The result of this chapter should provide the wavelet algorithm used in the WSTC.

4.1 Convolution-Based Decomposition

The wavelet decomposition algorithm, based on multiresolution analysis, is a recursive decomposition of the approximation coefficients that are generated in each stage of the analysis section shown in Figure 3-4. The filters, $h[n]$ and $g[n]$, are convolved with the approximation signal at each scale. The synthesis algorithm is a recursive reconstruction using the final approximation and the wavelet coefficients. Code used to implement the analysis and synthesis portions are given in Appendix C.

The spline envelope is considered to be at the 0th scale and higher order scales, $j > 0$,

Scale	Ideal	Daubechies (D_2)	Mallat N=25
d_1	256	258	268
d_2	128	131	146
d_3	64	67	85
d_4	32	35	55
d_5	16	19	40
d_6	8	11	32
d_7	4	7	28
d_8	2	5	26
d_9	1	4	25
a_9	1	4	25
total	513	537	705

Table 4.1: Number of coefficients at the different wavelet scales (d_k) and the final approximation (d_9) scale. A decomposition of a 512-length signal was decomposed with a convolution-based algorithm using the Daubechies N=2 compactly supported wavelet and Mallat's spline wavelet N=25.

represent coarser resolutions of the envelope. At each decomposition, the signal gets coarser until all that it left is the DC offset. The maximum depth of decomposition, using dyadic wavelet analysis, is the 9th wavelet scale.

Boundary problems are apparent when a convolution-based algorithm is used. The convolution at each stage increases the number of coefficients by $m-1$, where m is the length of the filter. Table 4.1 shows the number of coefficients generated by the convolution-based decomposition for both the Daubechies and Mallat wavelets. The first column shows the number of coefficients generated by the ideal case in which the signal is dyadically downsampled at every scale. The maximum decomposition scale for the envelope is 9 scales because the original envelope is a 512-point discrete signal. The number of coefficients using the convolution algorithm is larger than in the original signal, which defeats the purpose of the transform. In the ideal case d_9 represents the DC value and in $L^2()$ the DC is zero.

The reconstruction properties for Daubechies' compactly supported wavelet bases, D_2 and D_4 and for the spline wavelet with truncation lengths of 31, 25, and 19, were examined for those algorithms. The first set of experiments, were done using all the coefficients to obtain perfect reconstruction of the envelope. The next experiment was done by truncating the extra coefficients to determine their importance reconstruction of the envelope.

Two parameters were tested for each basis function: final depth of decomposition (FDD) and the number of scales used (NSU) for reconstruction. The FDD is the scale at which

the final approximation is kept. For example, in the ideal case, a FDD of 4 corresponds to a representation with the wavelet coefficients in the first 4 scales and the approximation coefficients at scale 4. To reduce the number of coefficients passed between the analysis and synthesis portions, the wavelet coefficients corresponding to the first l scales are zeroed. The NSU is defined as the number of non-zero wavelet coefficient scales used for synthesis. The NSU helps to evaluate what information is contained in the lower-scale coefficients.

4.1.1 Non-Truncation Experiments

The same algorithm was used for both the Daubechies and Mallat spline basis functions. The initial tests were performed for a FDD to the ninth scale.

The first and last points in the spectral envelope are not zero, so a large edge is produced in the signal. Figure 4-1 shows a typical spectral envelope for reconstructed voiced speech comparing two different Daubechies basis functions. The NSU ranged from 0 to 5. When all the low scales are used for the reconstruction ($l = 0$), the reconstructed signal is identical to the original. As the first few wavelet coefficient scales are zeroed, the edges of the reconstructed envelope begin to degrade because eliminating the low-scale wavelet coefficients removes the high-frequency components which are needed to represent the edge. All frequency components are necessary to reproduce that edge which is one porblem with finite duration signals.

The overall formant structure of the envelope is still intact even when the first 5 scales are discarded, but the formants in the reconstructed signal begin to take on the shape of the scaling function. Altered formant shape is most prominent in the D_2 case (Figure 4-1b), where the formant peaks look like the scaling basis functions.

The first four scales ($l = 4$), using the D_4 basis functions, can be removed without affecting the shapes of the formants and while maintaining a small difference error in the middle of the signal. Even though the overall formant structure is similar, the first formant is altered significantly and can cause poor speech synthesis. The lower-scale wavelet coefficients affect the reconstruction of unvoiced envelopes more than voiced envelopes (see Figure 4-2) because as the first few scales of wavelet coefficients are removed, some of the smaller peaks are not reconstructed.

Decompositions at different depths for spline wavelets of length 25- and 31-points are shown in Figure 4-3. No wavelet coefficients were zeroed, so the entire wavelet representation

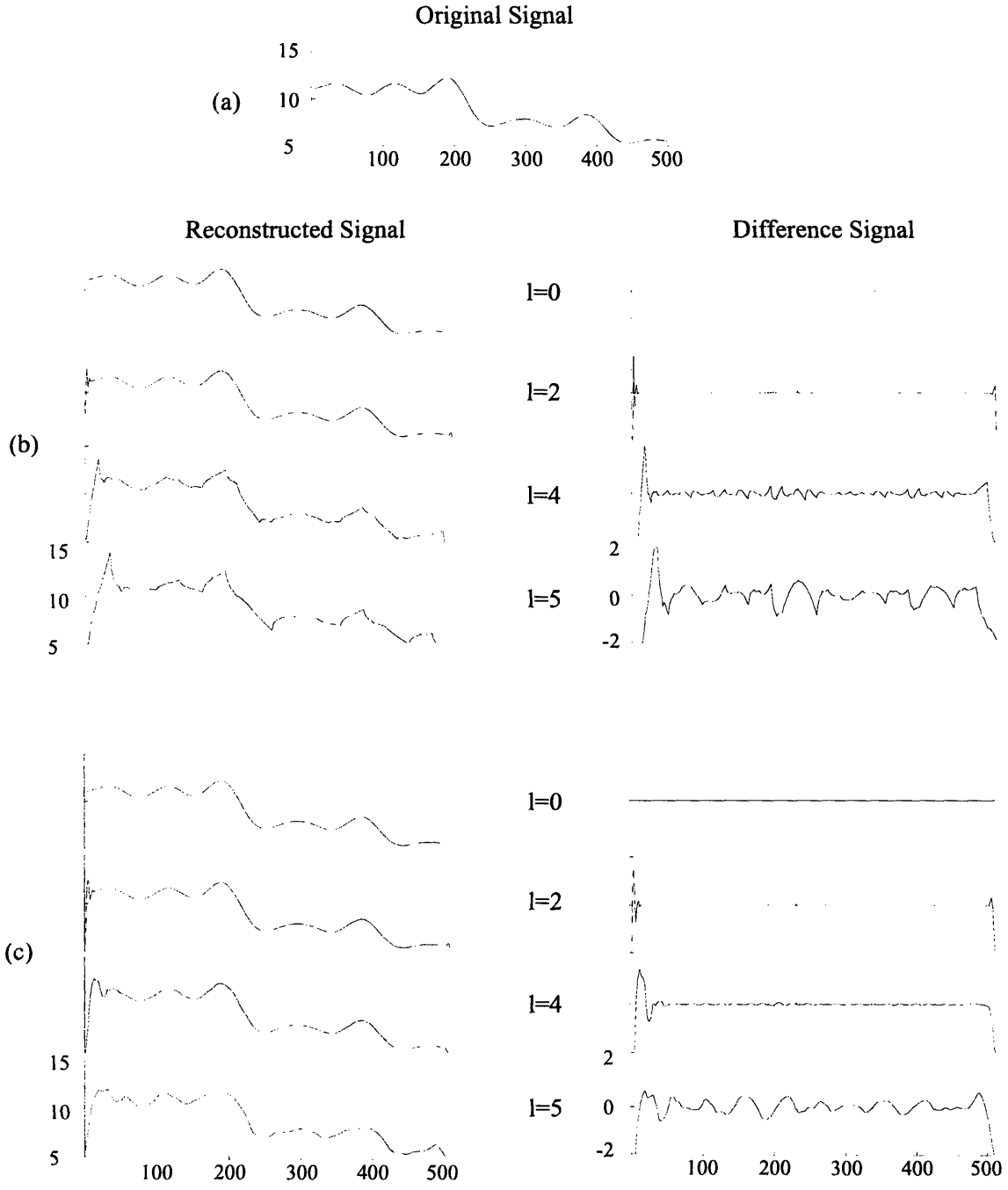


Figure 4-1: Reconstructed and difference error of an envelope representing voiced speech using two Daubechies compactly supported wavelets. (a) Original envelope for a frame of voiced speech. (b) Reconstruction using a compactly supported wavelet D_2 and (c) D_4. The first l wavelet scale coefficients are zeroed. All decompositions were done to scale 9 ($d = 9$).

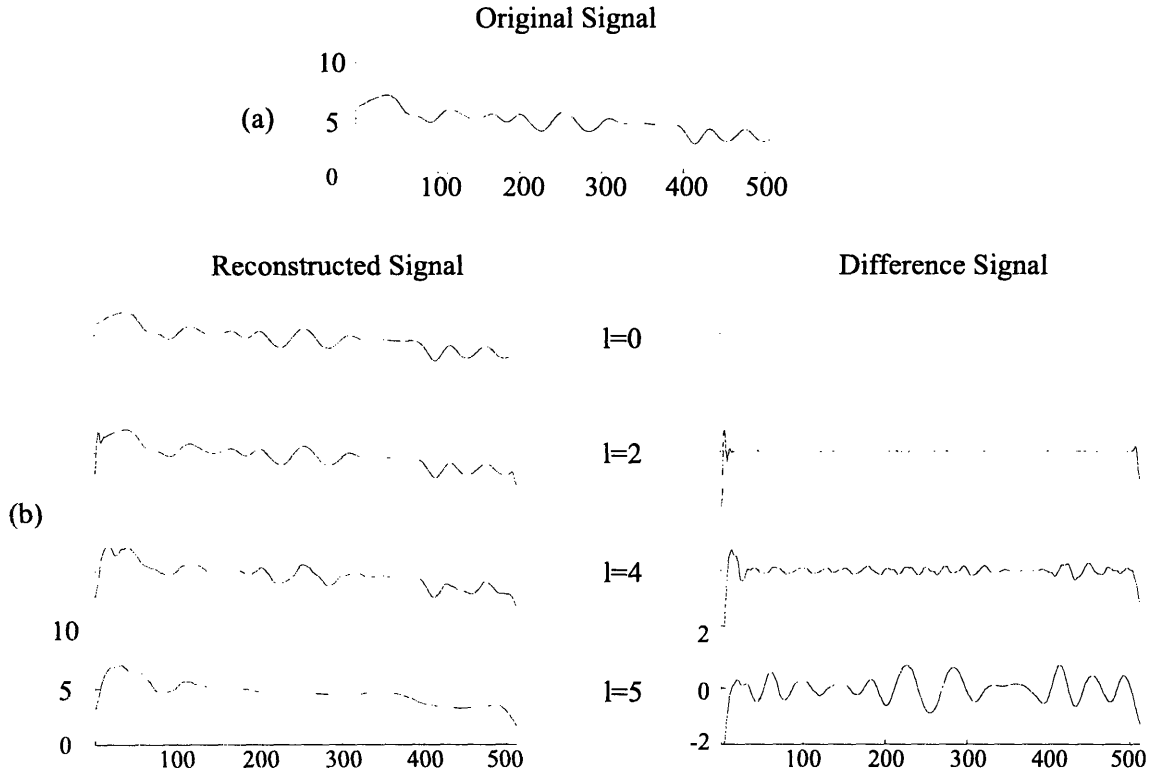


Figure 4-2: Reconstructed signal and difference error of an envelope representing unvoiced speech. (a) Original envelope for a frame of unvoiced speech. (b) Reconstruction using the Daubechies wavelet D_4. The first l wavelet scale coefficients are zeroed. All decompositions were done to scale 9 ($d = 9$).

is used. The envelopes were decomposed to a maximum scale depth of d. FDDs beyond the 4th or 5th scale were poorly reconstructed with large DC shifts, and altered formant structure.

At the 5th scale, the wavelet function is much larger than the envelope (Figure 4-4), so it is representing information beyond that of the envelope. Most of the information contained in the coefficients above the 5th scale is generated by the basis function in the form of the addition coefficients due to the convolution. Figure 4-4 shows how the wavelet basis scales with respect to the input envelope. Since these are perfect reconstruction filters, good reconstruction should be expected, but Figure 4-3 shows that the algorithm does not work when the signal is small compared to the filter length.

4.1.2 Truncation Experiments

One possible method to reduce the number of additional coefficients produced by the filters is to truncate the output of the convolution by $m - 1$, where m is the filter length. This

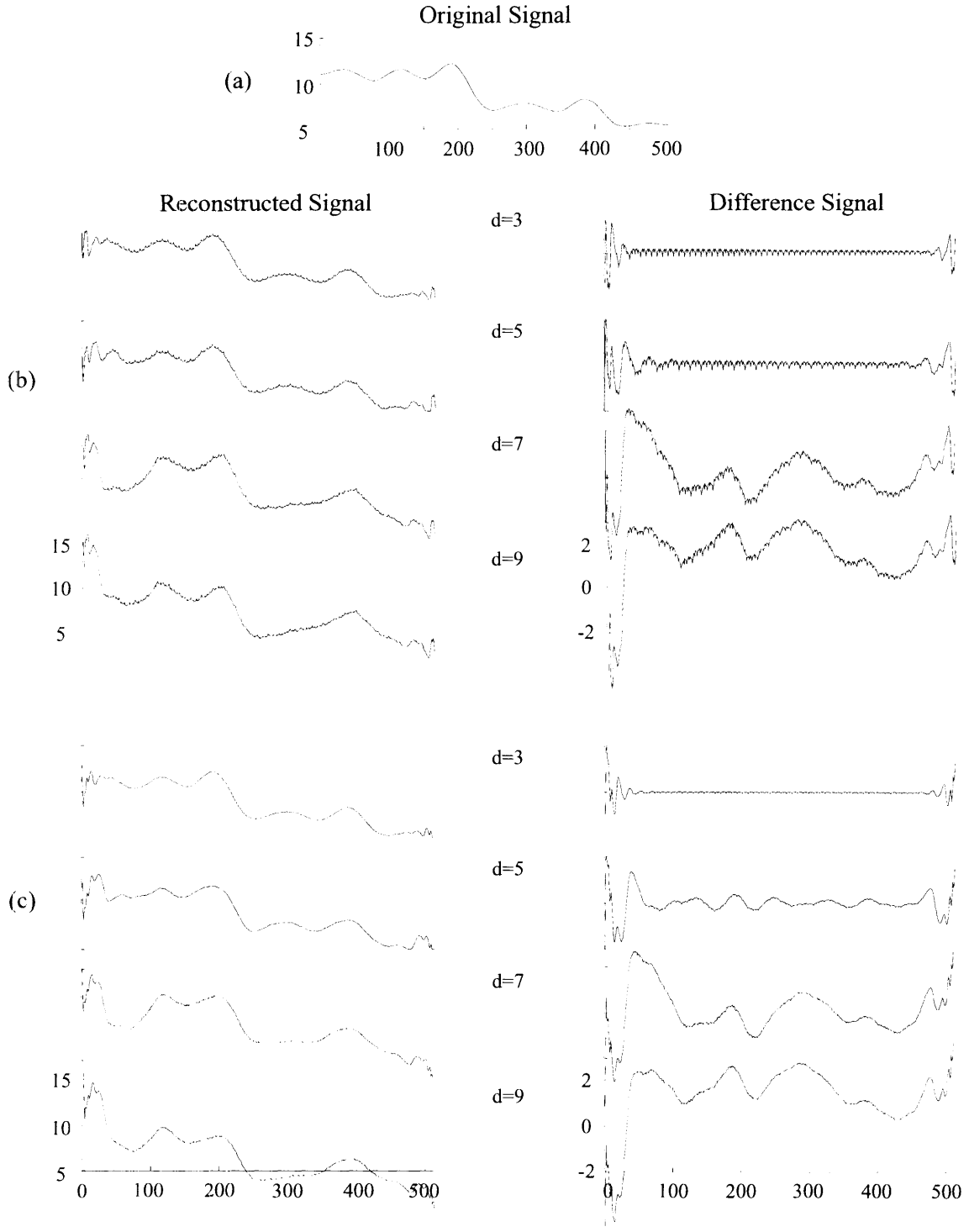


Figure 4-3: Reconstructed signal and difference error of an envelope representing voiced speech using the Mallat spline wavelet. (a) Original envelope for a frame of voiced speech. (b) Reconstruction using the Mallat spline wavelet with length 25 , and (c) 31. Reconstructions for different depths of decomposition ($d = 3, 5, 7, 9$) are shown. All the wavelet coefficient produced are used for reconstruction ($l = 0$).

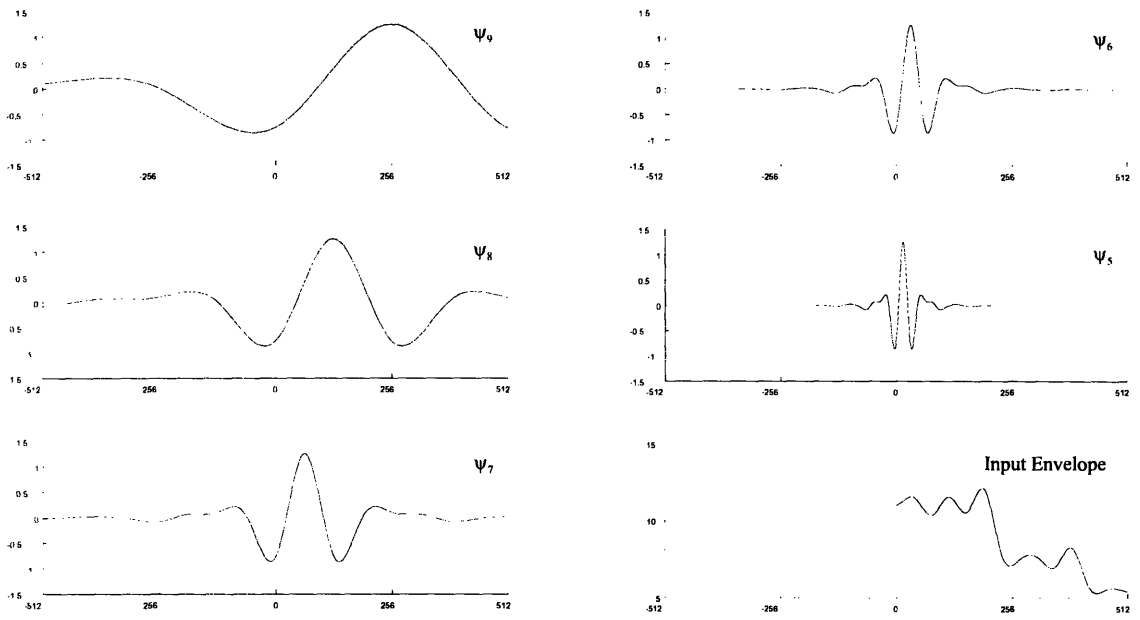


Figure 4-4: Spline wavelet basis at scales 9 to 5 compared to the input envelope.

algorithm produces large errors at the boundaries, and the decomposition is possible only to about the 3rd or 4th scale (for D_4) before the first formant is altered significantly. A large portion of the edge information is contained in the coefficients that have been truncated. An example of reconstructed signals at two different depths of decomposition scales is shown in Figure 4-5.

The boundary effects are large when the lower-scale coefficients are removed. A more efficient method is needed to eliminate the boundary effects in order to produce a low-rate WSTC which can reconstruct the spline envelope. These effects often alter the low-frequency speech spectrum, by altering the first two formants which results in poorly reconstructed speech.

Several problems exist with the convolution-based algorithm. For the Daubechies functions, all the wavelet scales are needed in order to reconstruct the edges, even though most of the wavelet coefficients at lower scales (0 to 4) have values near zero. The convolution generates additional coefficients, increasing the size of the representation which is opposite from the goal of reducing the representation. The spline wavelets do not perform well because of the finite duration of the signal and the infinite duration of the basis function.

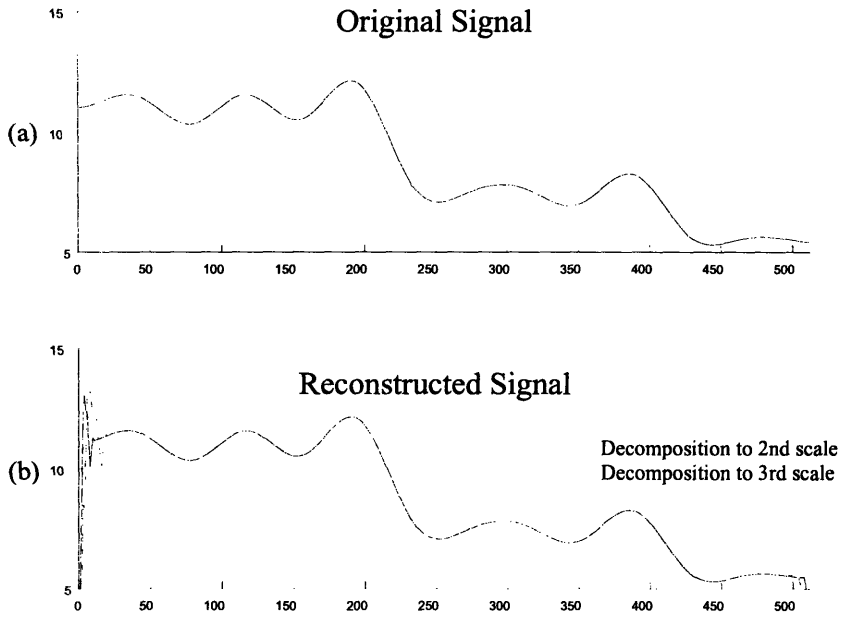


Figure 4-5: (a) Original envelope for a frame of voiced speech. (b) Reconstruction using the truncation convolution algorithm and Daubechies D_4 with a decomposition depth to scale 2 and 3.

4.2 Boundary Algorithms

Reconstruction of the envelope boundaries using the wavelet transform requires a large number of coefficients. The periodic and periodic-symmetric algorithms were evaluated to better represent the ends of the envelope. In both algorithms, the envelope is extended to represent the envelope over the entire space from $n = -\infty$ to $n = +\infty$ to eliminate the boundary effects. Extending the envelope is needed so the higher-scale wavelet basis functions overlap with useful information and project information contained within the original envelope. The spline wavelet basis at scale 5, for example, extends beyond the 512 interval of the envelope (Figure 4-4). Extending the envelope does not increase the representation because only those coefficients which represent the 512 points of the envelope are retained.

4.2.1 Periodic Envelope Algorithm

The periodic boundary matching algorithm makes the spline envelope periodic; therefore, the wavelet coefficients are periodic and the size of the wavelet representation does not increase because the algorithm is implemented using circular convolutions rather than the convolution described in the previous section. A circular convolution is also used in the synthesis portion

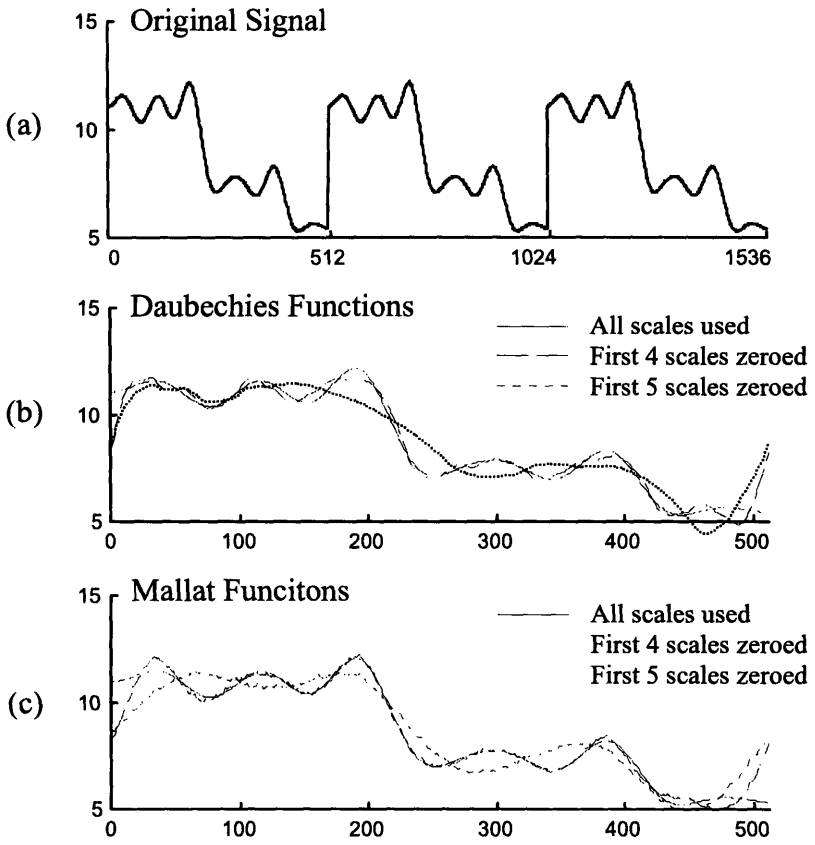


Figure 4-6: Periodic envelope algorithm with the envelope periodically replicated with a period of π. (a) Original and (b) Reconstructed signal using the Daubechies functions of regularity with $N = 4$. (c) Reconstructed signal using the Mallat spline basis function with length $N = 25$.

to reconstruct the envelope.

The periodic algorithm works well for signals in which there are no boundary edges or in which the boundaries are at the same DC level. The amplitude of the envelope at $n = 0$ is usually much larger than the amplitude at $n = 511$, creating a large edge at each boundary (Figure 4-6). The wavelet coefficients at all scales are needed to reconstruct these edges. The first formant can be altered by inadequate reconstruction of the boundaries when only a partial set of coefficients is used for reconstruction. Shifts in the higher-order formants are not as noticeable perceptually or visually.

4.2.2 Periodic-Symmetric Boundary Matching

The envelope is made symmetric about the origin and is then made periodic by using the circular convolution (Figure 4-7). The new wavelet representation has twice as many coef-

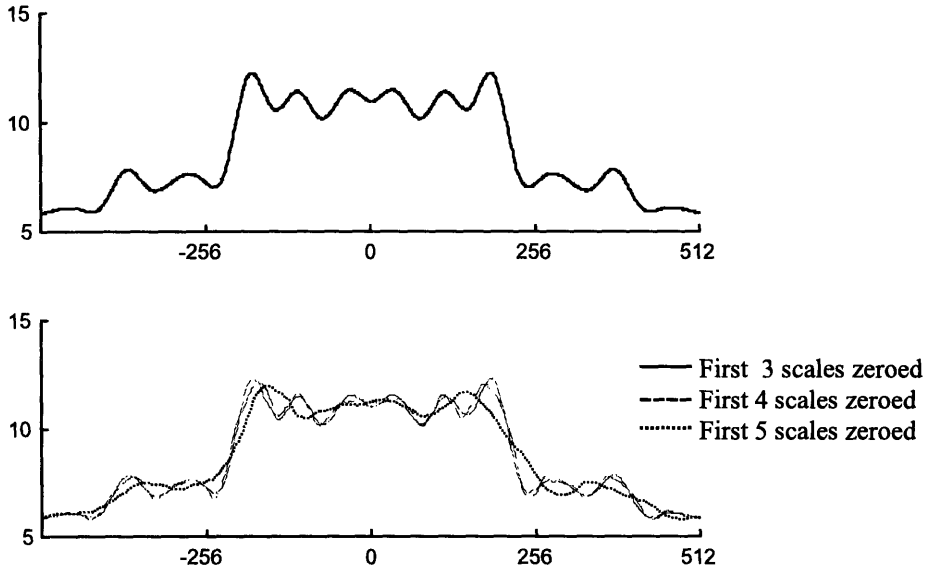


Figure 4-7: (top) Original symmetric-periodic envelope. (bottom) Reconstructed signal using the Daubechies D_4 basis function while zeroing the first l wavelet scales.

ficients as in the previously described algorithms because each period is a length of 1024 points. Using Daubechies' basis functions, the envelope can be reconstructed using the wavelet coefficients from scales 5 to 9 (Figure 4-7). Since the original signal is twice the size of the envelope, the number of coefficients used in this representation is 64.

The main difference seen in this reconstruction between the 4th and 5th scale is the ability to reconstruct the first peak in the envelope. When the 4th-scale coefficients are used (first 3 scales zeroed), the peaks are better represented, and the error between the original and reconstructed signal is small. Better reconstruction can be obtained using the lower-scale coefficients in addition to the higher-scale coefficients. The first peak is less prominent when the envelope symmetric, so the synthesized speech poorly matches the original speech due to a missing or reduced first formant.

The symmetric representation is redundant since half the wavelet coefficients represent the mirror image of the envelope. The wavelet coefficients obtained using Daubechies' wavelets are not symmetric because asymmetric basis functions produce asymmetric coefficients even if the signal is symmetric. If the wavelet coefficients were symmetric, then half of them could be discarded. Half of the coefficients were discarded as a test and the reconstructed signal is similar to the truncation case in which the boundaries are altered but the middle of the envelope can still be reconstructed, and there is a small DC shift in the signal.

A symmetric basis function should produce symmetric coefficients. Since the spline scaling

function is symmetric about the origin, the approximation signals at each scale should also be symmetric about the origin. While the continuous approximation signal is symmetric, the coefficients are not symmetric because they are downsampled versions of the continuous approximation signal. Figure 4-8 shows that the approximation signals at each scale are symmetric, but when they are sampled, the sampled points are not symmetric. For example, the 8th approximation has coefficients which are projections at $-512, -256, 0$, and 255 and the coefficients at 0 and -512 do not have the same value. For the approximation coefficients, all the coefficients except the ones at 0 and -512 are symmetric.

The wavelet function is symmetric about the axis $x = \frac{1}{2}$ resulting in a shift in the projection. This shift in the projection allows for the sampled wavelet coefficients to be symmetric. For example, at the 8th scale, the wavelet coefficients are projections of the approximation at points $n = -384, -128, 127$, and 383 shifted to the locations $n = -512, -256, 0$, and 255. Figure 4-8 shows that the four coefficients are symmetric, so two coefficients (at $n = -512$ and $n = -256$) are redundant and can be removed from the representation.

The values of the symmetric wavelet coefficients are not exactly the same, but the difference between symmetric coefficients is always less than -30dB and is usually less than -60dB. The difference is probably due to the truncation of the wavelet filter; as the length of the spline wavelet filter is increased, the coefficients are more symmetric. Synthesized speech using the actual coefficients from -512 to 511 was compared to a reconstruction using only the coefficients from 0 to 511, which were then forced to be symmetric. The results showed that forcing the wavelet coefficients to be symmetric for the spline wavelet does not affect reconstruction. Forcing the wavelet coefficients generated by Daubechies' wavelet basis to be symmetric produced large alterations in the envelope.

For voiced speech the reconstruction using wavelet coefficients from the 5th to 9th scales is sufficient to produce an envelope which is visually identical to the original spectral envelope. Using one-half of the wavelet coefficient from scale 5 to 9 ($d_{5,n}$ to $d_{9,n}$) and both of the approximation coefficients at scale 9 ($a_{9,n}$), the wavelet representation for the spline envelope is reduced to 33 coefficients, which is a significant reduction in the number of coefficients from the number of coefficients produced by the complete decomposition.

Figure 4-9 shows examples of both a voiced and a transitional envelope reconstructed using the wavelet representation beyond the fifth scale. Voiced speech is well represented using the 33 coefficients, but transitional speech is often poorly represented due to the lar-

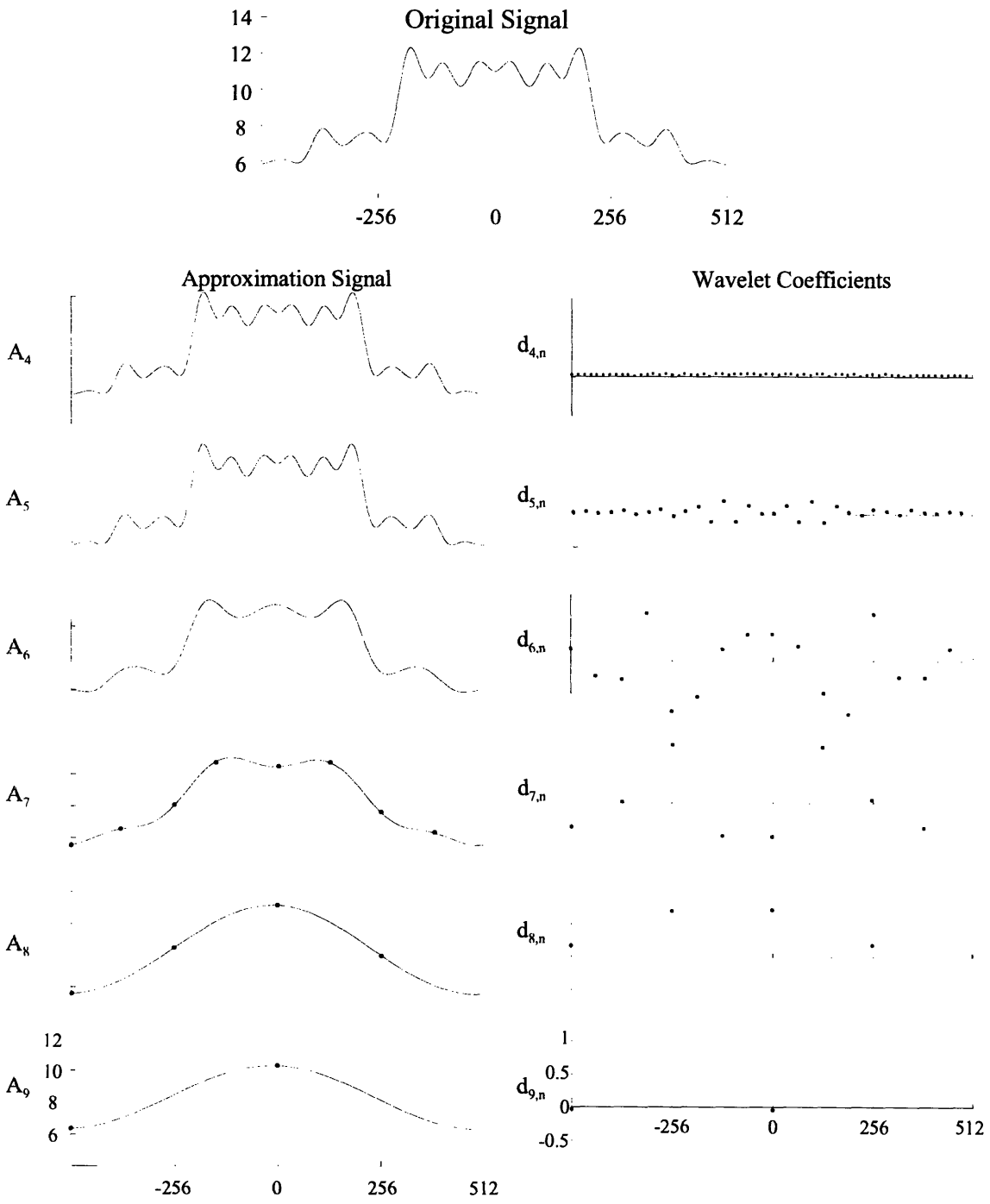


Figure 4-8: Original symmetric-periodic envelope for a frame of voiced speech. The approximation signals $A_4x(t)$ to $A_9x(t)$ are symmetric, but the sampled coefficients $(a_{j,n})$ are not. The detail or wavelet coefficients $d_{4,n}$ to $d_{9,n}$ are symmetric. Decomposition was done using the Mallat spline wavelet basis of $N = 25$.

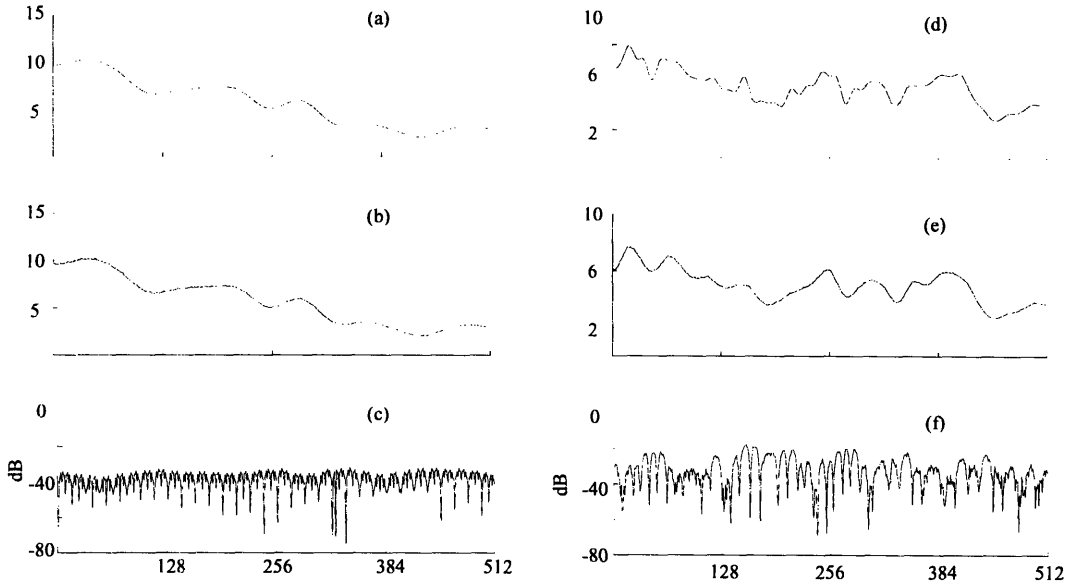


Figure 4-9: (a) Original voiced envelope, (b) reconstructed using the symmetric-periodic algorithm, and (c) difference error envelope for a frame of voiced speech. (d) Original transitional envelope, (e) reconstructed using the symmetric-periodic algorithm, and (f) difference error envelope for a frame of unvoiced/transitional speech. Reconstruction was performed using wavelet coefficients from scales 5 to 9 and the approximation at the 9th scale. The Mallat spline wavelet with length 25 is used.

ger number of peaks in the envelope. The peaks in the transitional envelope have smaller bandwidths, giving it a higher frequency content. Adding the wavelet coefficients at the 4th scale ($d_{4,n}$) doubles the coefficients, but the reconstructed envelope is almost identical to the original signal.

Fine ripples in the reconstructed envelope are produced by truncation of the filters $g[n]$ and $h[n]$ which produce small edges in the wavelet and scaling functions. These ripples should have little effect on the reconstructed speech, since the sine-wave system samples the envelope at the fundamental frequency and the ripples are small and closely spaced.

4.3 Discussion of Wavelet Algorithms

The two basis functions used in this chapter were Daubechies' compactly supported wavelets and a cubic spline wavelet. It is hypothesized that the cubic spline wavelet bases are better suited for representing the spline envelope because of their shapes are similar to the formants. Figure 4-10 shows the wavelet coefficients for both the spline wavelet basis and the D_2 wavelet basis.

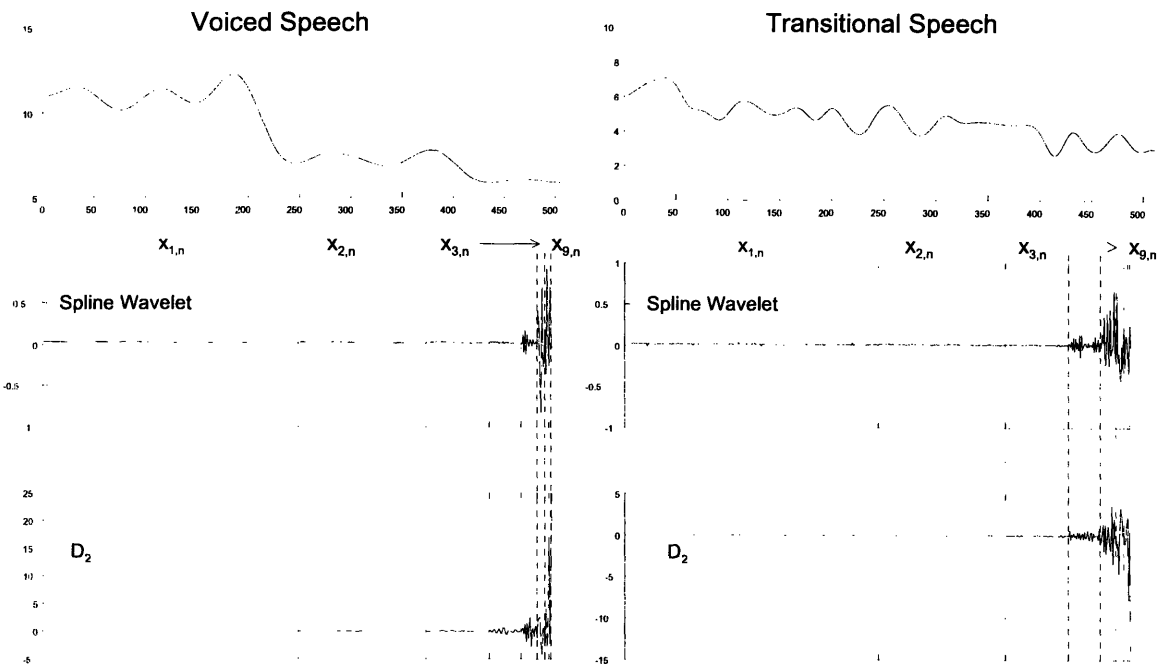


Figure 4-10: Voiced and Unvoiced spline envelopes. The wavelet coefficients are shown using the spline wavelet of length 25 (middle) and the D_2 wavelet (bottom).

Many voiced frames decomposed using the D_2 wavelet generate 4th-scale wavelet coefficients required for reconstruction whereas the spline wavelet for the same speech frame did not need the 4th-scale coefficients for reconstruction (Figure 4-10). During unvoiced and transitional frames, both wavelets produce sizable coefficients in the 4th frame when there are a large number of peaks in the envelope. Based on this, it seems that the spline wavelet is better than the Daubechies' wavelet for compressing voiced frames.

In addition to the size, the other noticeable difference between the two sets of wavelet coefficients is the range of values for the coefficients. The higher-scale coefficients have a much larger dynamic range for the D_2 wavelet which may make them more susceptible to quantization errors.

The first four wavelet scale coefficients were analyzed using the spline wavelet to make sure that the low-order scales could be discarded. The wavelet coefficients at the low-order scales, corresponding to rapid transitions and spikes, are small or zero for the spline envelope (Figure 4-11) because the envelope contains mostly low-frequency energy. Most of the envelope's information is contained in a subset of the wavelet coefficients; therefore, many of the lower-scale (below the 5th scale) coefficients can be eliminated. For signals with a large number of peaks, such as a fricative speech signal (Figure 4-11d), there are large coefficients

in the lower-order scales as well.

During voiced speech, the formant structures are very distinct, and the sine-wave system constructs a smooth envelope which contains fewer than 10 peaks. During unvoiced and transitional speech, however, the number of peaks in the envelope usually exceeds 10, and the bandwidth of each peak is smaller than the bandwidths during voiced speech. Peaks with smaller bandwidths, and envelopes with more peaks can have significant energy in the lower scale wavelet coefficients while voiced speech has its energy primarily in the highest few wavelet coefficients.

When the bandwidths of the peaks are small, coefficients from lower scales generate a significant part of the representation. The reason for this is that the bandwidth of the spline wavelet, which is used in Figure 4-11, is also smaller at the lower scales. Figure 4-4 shows the cubic spline wavelet at many scales along with a typical spline envelope. The bandwidth of the wavelet peak decreases as the wavelet scale decreases. The graphs are shown on the same axis scales. The bandwidths of the peaks of most voiced envelopes are about the same size as the bandwidth of the basis functions at scales 5 through 7; therefore, a basis decomposition with these scaling basis functions would strongly represent the formant structure.

4.4 Wavelet Algorithm in the WSTC

The symmetric-periodic algorithm reduces the number of wavelet coefficients while maintaining high-quality reconstruction. This algorithm was chosen to be the one used for the wavelet transform in the WSTC. Mallat's spline basis functions are used in the algorithm because they provide a greater reduction in the number of coefficients than Daubechies' basis functions. Spline basis functions are symmetric, so the periodic-symmetric algorithm can be used without increasing the number of coefficients.

The number of coefficients is reduced using the spline wavelet, but the coefficients must also be numerically stable so that they can be quantized. The properties of the coefficients are examined in the next chapter.

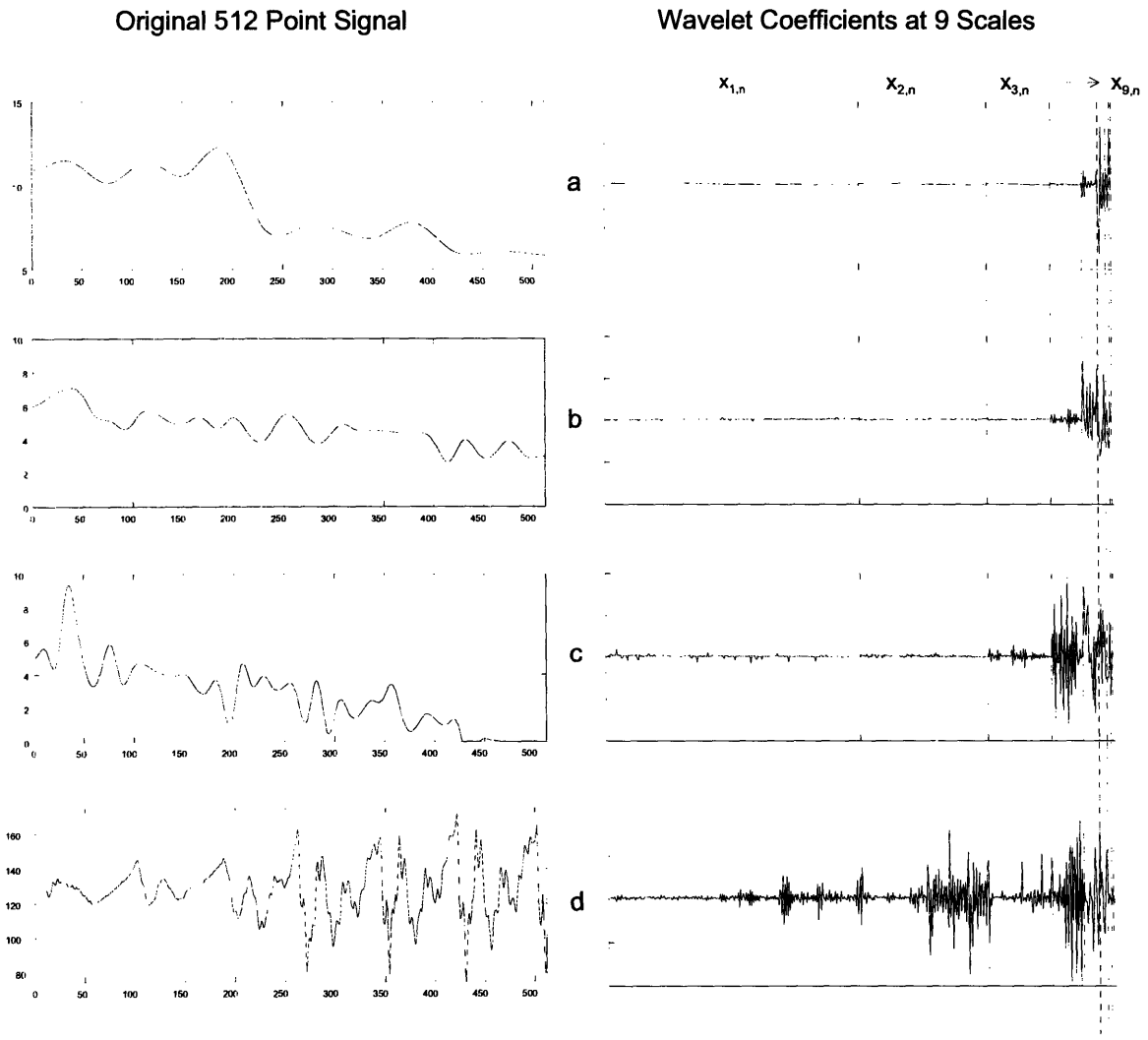


Figure 4-11: Three types of spline envelopes and a speech signal of /t/. (a) Voiced speech produces an envelope which has distinct formant structure. (b) Unvoiced speech contains many peaks. (c) Transitional speech has many more peaks and can sometimes be very flat. (d) The speech signal is 512 points of a /t/ utterance. The column on the right shows the wavelet coefficients for the signals. The coefficients are plotted on the same axis with the lower-scale coefficients on the left. The scales are separated by dotted lines. A cubic spline wavelet of length 25 was used for the decomposition.

Chapter 5

WSTC System

Design and testing of the WSTC system, which is based on the wavelet algorithm developed in the previous chapter, is explained in this chapter. A block diagram of the WSTC system is shown in Figure 5-1 with the shaded boxes representing changes made to the original STC system (Figure 2-4). The unshaded sections are common to both the STC and WSTC systems.

The distributions of the wavelet coefficients are analyzed to determine their numerical significance. Using information from the distribution and the structure of the basis functions at the different scales, several methods of quantizing the coefficients are explored. The WT algorithm is modified by extending it to the 10th scale, so the highest-scale coefficients are better quantized. The synthesis portion is used to evaluate the reconstruction of the spline envelope and to determine the effectiveness of the coding algorithms developed in the analysis section.

5.1 Analysis Section

The symmetric-periodic algorithm discussed in Section 4.2.2 is used to convert the spline envelope into wavelet coefficients. The Mallat spline wavelet of length $N = 25$ is the wavelet basis function used in the WT because it produces fewer coefficients than Daubechies' wavelet basis functions tested. The reduced representation may be due to similarities between the smoothness of the peaks in the envelope and the smoothness of the main peak in the spline wavelet and scaling functions. Figure 4-4 shows similarities between the peaks in the envelope and in the scaled wavelet functions.

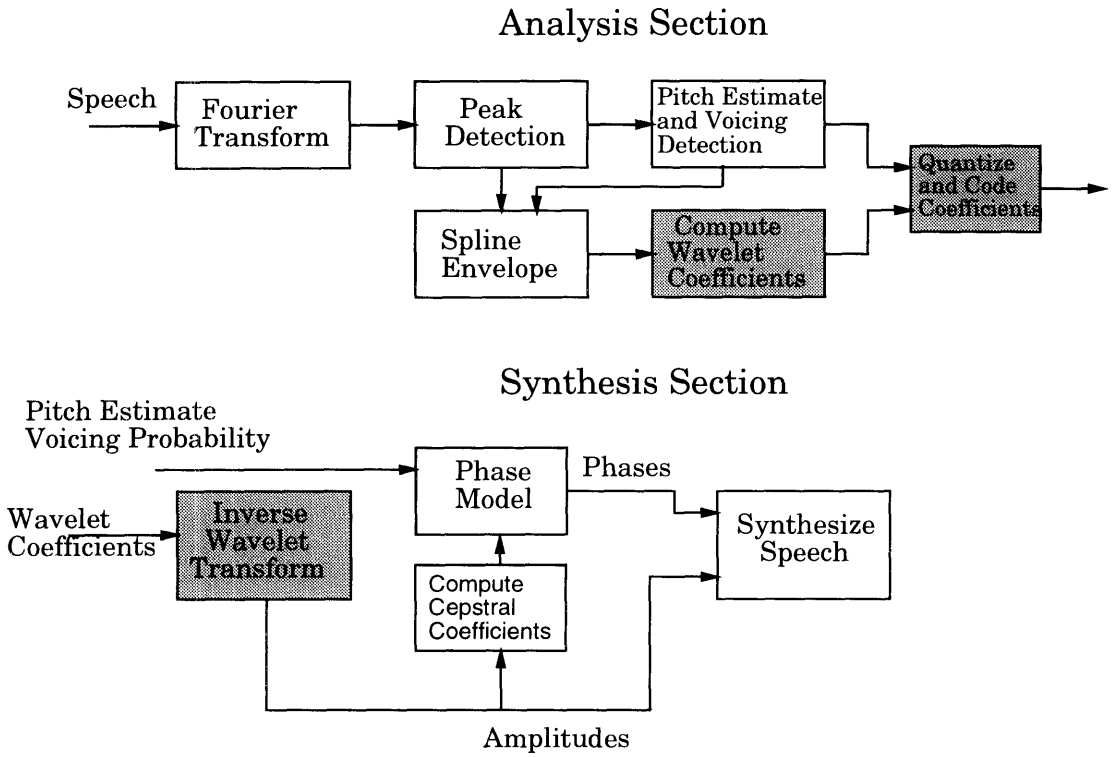


Figure 5-1: Block diagram of the main portions of the wavelet-based sinusoidal transform coder. The shaded components are the sections that have been added to or modified in the original STC system.

Tests were performed with the depth of decomposition to the ninth scale. All the wavelet coefficients across all scales were analyzed along with the two approximation coefficients at scale 9 ($a_{9,n}$). From this analysis, quantization methods for coding the coefficients were examined. A quantization scheme using 180 bits per frame or lower is attempted which is at the same rate as the STC system coding at 4800 bps with frame-fill.

5.1.1 Analysis of Coefficients

The maximum, minimum, mean, and variance of the coefficients at each scale, pooled across about 10,000 frames, are shown in Table 5.1.

The means and variances for the coefficients in the first four scales are close to zero. These coefficients can be removed from the representation without affecting the reconstruction, as shown in Chapter 4. The fourth scale is required for compelete reconstruction of many unvoiced and transitional frames, but is not used in the WSTC.

The variance progression of the pooled coefficients should follow that of Equation 3.44 if

Coefficients	# of Coef. Per Frame	Maximum Value	Minimum Value	Mean	Variance
$d_{1,n}$	256	0.290	-0.519	0.016	0.0001
$d_{2,n}$	128	0.302	-0.294	0.015	0.0001
$d_{3,n}$	64	0.290	-0.318	0.016	0.0003
$d_{4,n}$	32	1.290	-1.090	0.017	0.006
$d_{5,n}$	16	1.98	-2.21	0.015	0.027
$d_{6,n}$	8	1.85	-1.78	0.014	0.149
$d_{7,n}$	4	1.80	-1.98	0.0078	.2247
$d_{8,n}$	2	1.78	-1.84	-0.046	.6114
$d_{9,n}$	1	1.23	-1.63	-.157	.2311
$a_{9,n}$	2	14.8	0.80	8.45	2.67

Table 5.1: Maximum, minimum, mean, and variance for wavelet coefficients at a particular scale across frames.

the envelope is like a $1/f$ process. Figure 5-2 shows that for the middle scales, the variances follow such a progression which suggests that the envelope is like a $1/f$ process. There are some deviations from the progression. The variance at the 9th scale is not as large as the progression would predict, which means many of the envelopes share a common structure and low-frequency content. Wornell did not present any data showing the variance at such high scales [37]. Other $1/f$ data that he presented were similar to the progression seen here. At the first scale, the variance is larger than the progression would predict, which is probably due to the limited precision of the wavelet filter coefficients.

In addition to analyzing the large pool of coefficients, a subset of frames, containing both voiced and unvoiced speech, was analyzed on a frame-by-frame basis. The mean and variance of the coefficients at each scale were analyzed a single frame at a time. The single-frame analysis showed that the variances of the coefficients at the first four scales are very small. In transitional or unvoiced frames with a large number of peaks, the 4th-scale coefficients had a variance 10 to 100 times larger than the pooled variance.

This result suggests that voiced frames can be distinguished from unvoiced and transitional frames. The subsets of frames were further separated into voiced and unvoiced subsets. Many of the unvoiced frames had a larger 4th-scale variance, but for the majority of unvoiced frames, where there were few peaks, the variance did not deviate from the progression. The 4th-scale coefficients, alone, are not a good measure of voicing. The larger range of the 4th-scale coefficients is caused by larger but narrower peaks during transitional and unvoiced speech frames.

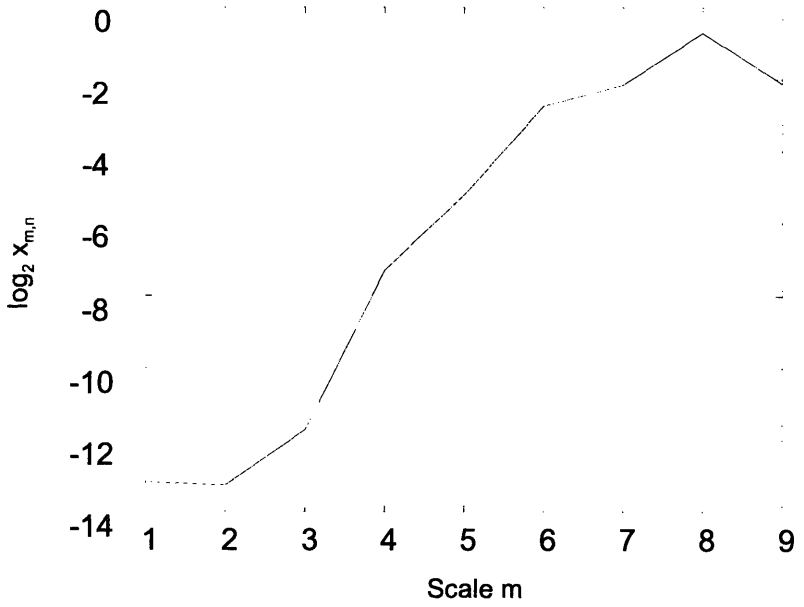


Figure 5-2: Variance progression for the scale-to-scale wavelet coefficients.

In the previous chapter, the coefficients in the first four scales were zeroed. When the coefficients are set to their mean value of 0.015, the fine ripples in the reconstructed envelope are reduced. There is also a slight improvement in the reconstructed envelope on the order of 2dB for most envelopes when non-zero values are used for the low-order scales. The mean value changes very slightly between different speakers by less than .001, so we can use the approximation of 0.015 for all speakers.

Histogram of Wavelet Coefficients

Figure 5-3 shows that the histograms for the coefficients used in the WSTC are Gaussian-like and have a dynamic range from -2 to 2. The 8th- and 9th-scale coefficient histograms have two peaks which look like a combination of two Gaussian distributions. At first it may seem that the two peaks in the histogram of d_8 are due to the individual coefficients, since there are two at that scale, but separate histograms of $d_{8,1}$ and $d_{8,2}$ still show a bimodal distribution.

When the frames are separated into voiced and unvoiced categories, there are no apparent differences in the coefficient distributions from the 6th to the 9th scales. The peak just above zero in the 8th scale histogram is slightly larger for unvoiced speech. The coefficients at the 5th scale had a larger variance for unvoiced and transitional speech than for the voiced speech. On a frame-by-frame analysis, the variance of the 5th scale coeffficients increased

suggesting that they might be useful predicting the voicing probability.

Projection Analysis

The wavelet transform is a basis decomposition, so the decomposition at each scale contains different information about the spline envelope. The wavelet coefficients for some samples frames of speech with the reconstructed envelopes are shown in Appendix D. The wavelet basis at scales 5 through 9 along with a sample envelope are shown in Figure 4-4. The basis functions get broader at higher scales, so the representations are coarser. The 5th to 7th scales represent smaller bumps or fine structure of the envelope. Peaks which are separated by less than 32 points are not well represented because at the 5th scale the coefficients are shifts of 32 points and those peaks separated by less than 32 points generate frequency content higher than the highpass cutoff of the frequency characteristic of the 5th-scale basis functions. The 32-point shift in the envelope corresponds to 256 Hz in the speech spectrum, so formants or structures separated by less than 256 Hz, are not well respresented. As a result, smaller amplitude peaks may be missed, and larger peaks may have a reduced amplitude or smaller bandwidth. An example of this can be seen in Frame 20 in Appendix D. In these cases, 4th-scale coefficients are needed to represent higher frequencies.

The location and bandwidth of the envelope peaks is important for synthesizing high quality speech which sounds similar to the original speech. Since the basis functions occur at discrete dyadic shifts at each scale, the location of the envelope peaks may not exactly overlap with the peak of the wavelet basis function. The coefficients at the lower scales adjust the location of the peak because of their finer sampling rate. Mallat showed a peak can be tracked across multiple scales and therefore this could be useful in adjusting the envelope reconstruction [21].

The width of the spline function at the 6th and 7th scales is similar to the widths of most of the envelope peaks. The peak bandwidths reconstructed from the 6th and 7th scales are modified by the 5th-scale coefficients. Removal of the lower scales means that corrections to the peak bandwidths and location will not be made in the reconstruction.

The 8th- and 9th-scale coefficients represent changes in the speech spectrum from low to high frequencies. Voiced envelopes show a two-step shape in which the mean amplitude at the lower 256 points is different from the mean amplitude at the higher 256 points (Figure 4-11). In these cases, one of the 8th-scale coefficients has a much larger value than the other.

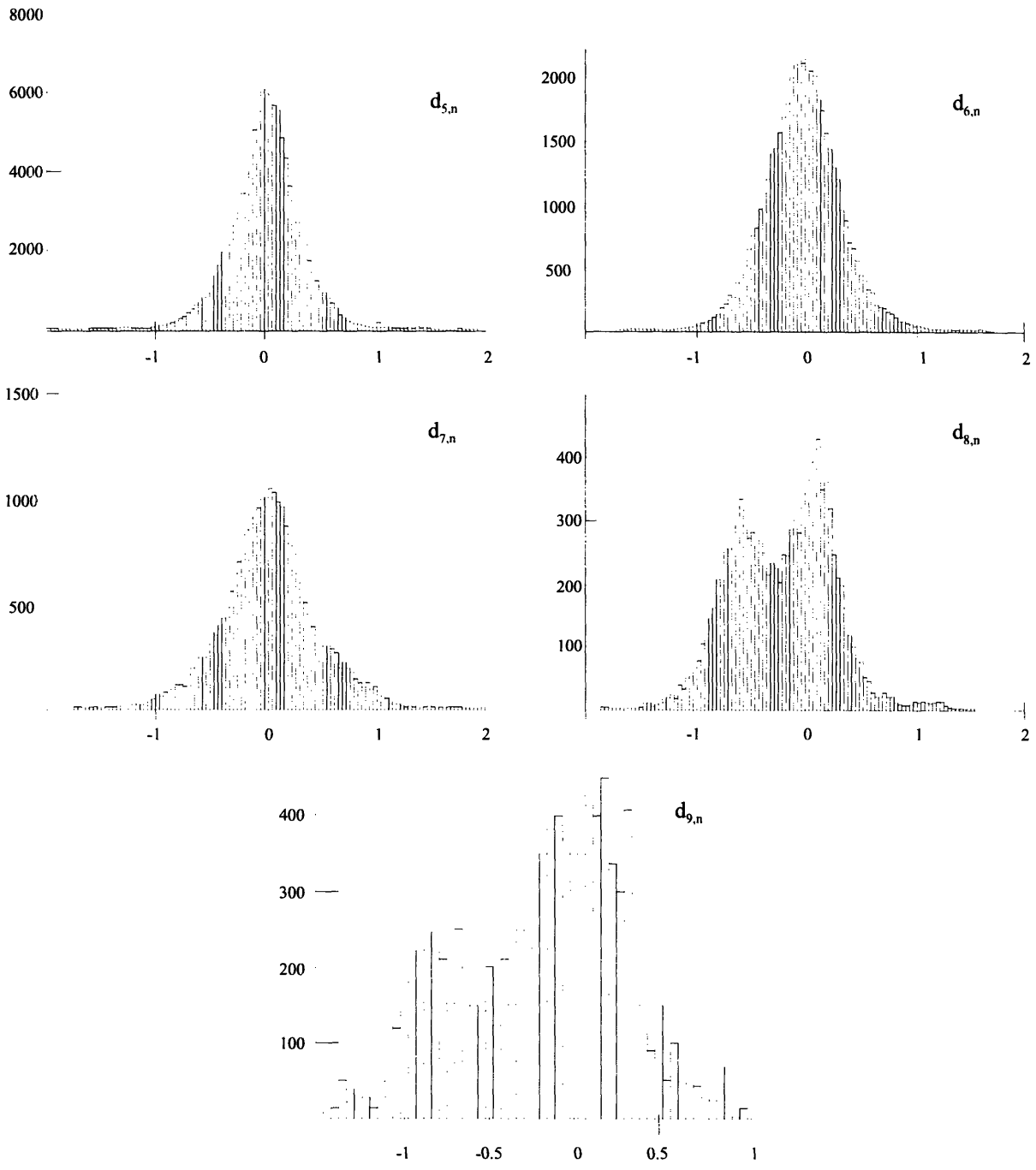


Figure 5-3: Histograms for wavelet coefficient which are used in the WSTC. The bin width for the 9th-scale coefficients is larger due to the smaller number of coefficients. Coefficients were computed across approximately 10,000 speech frames.

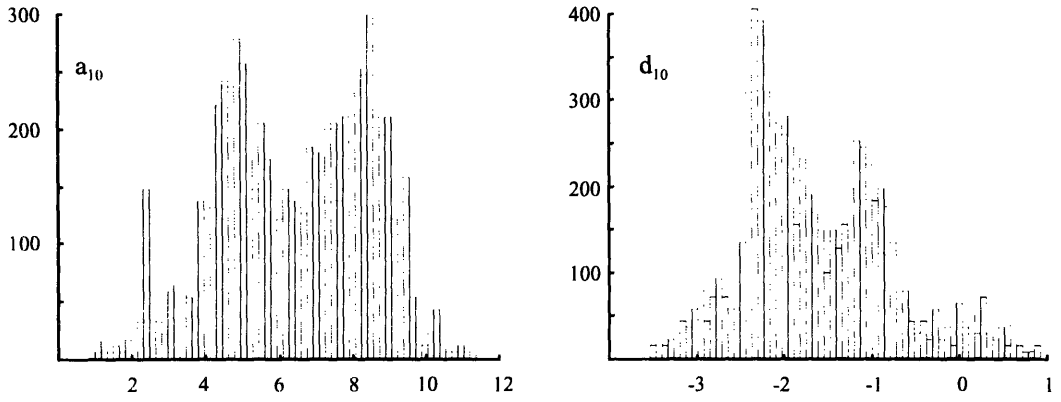


Figure 5-4: Histograms for approximation and wavelet coefficient at the 10th scale.

Approximation Coefficient

Both approximation coefficients at the 9th scale must be retained because they are not symmetric. The two coefficients represent the spectral tilt and average energy of the speech spectrum (see Figure 4-8). The coefficients have a non-uniform distribution which does not look Gaussian, with a range from 0 to 14. High-coding precision of these coefficients is needed to ensure that the speech energy is maintained. When random noise from -.4 to .4 was added to the coefficients, the reconstructed speech fluctuated in power, and the synthesized speech had loudness fluctuations. The quantizer must have a percision greater than .4, which corresponds to at least 7 bits per coefficient.

The two approximation coefficients are better represented by taking the average and the difference between them. The difference yields the spectral tilt, and the average gives the mean power of the speech. Implementation of another wavelet decomposition, to the 10th scale, is equivalent to taking the mean and the difference.

The approximation at the 10th scale represents the mean power of the speech spectrum. The range for a_{10}, the approximation coefficients, is 0 to 12; and a histogram of a subset of frames is shown in Figure 5-4. The 2 peaks in the distribution are not correlated to voicing probability. No correlations could be found between the peaks, with one possible exception: during periods of silence or transition, the values of the coefficients tended to be lower.

The wavelet coefficient d_{10} represents the spectral tilt. In most of the voiced envelopes, the mean value for the first half of the envelope is higher than for the second half. This difference is the spectral tilt of the envelope. Unvoiced and transitional frames usually have very flat envelopes. In this case, the 10th-scale coefficient is close to zero, or slightly positive.

The histogram of the coefficients shows two peaks which are correlated to voiced and unvoiced frames with the peak near -1 is composed mostly of unvoiced speech. The wavelet coefficient, d_{10}, has a larger range than the other wavelet coefficients, from -3.5 to .5, and its mean is centered near -2.

5.1.2 Quantization of Coefficients

For 4800 bps system, there are 180 bits per frame available for coding the coefficients. The 10th scale coefficients (d_{10} and a_{10}) are coded with 16 bits. The remaining bits can be evenly distributed among the remaining 31 wavelet coefficients. This would allow for 5 bits per coefficient. Since a lower rate is desired, 4 bits were allocated for the 31 wavelet coefficients (3 bits for the value and one sign bit) and 12 bits were allocated for the 10th scale. If the frame rate is 50 Hz with frame-fill, the coder rate is 3700 bps.

The next step, after computing the coefficients, is to quantize them for the synthesis system. Quantization can be done in three ways: uniform quantization, logarithmic division, and inverse log.

The range of the coefficients is between -2 and 2, but 97% of them are between -1.5 and 1.5, which is therefore chosen to be the quantization range. The coefficients outside this range corresponded to transition frames which contained a large edge, or to frames in which the differences between formant peaks were large. The quantization results are shown in Table 5.2.

Code	Uniform	Logarithmic (1)	Logarithmic (2)	Inverse Log (1)	Inverse Log (2)
000	.1875	.0053	.0234	.6980	.2561
001	.3750	.0346	.0937	.9008	.4617
010	.5625	.1043	.2109	1.046	.6517
011	.7500	.2279	.3750	1.162	.8322
100	.9375	.4181	.5859	1.262	1.006
101	1.125	.6862	.8437	1.349	1.175
110	1.312	1.043	1.148	1.428	1.339
111	1.500	1.500	1.500	1.500	1.500

Table 5.2: Quantization breakpoints for three methods of quantization. One additional bit is used for coding the sign.

The first scheme tested was the uniform quantization. The range was quantized uniformly into 16 bins from -1.5 to 1.5. The distributions for the quantized coefficients from scales

5 through 7 remained Gaussian with a smaller variance, with the cofficients having used only one or two bits. The quantized coefficients in the 8th and 9th scales formed uniform distributions.

The logarithmic quantizer allocates more bits to better quantize coefficients near zero. This quantization scheme flattens the Gaussian distribution, making the distribution of the quantized coefficients more uniform. The inverse logarithm quantizer yields higher precision for the coefficients at the edges of the range. Most of the quantized coefficients are put into the lowest two or three bins for the lower scales.

Using a single quantization scheme for all coefficients does not reconstruct the envelope well because the scales represent different features of the envelope. The 5th-scale coefficients require higher precision at the lower values because they represent finer fluctuations in the envelope. The first logarithmic quantizer is used for the 5th scale. This quantizer over-represents the lower values for coefficients in the 6th- and 7th-scale coefficients, so they need to be represented at slightly higher magnitudes. The second logarithmic quantizer (2) is used to quantize these coefficients. The 8th scale is best quantized using the second inverse logarithmic quantizer (2). The 9th-scale coefficient is best coded using a linear quantizer. All these coefficients are coded with 4 bits. The 10th-scale wavelet coefficient and approximation coefficient are coded using a uniform quantization scheme, but with a different range than the other scales. The range for the wavelet coefficient, -3.5 to .5, is quantized using 5 bits. The approximation coefficient is quantized to 7 bits uniformly distributed from 0 to 12.

5.2 Synthesis Section

The synthesis section is broken into two steps. The first step in the synthesis section is to reconstruct the 9th-scale approximation from a_{10} and x_{10}. The wavelet synthesis takes the wavelet coefficients up to scale 9m and the 9th-scale approximation is taken as the input. The output is the reconstructed envelope. The coefficients are made symmetric at each stage for reconstructing the symmetric-periodic envelope. Instead of zeroing the wavelet coefficients at the lower four scales, the wavelet coefficients at these scales are set to a value of 0.015. Which is the mean value for a large set of envelopes tested. Additional coding bits can be used in the future to code the mean value for these coefficients.

Appendix D shows a set of consecutive frames which are reconstructed using the WSTC.

The unquantized wavelet coefficients are plotted below the original and reconstructed envelopes.

For voiced envelopes, all the peaks are represented. There are two main errors in the reconstruction. The envelope peaks and valleys are not reconstructed exactly because of quantization errors. A low sampling rates results in misalignment of the wavelet peaks and formant peaks which has the effect of making the formant bandwidths either wider or narrower. In some cases, there is also a small shift in the maximum frequency of the formant, which can alter the synthesized speech. This shift is prominent in many of the frames in Appendix D, as seen in frame 11, where the amplitude of the first formant peak is larger than the original and the frequency of the maximum is also shifted by a few points. The larger amplitude makes the bandwidth of the formant larger than it should be. In frame 12, the local minima between the first two formants is smaller in the reconstructed envelope. This may cause a reduction in the formant bandwidths.

In some of the voiced envelopes, there is a small peak at the beginning of the envelope. This peak is probably due to the first and second formants being very close to each other and at a low frequency. Because the envelope is made symmetric, this peak is not represented in the reconstruction if it is small enough and close to zero. Figure 4-7 shows an example of a frame in which the inital peaks are lost in the reconstruction if the lower-order scales are removed. Loss of the peaks is a problem when there is also a larger peak near the inital peak, because the coefficients represent the larger peak. This problem is more common in unvoiced speech, such as frame 63 in Appendix D, where the first little peak is not reconstructed. The small peak can also be exaggerated, which causes quality differences in the synthesized speech.

During unvoiced and transitional speech, the variance of the 5th-scale coefficients increases even though the mean does not. The increased number of peaks in unvoiced speech means that finer detail coefficients are needed. The logarithmic quantizer does not perform well for unvoiced speech. Uniform quantization for all scales is a good method for unvoiced speech. The peaks are exaggerated for unvoiced speech, as seen in frame 18. When there are too many peaks, usually greater than 16, many of them are removed from the reconstruction, as is seen in frames 20 and 37.

The next chapter describes how these errors affect the synthesized speech and how they compare to the STC system with similar coding rates.

Chapter 6

Evaluation of the WSTC

This chapter describes a comparison of the speech synthesized by the WSTC with speech synthesized by the existing STC using comparable coding rates. The evaluation was performed for the WSTC system at a coding rate of 3700 bps and for the STC system at a rate of 4800 bps. The WSTC system was not tested at 4800 bps because when the tests were performedm the coder rates were miscalculated. During the tests, the WSTC coder rate was believed to be slightly larger than 4800 bps, but subsequent calculations, shown in the previous chapter, show that the rate was actually 3700 bps.

6.1 Simulation

The comparison was done using a non-real-time simulation. Both coders were tested on a Sun-4/370 computer using speech phrases from a Lincoln Laboratory database. The synthesized speech from both coders was compared to the input speech and evaluated by the author for clarity and quality. Similar evaluations were performed by other untrained listeners who were presented a subset of the phrases.

The STC system was set to a zero-phase system with a 20-ms frame interval at a rate of 4800 bps with frame-fill activated. Only the pitch, voicing probability, and amplitude were used for speech synthesis. It has been shown that high-quality synthetic speech can be achieved using just the sine-wave amplitudes [23].

The WSTC system was set to a 20-ms frame interval with frame-fill. The pitch and the voicing probability generated by the STC system were used and the amplitudes represened as the 32 wavelet coefficients from $d_{5,n}$ to $d_{10,n}$ and the approximation coefficient $a_{10,1}$. When

these coefficients were passed to the synthesis system without quantization, the synthesized speech sounded almost indistinguishable from the input speech. The wavelet-based system is able to code the sine-wave amplitudes using 33 coefficients, and high-quality speech is recoverable from these coefficients.

The same WSTC system was implemented with the coefficients quantized so the coding rate of the system was 3700 bps. Problems with the reconstructed speech are primarily due to the quantizer algorithms developed in the previous chapter.

6.2 Evaluation

Speech synthesized by the WSTC system generally sounds similar to speech reconstructed using the STC system, but the STC-generated speech sounds slightly better. The weaker vowels generated by the WSTC system sounded more muffled and are not as crisp as the STC vowels. These sections of speech were analyzed on a frame-by-frame basis and the first peak or a small structure in the first formant was altered significantly. Muffling occurred in many cases where the first peak was attenuated or removed and during back vowels, for which the frequency of the first formant, F1, is smaller than 500 Hz. When the formant is below 500 Hz, it is represented by only one coefficient because of the sampling, and its peak location can be easily altered by quantization error.

Fricatives synthesized by both systems sound similar because the unvoiced envelope is poorly represented in both systems. The poor representation is probably due to both the harmonic model and a lack of fine structure in the reconstructed envelope. The harmonic model fixes the maximum number of sine waves used in synthesizing speech. Envelopes for unvoiced and transitional speech, in general, have more peaks than voiced speech. Using a limited set of coefficients does not allow for complete reconstruction of all the peaks. Some of the frames (20 and 37) show reconstructed envelopes in which many of the peaks are lost due to the wavelet reconstruction.

Both systems have similar clarity problems. Many of the reconstructed speech segments sound "muffled" or "nasalized." Unvoiced speech often sounds "buzzy" and like a computer synthesized voice. The lack of clarity could be due to the lack of the phase information in reconstructing the speech. Another possible explanation is the change in the formant bandwidths and amplitudes in the spline envelope. McAulay and Quatieri suggest that the

muffling and buzzing could be due to smoothing the formant null [23].

Comparing voiced and unvoiced speech, the synthesized unvoiced speech sounds worse than the synthesized voiced speech. Both systems generate a hiss or metallic sound in the unvoiced speech, but occasionally the WSTC-synthesized speech sounds better.

To understand differences between the two systems, reconstructed envelopes for subsets of unvoiced speech were compared. The WSTC-synthesized envelopes have more peaks reconstructed than the STC-synthesized envelopes. The peaks which were better represented in the WSTC often had maxima near the dyadic shifts of the 5th scale coefficients ($n = 0, 32, 64, 96, 128, ...$), which suggests that the wavelet projections form a good representation of the peaks. Despite the loss of some of these peaks, there was only a small noticeable difference in the synthesized speech. The loss of 4 to 5 peaks in the envelope can still produce intelligible unvoiced speech.

There is no noticeable difference between the two systems for male and female speakers. Both systems perform slightly worse for the two female speakers, due to a higher fundamental frequency. The same harmonic model is used in both the WSTC and STC, so similar results are expected for both systems.

These evaluations are all based primarily on a single listener. A small subset of the original speech, STC-synthesized speech, and WSTC speech were played to other people within the group. The other listeners were asked their opinion regarding differences which sounded subtle to the primary listener. The secondary listeners were not trained, so not all the qualities, such as brightness or vibrancy, were evaluated. To completely characterize the differences, several trained listeners should give their subjective evaluations of the synthesized speech. Another possible measure is to produce a psychoacoustical experiment in which the synthesized speech and the original speech are compared. The main problem with such psychoacoustical tests and asking for subjective information is that people may use different criteria for perceiving speech.

A quantitative assessment of synthesized speech could be performed using the diagnostic acceptability measure (DAM) and the diagnostic rhyme test (DRT). Time constraints on the project did not allow for the evaluation of the WSTC using these standard tests.

6.3 Conclusions

The spline wavelet basis functions have peaks with bandwidths similar to those of the peaks in the spectral envelope (Figure 4-4). The spline wavelet decomposes the spline envelope into fewer coefficients than the compactly supported wavelets, making it a better representation for low=rate coding. The spline wavelet may be better because both the envelope and basis functions are made using cubic spline, but this was not proven.

A basis decomposition – rather than the cepstrum, LPC coefficients or other representations of the envelope – allows for a better understanding of which features of the spline envelope correspond to perceptual difference in the synthesized speech as compared to the original speech. The 5th-scale coefficients represent the fine structures in the envelope and increasing bit allocation to these coefficients results in less buzzing or hissing in the unvoiced speech. The wavelet coefficients at the 10th scale suggest that the envelopes have a common tilt. Tests can be performed to find out what happens to the perceived speech when the tilt is set to a constant. In the future a systematic alteration of the different wavelet scale coefficients can be done to determine the effect on the synthesized speech. A systematic adjustment of the wavelet coefficients on each scale could give additional insight into how the spectral envelope relates to perception of the synthesized speech.

The study described in Appendix A shows that small changes in bandwidth are noticeable for a simple synthesized utterance. Since the formant bandwidth changes in the spline envelope are much larger than the 4dB and there is still little difference in the way the synthesized speech is preceived, the first formant bandwidth changes are probably not the only perceptual cues used in perceiving continuous speech. The structure of the spline envelope can be altered significantly, and yet the synthesized speech is still comprehensible.

Mallat and Zhong showed that the derivative first order spline wavelet is similar to an edge, so the wavelet transform finds the edges of images. The edge can be tracked across a set of scales to determine the exact location and strength of an edge [21]. Similarly, future work on the WSTC system could provide a method to adjust the location or bandwidth of the spectral peaks using the lower wavelet scales. In initial experiments, 4th scale coefficients were added around the major peaks in the 5th and 6th scales to make the peaks narrower. The goal was to make the formant bandwidths narrower to determine whether the muffled voiced speech was due to formant broadening. The additional 4th scale coefficients caused

the speech to sound metallic and synthesized. More work needs to be performed to determine the appropriate narrowing.

An iterative modification to the wavelet coefficient can be used in the WSTC where the error between the original and synthesized speech is used to modify the quantization of the coefficients. If a large peak in the spline envelope usually shows up as a coefficient across several scales, quantization errors can sum across the scales producing an exaggerated peak. If the error is used to adjust the coefficients across scales, the amplitude of the envelope peaks may be better represented and the errors in the reconstruction could be minimized. Testing of the quantization algorithms and bit allocation is needed to obtain better envelope reconstruction.

The distribution of the wavelet coefficients appears to have a Gaussian nature and follow a variance progression of a $1/f$ process. From the distribution of the coefficients, certain parameters of the $1/f$ process, such as the mean and variance, may be coded to make the representation more robust. Stochastic coding methods may be able to perform better than the quantizer developed in this thesis. Coding methods were not the main focus of the evaluation of the WSTC, but they could be explored in future work to reduce coding rates for the WSTC system.

The performance of the WSTC coder developed in this thesis is similar to the STC system at 4800 bps. Lower rates are difficult to obtain because the coefficients need at least 3 bits per coefficient to reconstruct the envelope. The envelope was not warped as is done in the STC, because it would not yield dyadic sampling. Warping could be achieved by discaring half of the 5th-scale coefficients corresponding to points $n = 256$ to $n = 511$ which would remove the finer structure at the higher frequencies. It has been shown that the ear is less sensitive to frequency differences at frequencies above 2.5 kHz, so this would be possible without having a large effect on the perception of the synthesized speech.

The wavelet-based sinusoidal coder provides another approach to coding speech at low rates. The system can represent the sine-wave amplitudes using 33 wavelet coefficinet which are uncorrelated and could lead to better coding algorithms. In this thesis, simple quantization of these coefficients leads to a low-rate coder which is comparable in quality to the exisiting STC system. Future analysis of the wavelet coefficients may lead to insights into which features of the envelope correspond to perceptual errors in the synthesized speech.

Appendix A

Formant Bandwidth Tests

A.1 Introduction

In the 1950s, Flanagan performed psychoacoustic experiments to determine the precision with which different formant parameters needed to be coded in speech coders. Flanagan performed just-discriminable differences for the formant frequency, over-all amplitude level (Figure A-1), and second formant amplitude (figure A-2) [12, 11, 13]. The results of these experiments were used to develop a formant-coding speech compression system. These studies did not explore the affect of formant bandwidth alteration on human perception and how it affects the quality of speech.

Flanagan found that the difference limen (DL), which is the noticed difference 50 percent of the time, for the formant frequency was about 12%. The DL for the overall amplitude is $\pm$ 1.5 dB and the DL for the amplitude of F2 is about $\pm$ 4 dB. Speech quality is most sensitive to variations in the formant frequency, and then to variations in the over-all amplitude.

A.2 Psychoacoustic Experiments

To determine the just-noticeable difference for the bandwidth of the first four formants, a simple AB discrimination test was administered to 8 untrained listeners.

The synthesized utterance (/a b a/) were produced using the Klatt synthesizer at MIT. The "normal" utterance was developed by Professor K. Stevens. The bandwidths of the first four formants were all changed with respect to the initial bandwidths of the formants in the normal utterance. The normal synthesized utterance has formant bandwidths of 90 Hz,

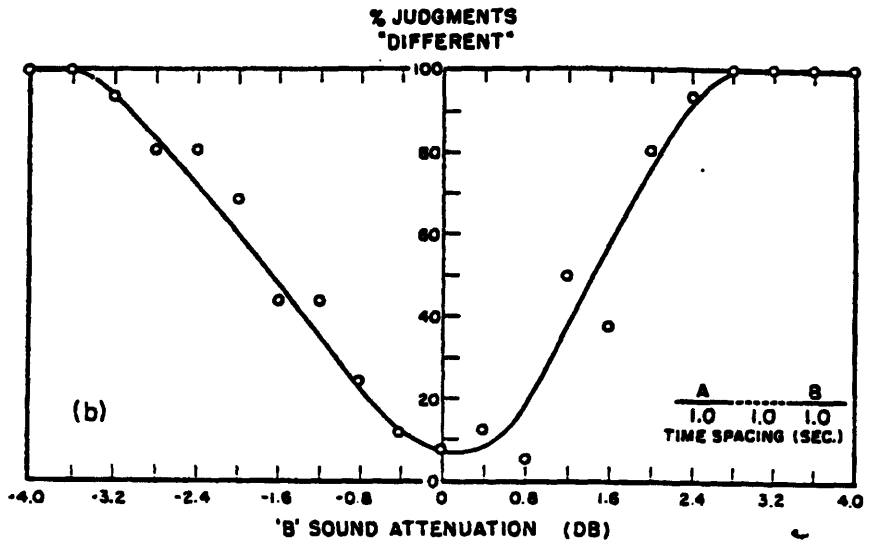


Figure A-1: Curves showing the percentage of trials called different as a function of the overall amplitude level [11].

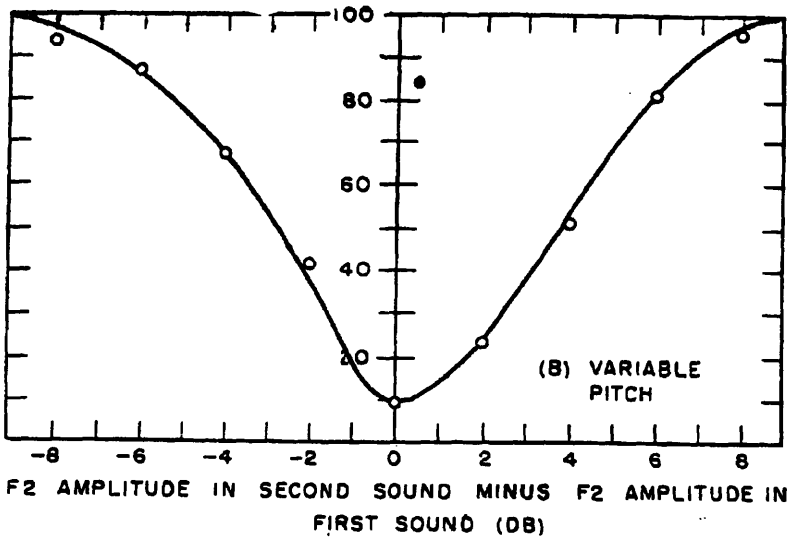


Figure A-2: Curves showing the percentage of trials called different as a function F2 amplitude [13].

120 Hz, 150 Hz, and 300 Hz for the first four formants respectively. Nine utterances were generated using bandwidth differences of -12, -6, -2.5, 0, 2.5, 4.9, 6, and 8 dB, corresponding to bandwidths that are 25%, 50%, 75%, 100%, 133%, 175%, 200% and 250% of the normal bandwidths.

The voicing amplitude parameter was varied along with the bandwidth to keep the RMS level of the spectra at three different points in the utterance (the first /a/, the burst of /b/, and the second /a/) within 1 dB of the "normal" utterance. The amplitude of F4 was the only formant amplitude that varied significantly.

Each trial in the test was an AB discrimination test, in which a pair of utterances with different formant bandwidths were presented. During each trial, utterance A was always the normal and the formant bandwidth of utterance B was altered. The experiment was administered as two tests. In one test, the narrow test, the bandwidths were made narrower. In the second AB test, the broad test, the bandwidths were the same or larger. The first just-noticeable test was performed to determine noticeable differences between the normal synthesized utterance and the altered utterances with smaller bandwidths (called the narrow test). The second test compared the normal utterance to the utterances with broader bandwidths (called the broad test).

The listeners were asked if the utterance B was the same or different from utterance A. Each test consisted of 25 AB trials. The listeners were briefly trained by being presented the normal sound six times in succession. Just before each test three sample trials were given where the utterances were at the extremes of the bandwidth change.

In addition to the AB discrimination test, the listeners were asked informally to describe the quality of the altered speech. This was asked to determine what questions might be asked in future experiments. This query also was expected to determine whether the listeners were hearing quality differences comparable to those heard in the speech coder.

A.2.1 Results

Plots were made for the total percentage of the judgments that were called *different* as a function of narrowing and broadening the formant bandwidths (Figure A-3. The difference limen is near $\pm$4dB.

One measure of variance was determined to be the number of times the AA combinations was judged differently. There was only one subject that responded with a different for a AA

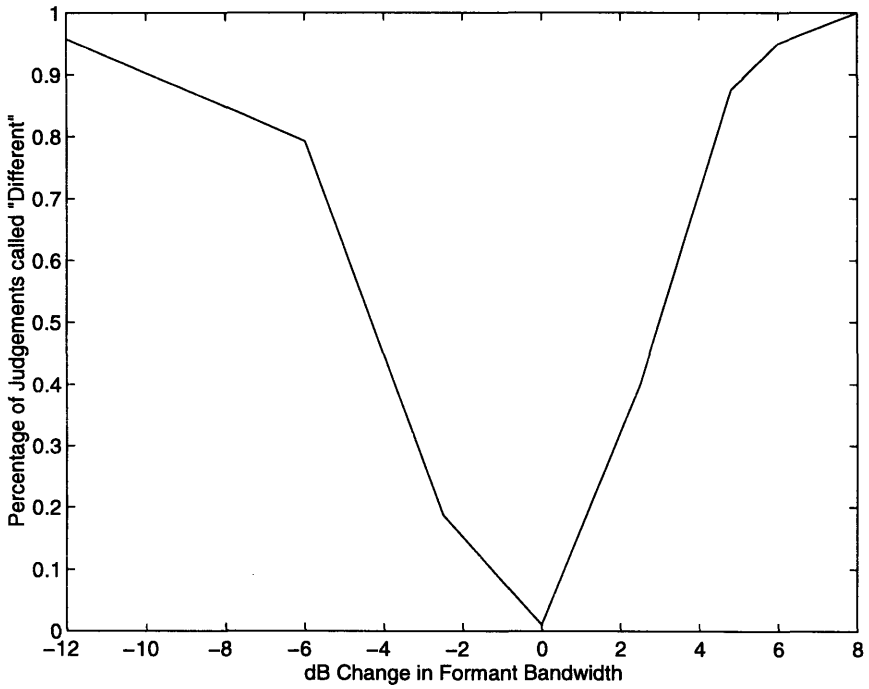


Figure A-3: Percentage of judgments called *different* as a function of formant bandwidth.

trial. But since that was one response out of 5 trials, the variance is small and the data was retained.

Each AB combination was presented 6 times for the narrow test and 5 times for the broad test. For example, subject 1 responded 3 times out of 5 that the 2.5 dB increase sounded the same as the original. Since a 2.5dB increase is at about a 30% noticability difference, a high variance is expected. At the extremes, very little variance is expected. At most the variance is 1 or 2 differing from the majority of the 5 responses. More trials need to be done to obtain better variance measures.

When asked about the what qualities were different between the utterances, many of them stated that if they had to choose a word for the broadened utterances, they would say that the utterance sounded muffled, but not softer. For the narrower utterances, several subjects stated that they did not sound very natural, or sounded harsh and breathy like someone with a false larynx.

A.2.2 Discussion

Before the experiment was conducted, the spectrum for three different locations of each of the utterances were examined to confirm that the formant and RMS amplitudes did not

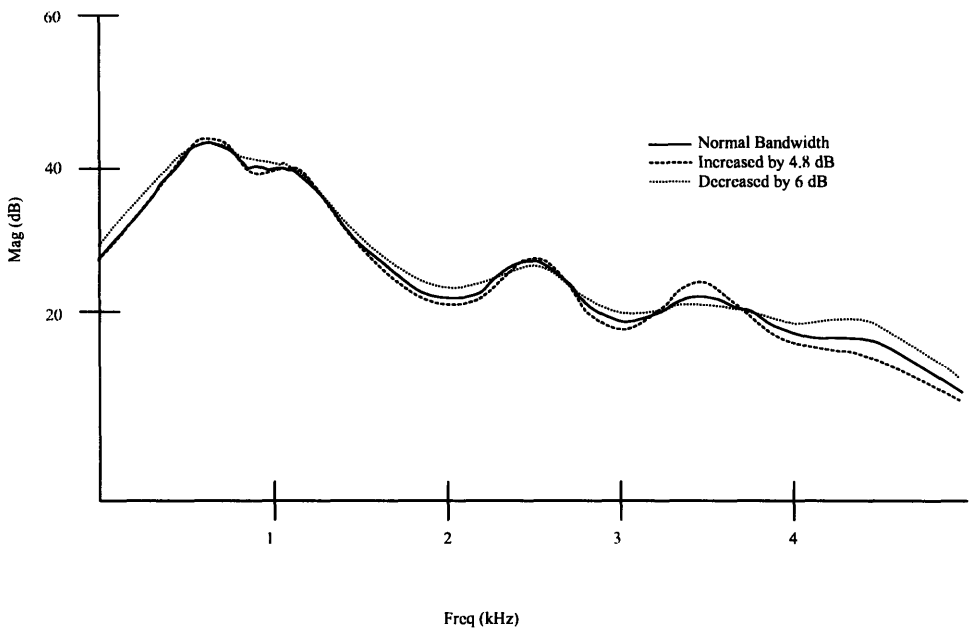


Figure A-4: Speech spectrum of the /a/ in the utterance /a//b//a/. The formant bandwidth is altered by -6dB and 4.8dB.

vary by more than 1 dB. The magnitude of the spectrum for the normal, -50%, and 175% cases are overlayed in Figure A-4. There are two main differences that are seen. First the amplitude between the formants are increased for broader bandwidths and decreased for narrower bandwidths. Second, the higher order formants, F3 and F4, appear to be more distorted for the broader bandwidths. These results are similar to the effects seen in the envelope reconstruction in the sinusoidal low-rate coder.

The results for the experiment showed that the DL is near $\pm$4dB for the formant bandwidths. The judgment graph was steeper for broader bandwidths. A finer resolution between -4 and 4 dB changes should be performed in the future to determine a better curve for the noticeable as noticeable as the F2 amplitude. To make sure that the listeners were listening to just the change in the bandwidth and there wasn't a side effect of altering F2 amplitude, the spectra of the synthesized speech was checked to make sure the F2 amplitudes were never more than 1 dB different from the original. The results are attributed primarily to the changes in formant bandwidths.

The Utterance

It would have been useful to determine which formant bandwidths contributed the most to the altered signal. It seems that since the first formant affects the remaining formants, it's bandwidth should be the most important in quality of the synthesized speech. Some utterances were generated where only the first or the third formant bandwidth was altered. A quick review of these utterances resulted in no noticeable difference between varying the third formant bandwidth and not varying any bandwidths was was noticed. A more complete set of experiment on individual formant bandwidths needs to be conducted.

The utterance /a b a/ was a good utterance to use because of the /a/. The first two formants of /a/ are very close to each other. This means that the energy between the formants will rise much more rapidly than for any other vowel when the bandwidths are broadened, and will fall much quicker for narrowing of the bandwidths.

A.3 Conclusion

Just-noticeable experiments were performed to determine the formant bandwidth variations that were perceived as "different." Another set of tests were done where the amplitude of F1 was inadvertently increased by broadening F1. The results showed that formant amplitude changes masked bandwidth changes because the DL was larger for the experiment where both amplitude and bandwidth were changed than for the experiment where just the bandwidth was changed. Changes in formant bandwidths by about 3dB or by about 40% are noticeable about 50% of the time. When the amplitude masked the bandwidth changes, the noticeable difference was closer to 2 dB. Small changes in bandwidth of about 10% to 20% will also be noticeable about 25% of the time.

These results show that the auditory system is able to detect small changes in bandwidth. But the criterion used by the listeners to judge the difference may not be the same for all the listeners. This creates a flaw in the analysis of the results. Overlaying spectrum plots (shown in Figure A-4) shows that there are large changes in the shapes and amplitudes of the 3rd and 4th formant peaks. The listeners could also be cuing on the transitions between the /b/ and the /a/. Since the reconstruction of the spline interpolated spectral envelope in the STC often broadens the first formant by more than the 4 dB seen in these experiments. The auditory system is probably using additional cues, such as additional phonemes, or transitions, to

correct for the perception of the broadened formant. The attention paid to the broadened formant may not be the key element in perceiving clarity, but may be a contributing factor to finer quality issues.

Appendix B

$1/f$ Property of the Spline Envelope

The cepstrum of the speech signal is written as the inverse Fourier transform of the log magnitude of the frequency response:

$$c_n = \frac{1}{\pi}\int_0^{\pi} \log A_s(\omega)\cos(n\omega)d\omega \qquad n = 0, 1, \ldots. \tag{B.1}$$

Cepstral coefficients c_m have certain general properties [26]. The main property relavent to this thesis is that the cepstrum decays at least as fast as $1/n$ [26]. Therefore, the log magnitude of the Fourier transform of speech has a $1/n$ like cepstrum.

If instead of taking the inverse Fourier transform of the log magnitude produces a $1/n$ signal, then the Fourier transform of the log magnitude results in a spectrum that is $1/f$. The discrete Fourier transform (DFT) is as follows:

$$y'[n] = \sum_{k=0}^{N-1} x'[k]e^{-j(2\pi/N)kn} \tag{B.2}$$

and the inverse DFT is

$$y[n] = \frac{1}{N}\sum_{k=0}^{N-1} x[k]e^{j(2\pi/N)kn} \tag{B.3}$$

The two equations are similar except for a factor of $1/N$, and the negative j in the exponential. For $x'[n]$ real and even, $y'[n]$ will be real and even. The same holds for $x[n]$ and $y[n]$. If $x'[n] = x[n]$ and it is real and even, then $y[n] = \frac{1}{N}y'[n]$.

The inverse Fourier transform of the log magnitude is similar to the Fourier transform of the log magnitude except for a $1/N$ factor. The log magnitude ($\log A_s(\omega)$) is real and even,

so the inverse DFT, will yield the complex cepstrum which is also real and even. The DFT of $\log A_s(\omega)$ will be the same except for a factor of N which is the length of the signal. If the log magnitude response is the signal being analyzed, it will have a $1/f$ response. This means that the speech envelope is a $1/f$ signal.

Appendix C

Matlab Code

This appendix contains the Matlab code used for testing out the wavelet algorithms from Chapter 4. The C code used in the WSTC system was a direct translation of the following Matlab code.

C.1 Convolution-Based Algorithm

```
\begin{small}
function [approx,wave_cof,len] = con_ana(x,h,g,scal);
%
% Analysis section of the wavelet transform using the
% convolution method
%
%Input: x -- signal
% h -- low pass filter
%       g -- high pass fitler
%       scal -- number of scales
%
%
%Output: approx -- final approximation signal at scale n
%  wave_cof -- wavelet coefficients
%  len - length of coefficients at each scale
len=[];
wave_cof=[];
approx=x;
for i=1:scal,
t=conv(g,approx);
approx=conv(h,approx);
t=t(1:2:length(t));
```

```
approx=approx(1:2:length(approx));
len=[len length(t)];
wave_cof=[wave_cof t];
end;


function out = con_syn(wv,apr,len,h,g,scal,fin,m);
%
% Synthesis section of the wavelet transform using the
% symmetric method. The synthesized approximation at each scale
% is truncated by length(h) in the front and end.
%
%Input: wv -- wavelet coefficients from con_ana
% apr - final approximation at scale scal
% len - length vector of length at each scale
% h -- low pass filter
%       g -- high pass fitler
%       scal -- number of scales the signal was decomposed
% fin - final scale in which the wavelet coefficients
% are to be use, after which they are zeroed.
% m -- scaling factor for wavelet (Usually 2 for B-L)
%
%
%Output: out -- Reconstructed signal
%  wave_cof -- wavelet coefficients
%  len - length of coefficients at each scale
ap=apr;
for i=scal-1:-1:0,
if i>=fin
t=wv(sum(len(1:i))+1:sum(len(1:i+1)));
tr=conv(g,m*srexpand(t,2));
ap=tr+conv(h,m*srexpand(ap,2));
else
ap=conv(h,m*srexpand(ap,2));
end
if i==0
 out=ap(length(h):512+length(h)-1);
else
ap=ap(length(h):len(i)+length(h)-1);
end
length(ap);
end
\end{small}
```

C.2 Periodic WT Functions

```
function [approx,wave_cof,len] = per_ana(x,h,g,scal,n);
%
%
% Analysis section of the wavelet transform using the
% convolution method
%
%Input: x -- signal
% h -- low pass filter
%       g -- high pass fitler
%       scal -- number of scales
% n -- offset of the basis frunction from zero
%
%Output: approx -- final approximation signal at scale n
%  wave_cof -- wavelet coefficients
%  len - length of coefficients at each scale

len=[];
wave_cof=[];
approx=x;
for i=1:scal,
t=cconv(g,approx,n(2)); % Circular Convolution
approx=cconv(h,approx,n(1)); % Circular Convolution
t=t(1:2:length(t)); % Downsample
approx=approx(1:2:length(approx)); % Downsample
len=[len length(t)];
wave_cof=[wave_cof t];
end;

function out = sym_syn(wv,apr,len,h,g,scal,fin,m,n);
%
% Synthesis section of the wavelet transform using the
% symmetric method. The synthesized approximation at each scale
% is truncated by length(h) in the front and end.
%
%Input: wv -- wavelet coefficients from con_ana
% apr - final approximation at scale scal
% len - length vector of length at each scale
% h -- low pass filter
%       g -- high pass fitler
%       scal -- number of scales the signal was decomposed
% fin - final scale in which the wavelet coefficients
% are to be use, after which they are zeroed.
% m -- scaling factor for wavelet (Usually 2 for B-L)
% n -- offset of the basis frunction from zero
```

```
%
%Output: out -- Reconstructed signal

ap=apr;
for i=scal-1:-1:0,
if i>=fin
t=wv(sum(len(1:i))+1:sum(len(1:i+1)));
tr=cconv(g,m.*srexpand(t,2),n(4)); % Upsample and Circ Conv
ap=tr+cconv(h,m.*srexpand(ap,2),n(3)); % Upsample and Circ Conv
else
ap=cconv(h,m.*srexpand(ap,2),n(3));
end
end
out=ap;
```

C.3 Periodic-Symmetic Functions

```
function [approx,wave_cof,len] = wstc_ana(x,h,g,scal,n);
%
% Analysis section of the WSTC system.  The algorithm used is
% the symmetric-periodic algorithm where the signal is made
% symmetric and then a circular convolution is used.
%
%Input: x -- Spline Envelope (input)
% h -- low pass filter
%       g -- high pass fitler
%       scal -- number of scales of decomposition
%       n -- offset of the filter from zero
%
%Output: approx -- final approximation signal at scale n
%  wave_cof -- wavelet coefficients
%  len - length of coefficients at each scale

len=[];
wave_cof=[];
approx=[fliplr(x) x];  % Make the input symmetic about 0
for i=1:scal,
t=cconv(g,approx,n(2)); % Circular Convolution
approx=cconv(h,approx,n(1)); % Circular Convolution
t=t(1:2:(length(t)/2)); % Downsample and take 1/2 Coefficients
approx=approx(1:2:length(approx)); % Downsample
len=[len length(t)];
wave_cof=[wave_cof t];
end;
```

```
function out = wstc_syn(wv,apr,len,h,g,scal,m,n);
%
% Synthesis section of the WSTC using the symmetric-periodic
% algorithm.  The wavelet coefficients are made symmetric and
% then a circular convolution is used.
%
%Input: wv -- wavelet coefficients from wtsc_ana
% apr - final approximation at scale scal
% len - length vector of length at each scale
% h -- low pass filter
%       g -- high pass fitler
%       scal -- number of scales the signal was decomposed
% m -- scaling factor for wavelet (Usually 2 for B-L)
%       n -- offset of the filter from zero
%
%
%Output: out -- Reconstructed signal
%
ap=apr;
for i=scal-1:-1:0,
t=wv(sum(len(1:i))+1:sum(len(1:i+1)));
t=[t fliplr(t)]; % Make Wavelet Coefficient Symmetic
tr=cconv(g,m.*srexpand(t,2),n(4));
ap=tr+cconv(h,m.*srexpand(ap,2),n(3));
end
out=ap(1:length(ap)/2);
```

C.4 WSTC System

```
load data/spec.dat -mat % Load envelope Data
load data/b_l.dat -mat % Load Wavelet Basis Functions
off=11; % Truncate Wavelet basis to
b_l_set; % N=25
m=[];
test=[];
wavelet=[];
ap=[];
for z=1:90, % Sample 90 frames
set_quan; % Set the Quantizers
input=inp((z*512)+1:((z+1)*512)); % Get envelope information
[approx,wavel,len]=wstc_ana(input,h,g,9,n);
plotwstc; % Ploting Routines
wquant; % Quantize Coefficients
t=wavelet(length(wavelet));
t=cconv(g,2.*srexpand(t,2),n(4)); % Decompose 10th Scale
```

```
approx=t+cconv(h,2.*srexpand(ap,2),n(3));
sigout=wstc_syn(wavelet,approx,len,h1,g1,9,2,n);
%
% Sigout is the reconstructed signal
%
end

wquant.m
% Quantizer in the Analysis portion
%
% Set the initial 4 Scales to 0.015
%
wavelet=ones(1,sum(len(1:4)))*.015;
%
% Logarithmic (1) Quantizer for Scale 5
%
wavelet=[wavelet quantize(wavel(1+sum(len(1:4)):sum(len(1:5))),log1)];
%
% Logarithmic (2) Quantizer for Scale 6 and 7
%
wavelet=[wavelet quantize(wavel(1+sum(len(1:5)):sum(len(1:7))),log2)];
%
% Inverse Log (2) Quantizer for Scale 8
%
wavelet=[wavelet quantize(wavel(1+sum(len(1:7)):sum(len(1:8))),ilog2)];
%
% Uniform Quantizer for Scale  9
%
wavelet=[wavelet quantize(wavel(1+sum(len(1:8)):sum(len(1:9))),uni)];

t=cconv(g,approx,n(2)); % Decompose the approximation to scale 10
ap=cconv(h,approx,n(1));
t=t(1:2:(length(t)/2))
t=squant(t,(-3.5:.1875:.5)); % Quantize wavelet coeff from -3.5 to .5
wavelet=[wavelet t];
ap=ap(1:2:length(ap));
ap=squant(ap,(.1875:.1875:12)); % Quantize approx coeff from 0 to 12
```

C.5 Support Functions

```
function vec=ccconv(h,x,n)
% Cicrular Convolution h is the filter, x is the signal, n is the number
% of negative offset.

if length(x)>=length(h),
```

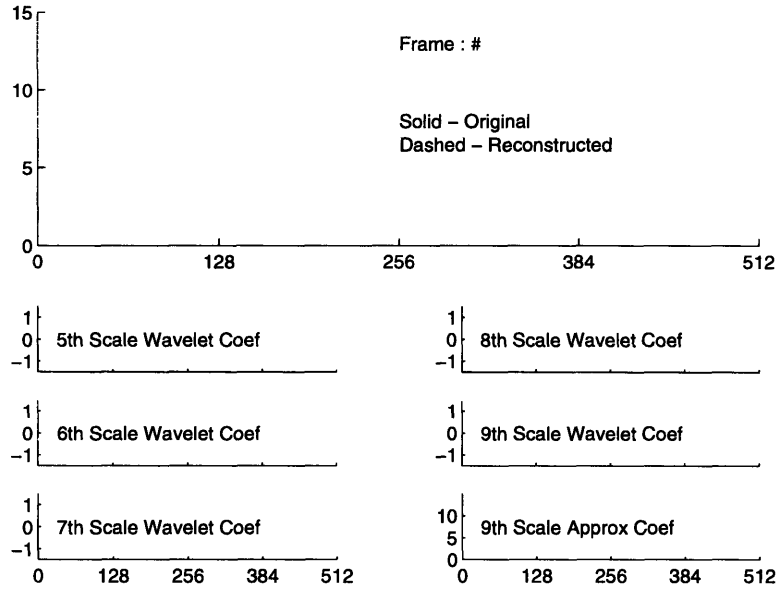
```
H=fft(h,length(x));
X=fft(x,length(x));
vec=real(ifft(H.*X,length(x)));
n=n-floor(n/length(vec))*length(vec);
if n>0,
vec=[vec(n+1:length(vec)) vec(1:n)];
end
if n<0,
vec=[vec(length(vec)-n:length(vec)) vec(1:length(vec)-n-1)]
end
end

if length(x)<length(h),
sj=n+1-length(h);
h=fliplr(h);
for i=0:length(x)-1
sum=0;
for j=sj:length(h)-1+sj
sum=sum+h(j-sj+1)*x(modulo(i+j+(10*length(x)),length(x))+1);
end
vec(i+1)=sum;
end
fliplr(h);
end

set_quant.m
%
% Setup the Quantizer Tables
%
i=exp(log(1.5)./2.7183);
i=i/8:i/8:i;
log1=i.^2.7183;

i=exp(log(1.5)./2);
i=i/8:i/8:i;
log2=i.^2;

uni=.1875:.1875:1.5;

i=exp(log(1.5)./(1/2.7183));
i=i/8:i/8:i;
ilog1=i.^(1/2.7183);

i=exp(log(1.5)./.85);
i=i/8:i/8:i;
ilog2=i.^.85;

function t=quantize(x,q);
```

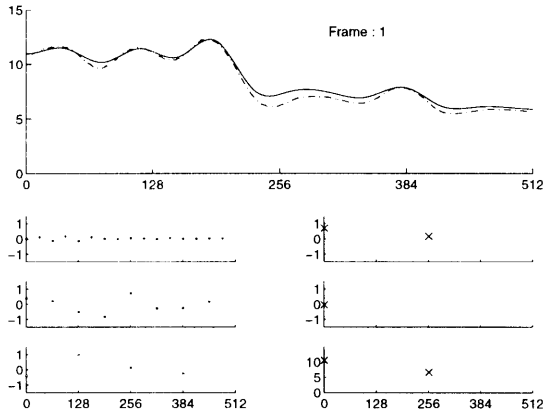
```
%
% Quantize the vector x using the vector q which is centered around 0
%

qlen=length(q);
t=zeros(size(x));
q=[0 q];
for i=2:qlen
t=t+((x>q(i-1))&(x<q(i)))*q(i);
t=t+((x<(-q(i-1)))&(x>(-q(i))))*(-q(i));
end
t=t+(x>q(qlen))*q(qlen+1);
t=t+((x<(-q(qlen)))*(-q(qlen+1)));
```

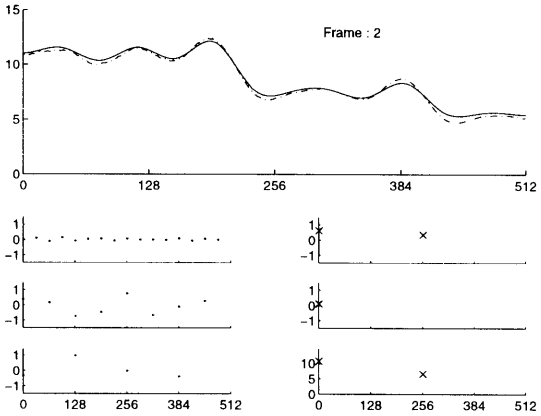
Appendix D

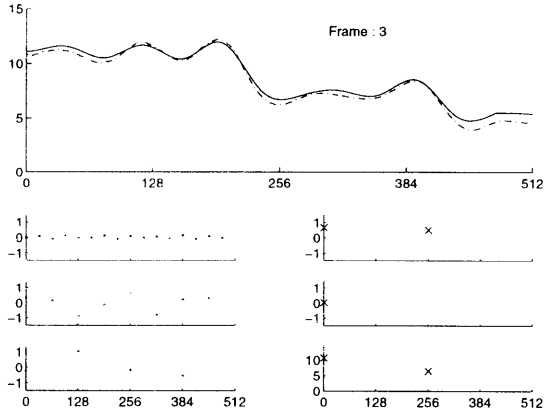
Raw Data

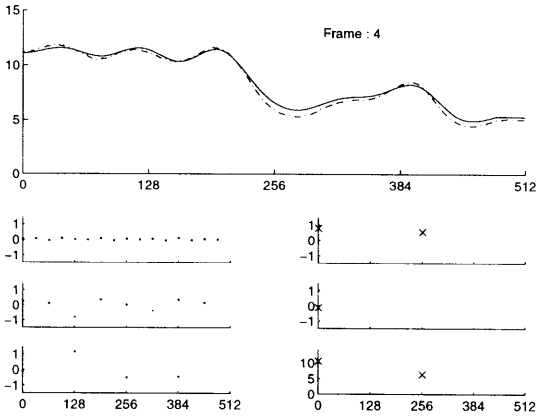
This appendix contains a sample set of frames which were decomposed and reconstructed using the WSTC. The original envelope is the solid line. The dashed line is the reconstructed envelope using the complete WSTC system. The unquantized wavelet coefficients are given below along with the approximation at the 9th scale. From the approximation at the 9th scale we can see the spectral tilt and the DC value of the enevelope. The graph below shows the labeling of the figures for the frames.

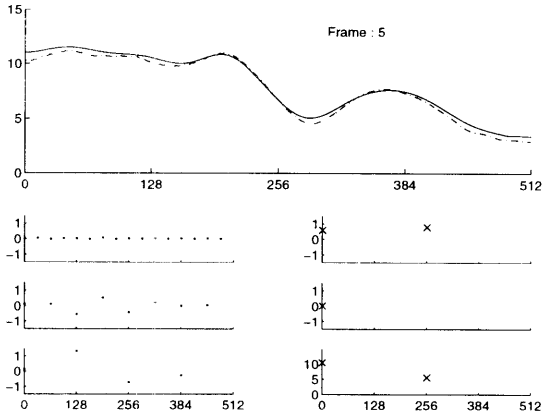


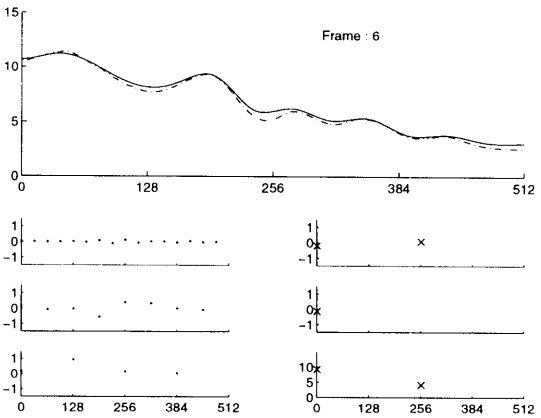

Frame : 1

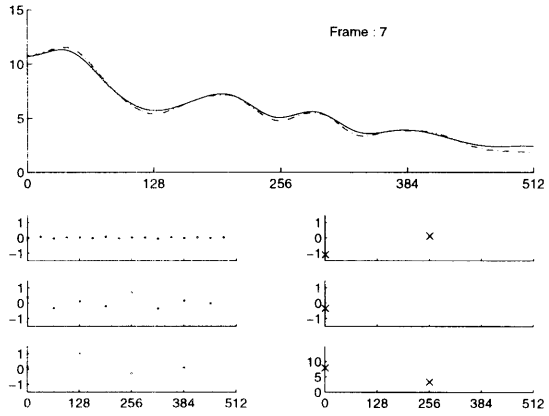

Frame : 2

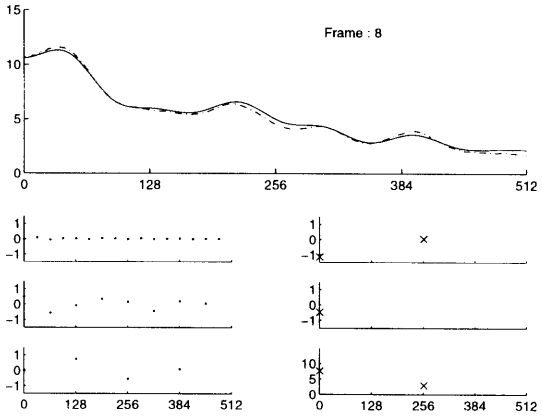

Frame : 3

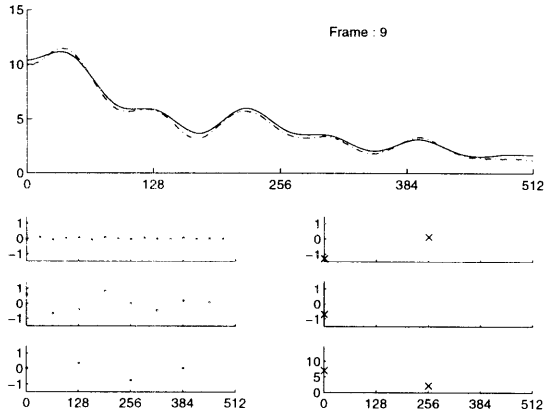

Frame : 4

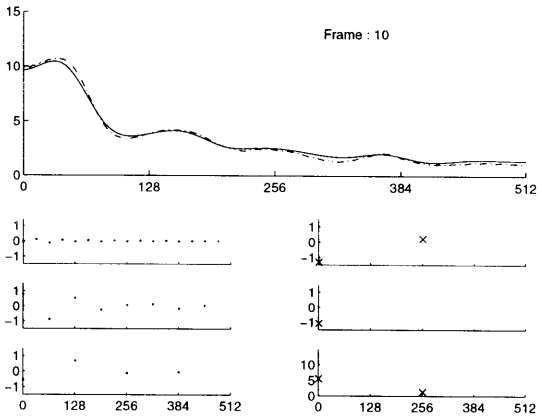

Frame : 5

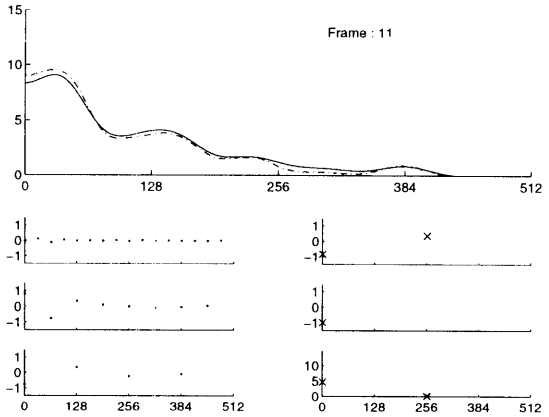

Frame : 6

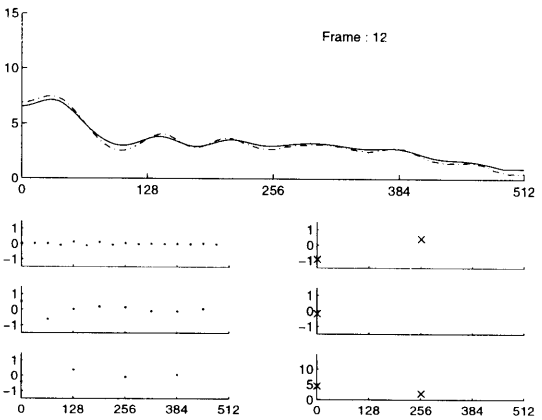

Frame : 7

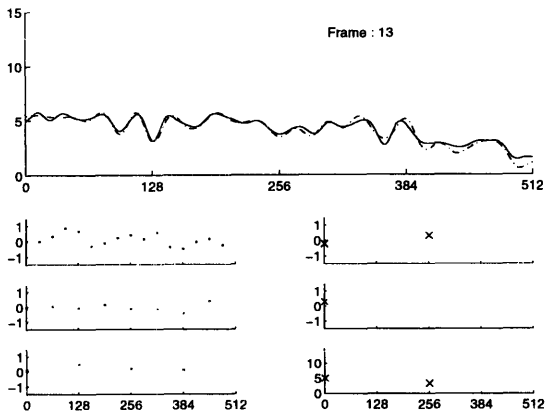

Frame : 8

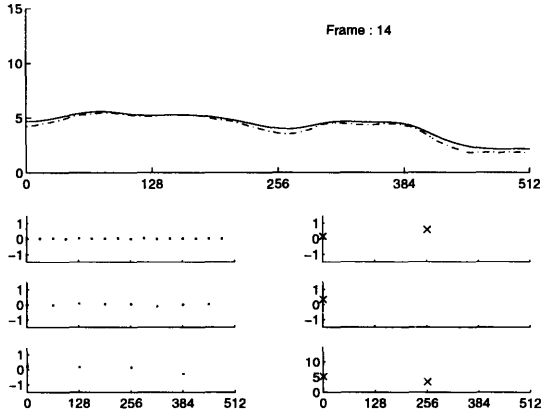

Frame : 9

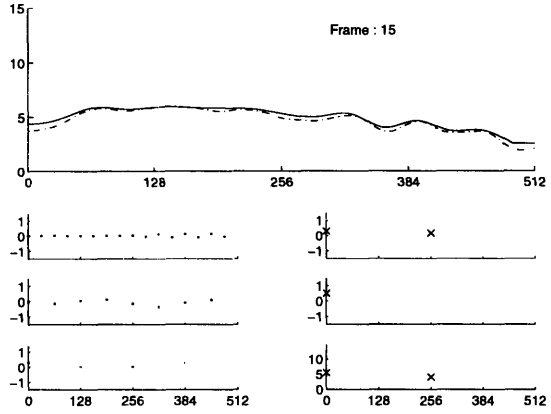

Frame : 10

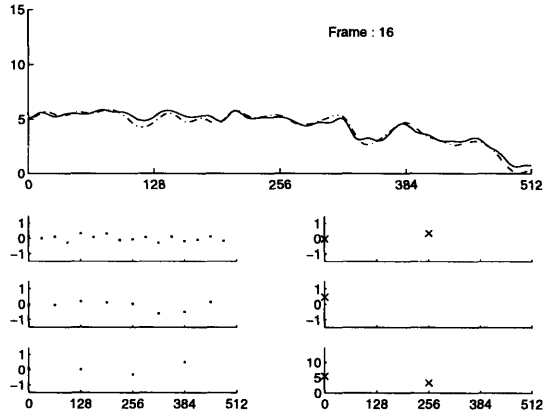

Frame : 11

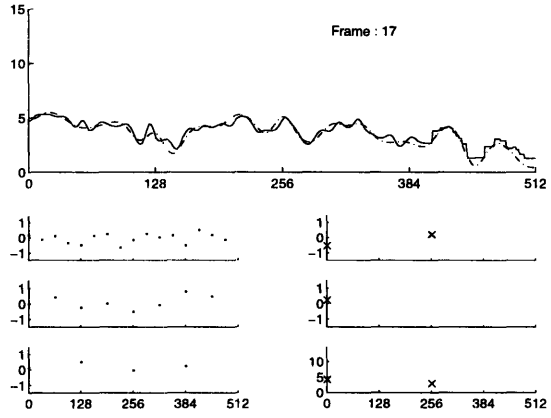

Frame : 12

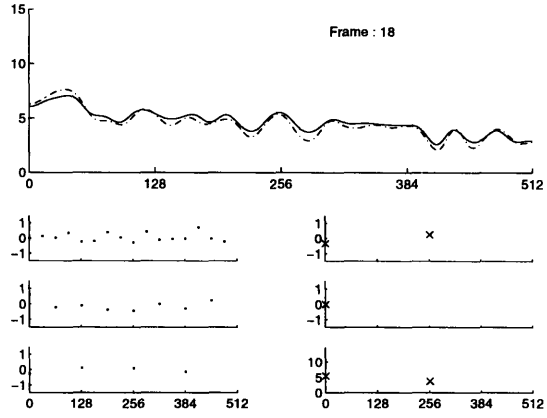

Frame : 13

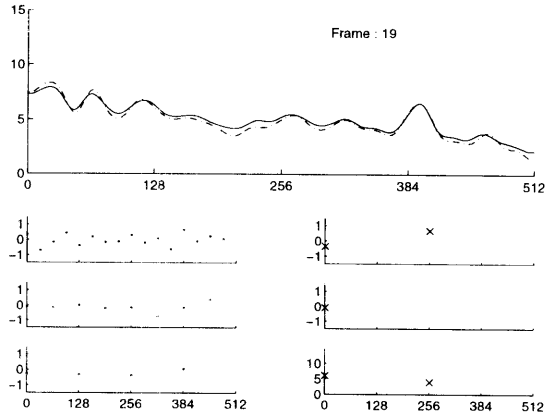

Frame : 14

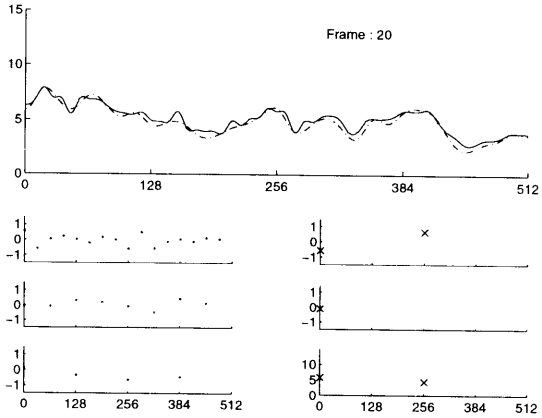

Frame : 15

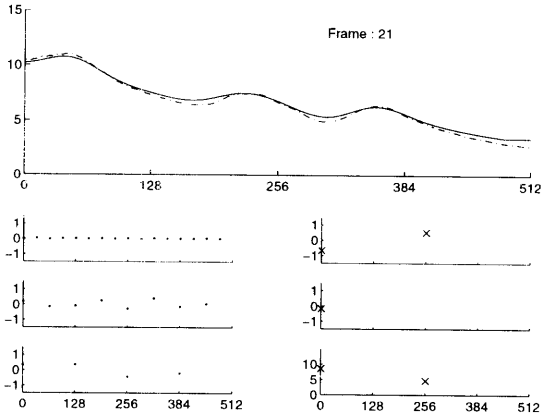

Frame : 16

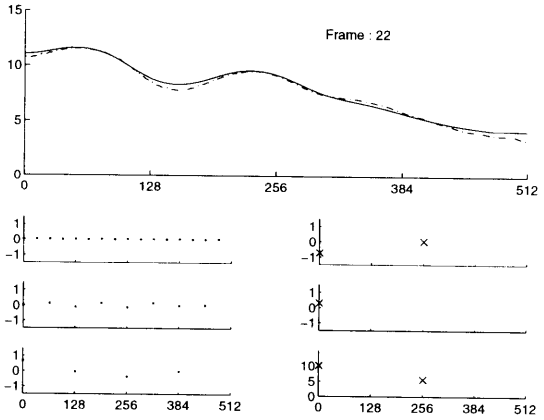

Frame : 17

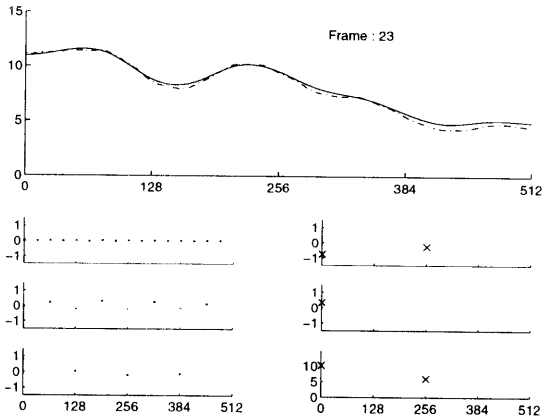

Frame : 18

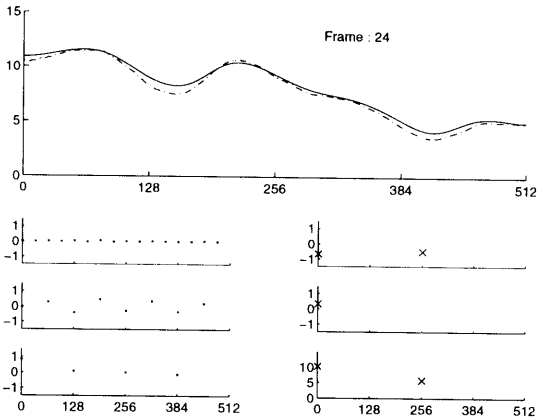

Frame : 19

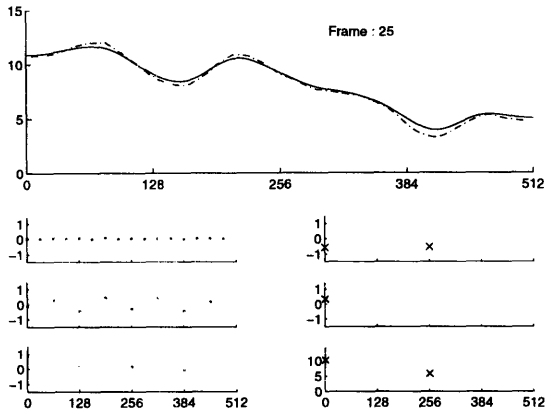

Frame : 20

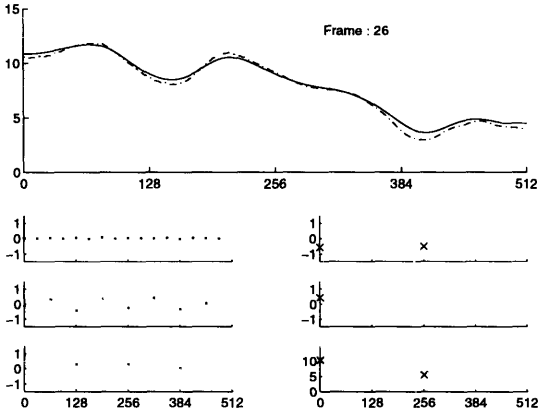

Frame : 21

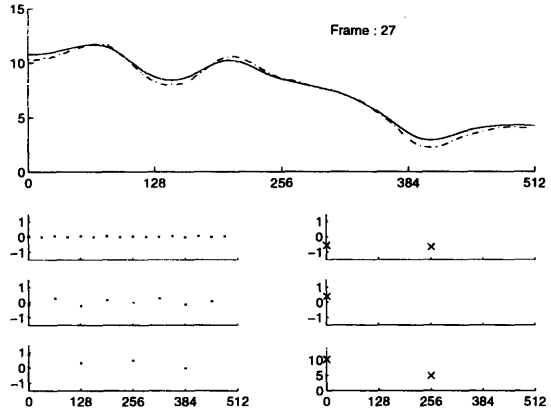

Frame : 22

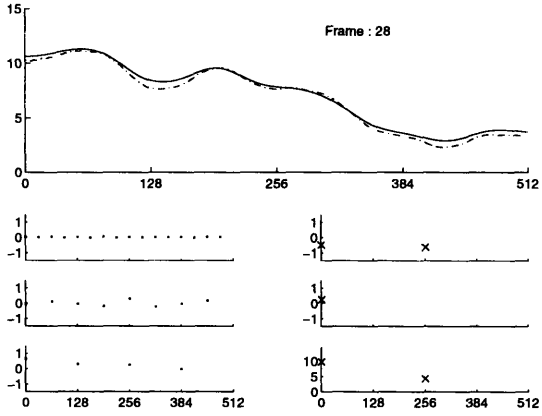

Frame : 23

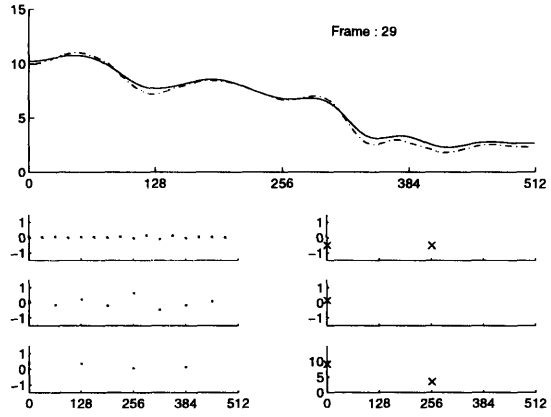

Frame : 24

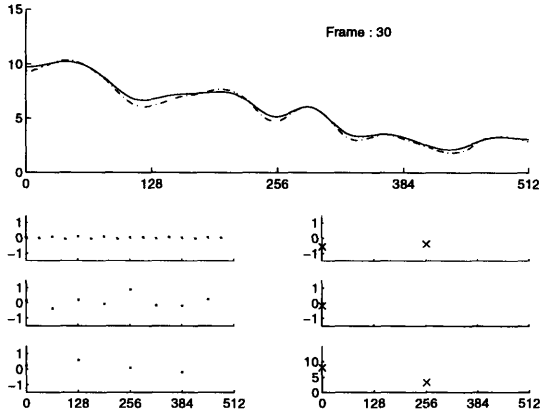

Frame : 25

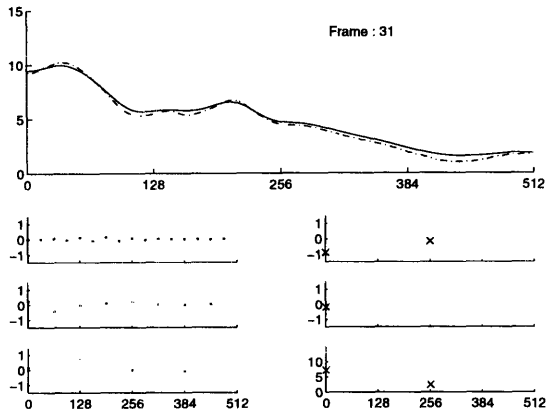

Frame : 26

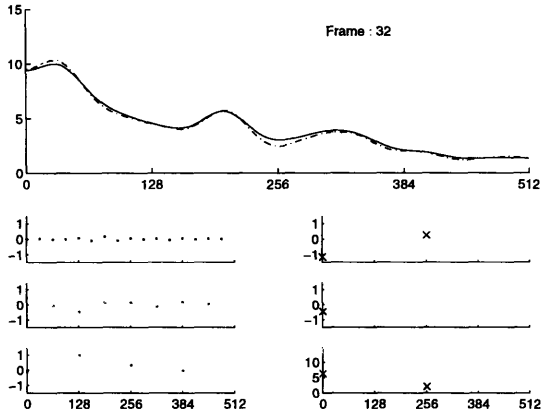

Frame : 27

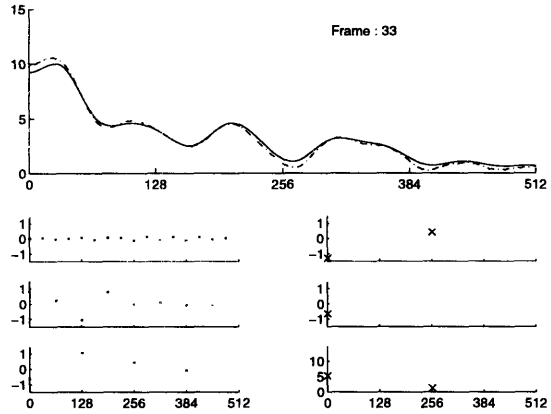

Frame : 28

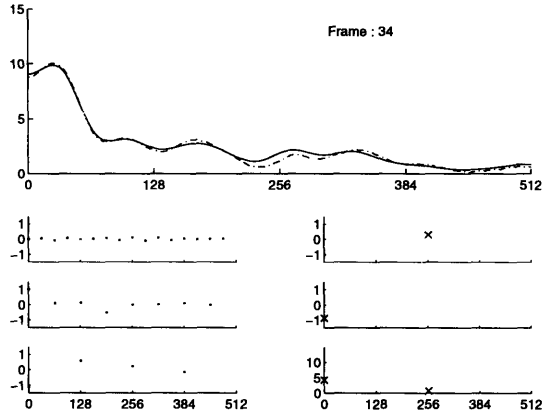

Frame : 29

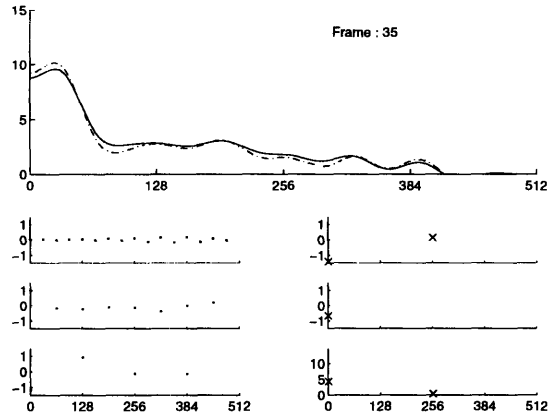

Frame : 30

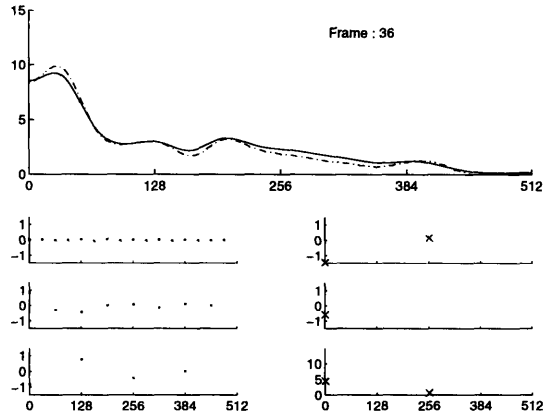

Frame : 31

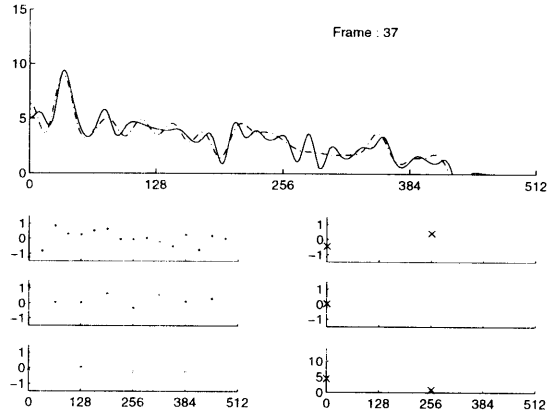

Frame : 32

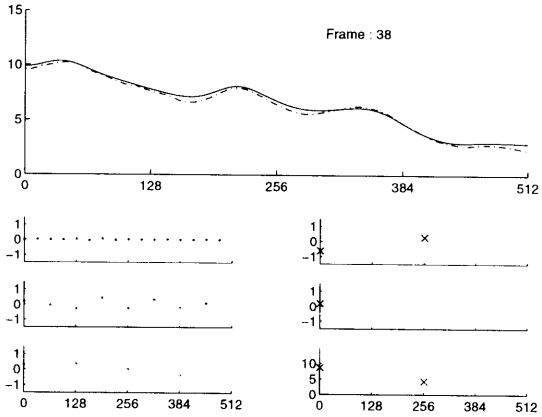

Frame : 33

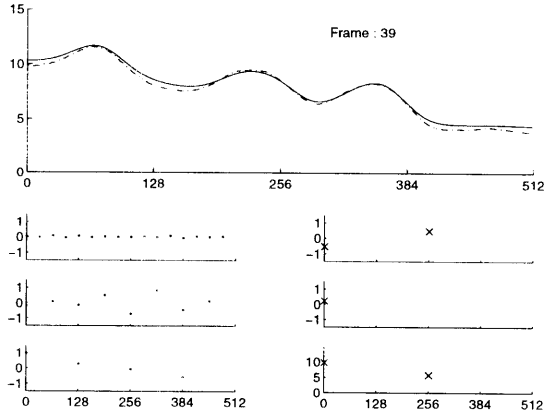

Frame : 34

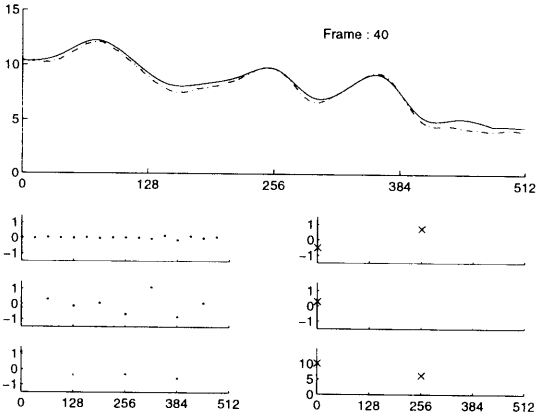

Frame : 35

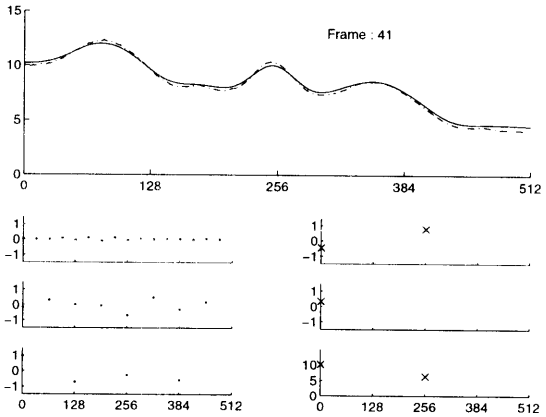

Frame : 36

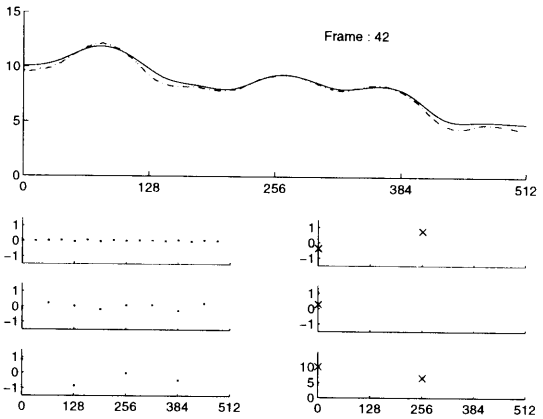

Frame : 37

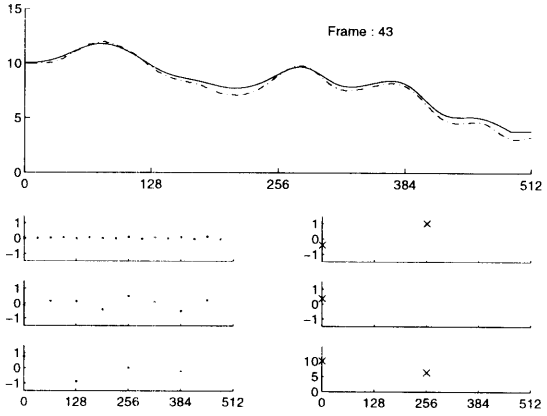

Frame : 38

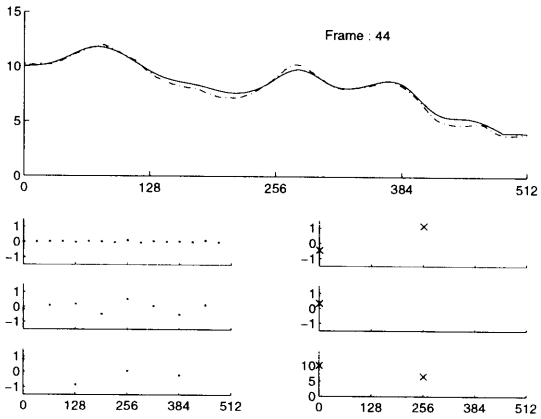

Frame : 39

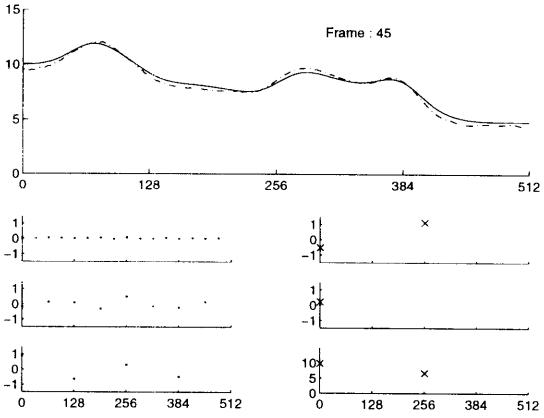

Frame : 40

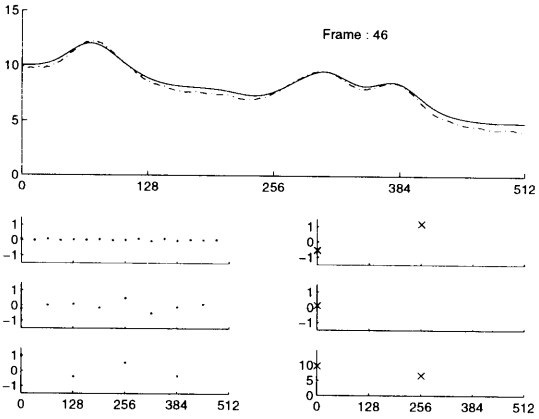

Frame : 41

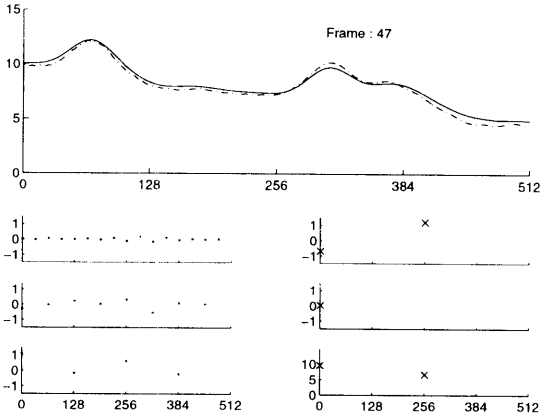

Frame : 42

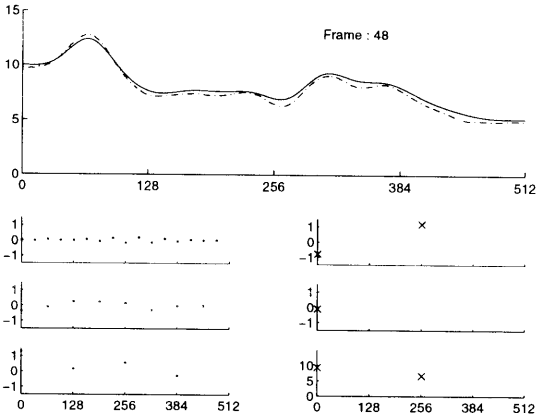

Frame : 43

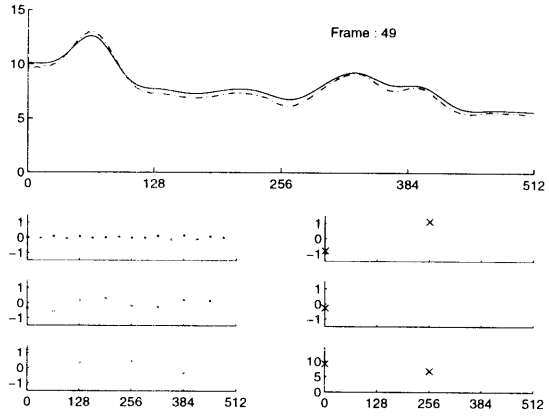

Frame : 44

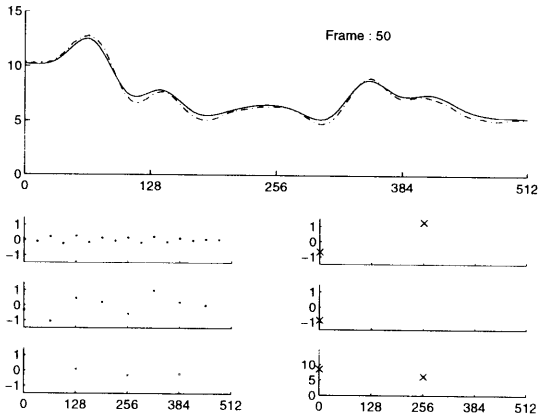

Frame : 45

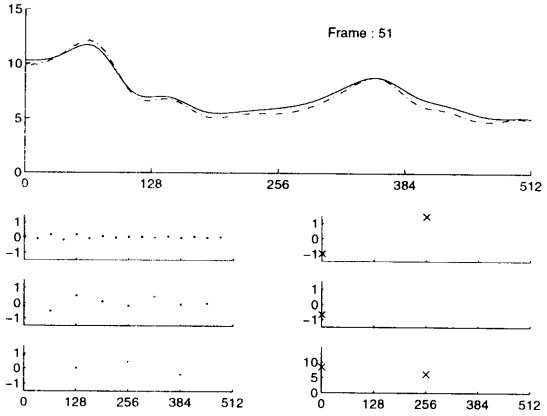

Frame : 46

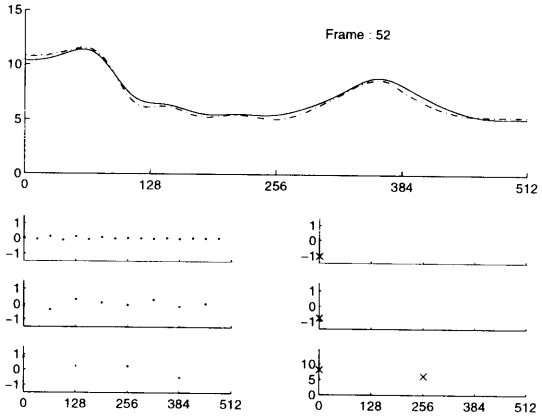

Frame : 47

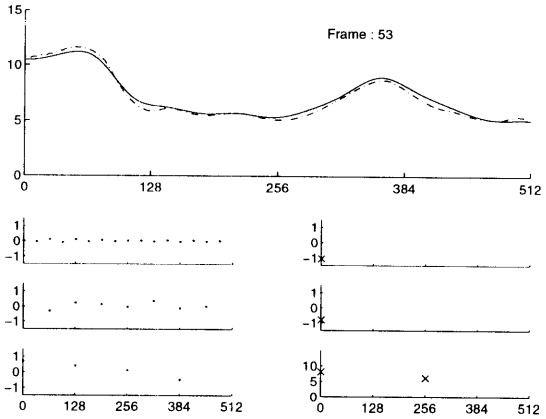

Frame : 48

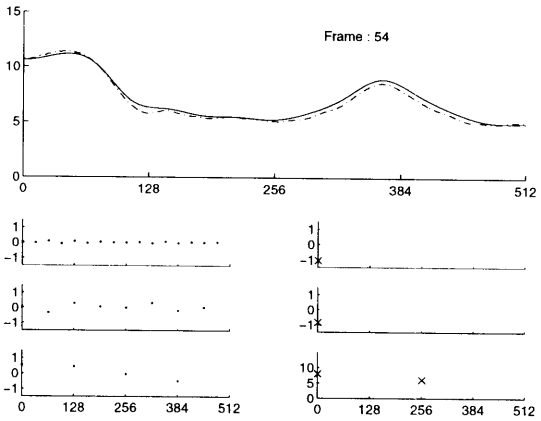

Frame : 49

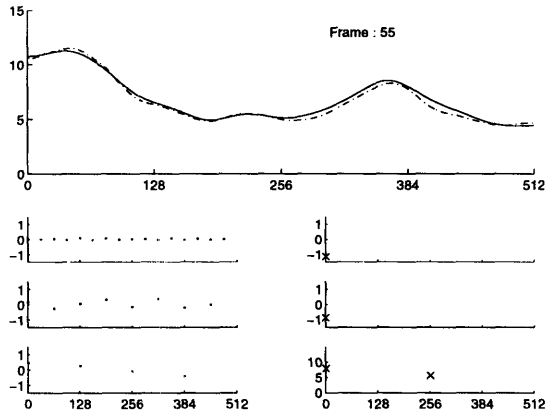

Frame : 50

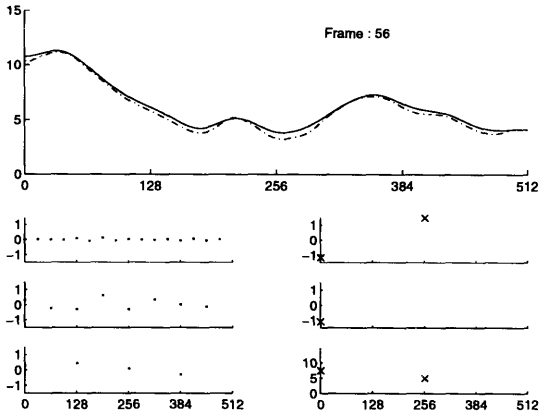

Frame : 51

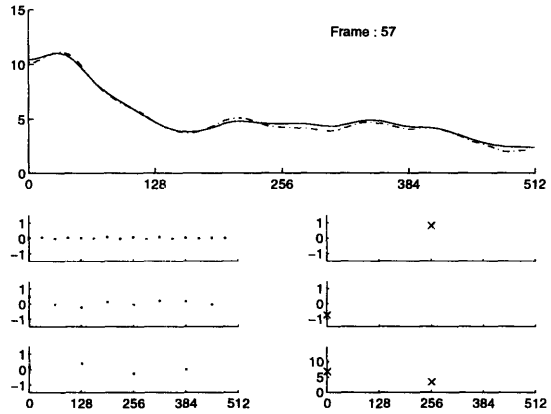

Frame : 52

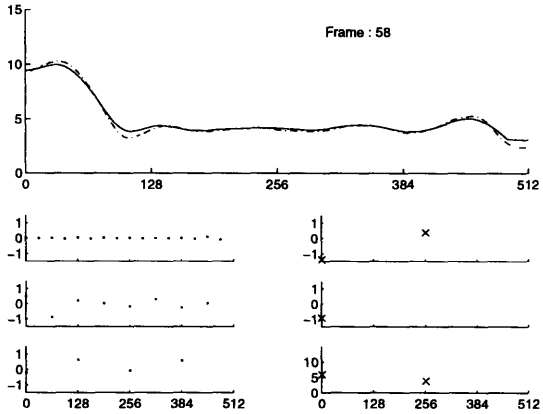

Frame : 53

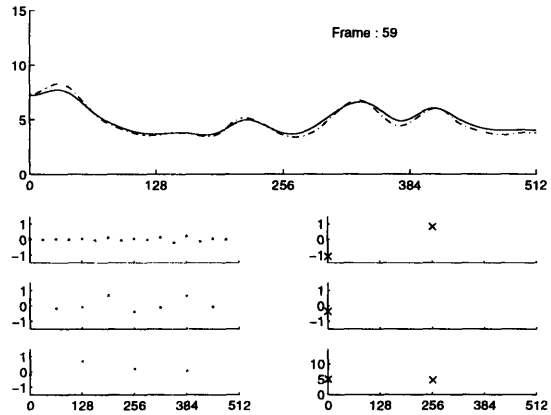

Frame : 54

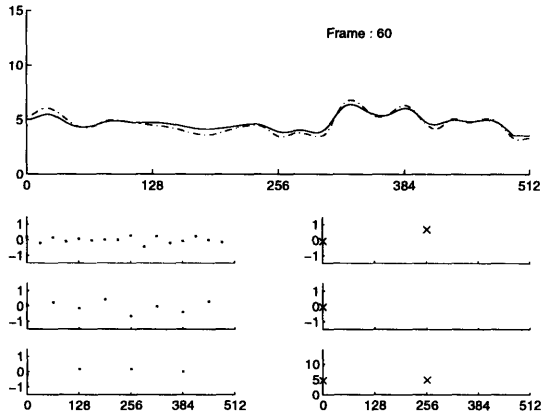

Frame : 55

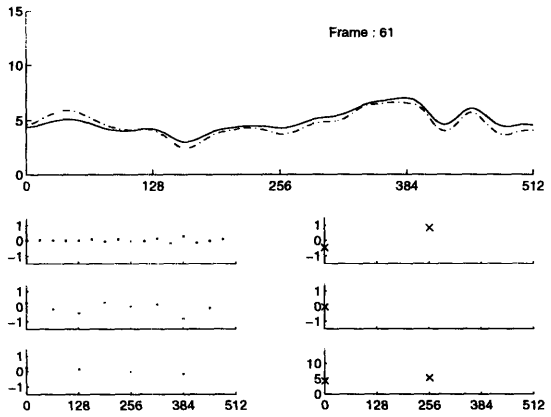

Frame : 56

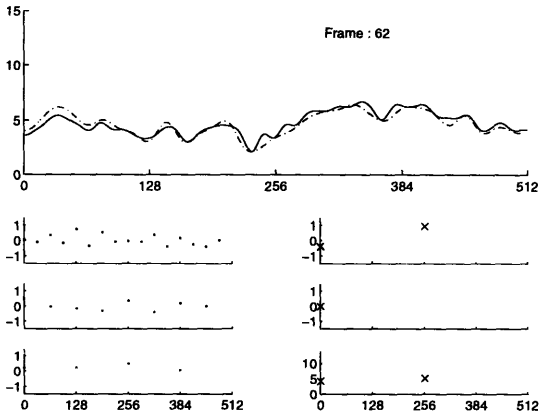

Frame : 57

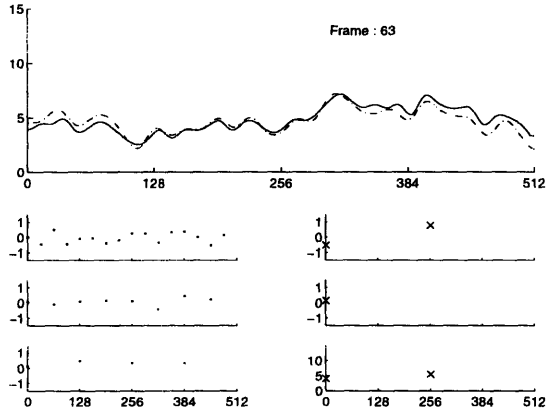

Frame : 58

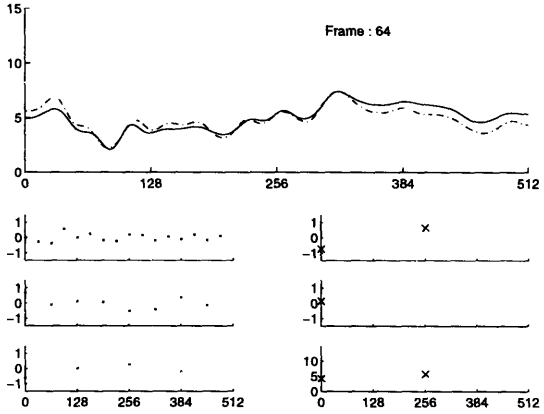

Frame : 59

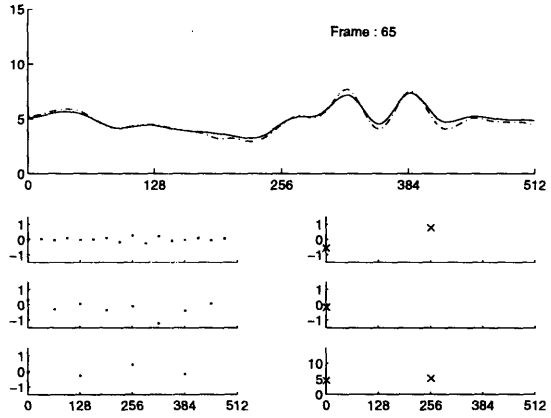

Frame : 60

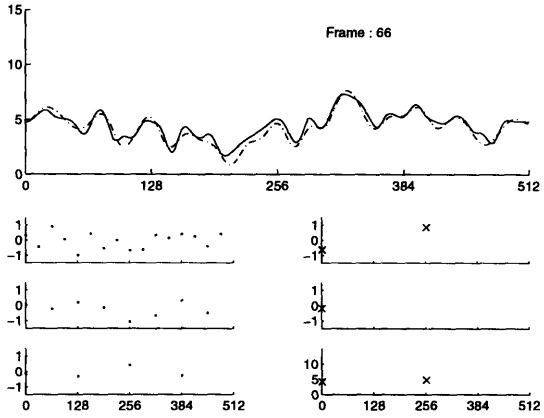

Frame : 61

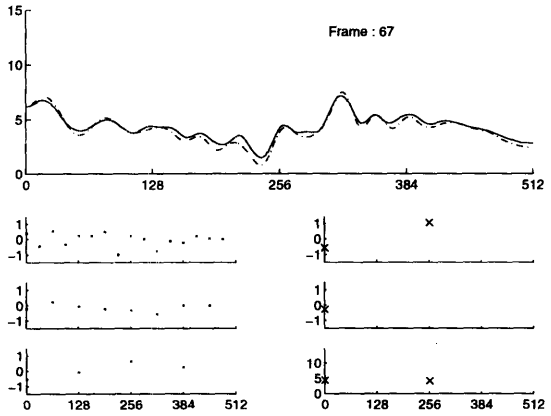

Frame : 62

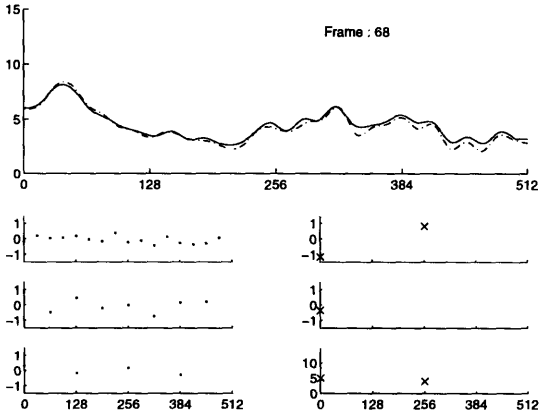

Frame : 63

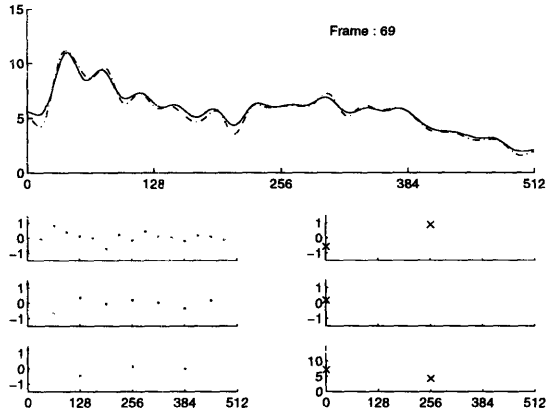

Frame : 64

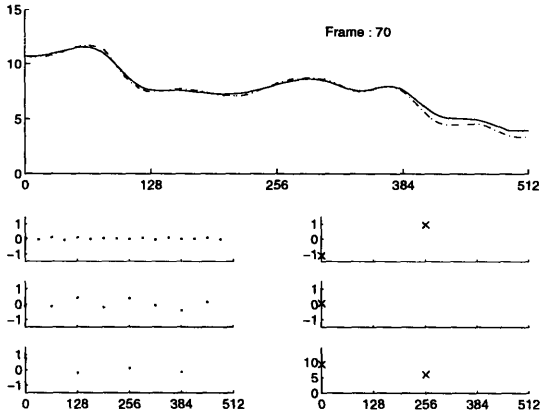

Frame : 65

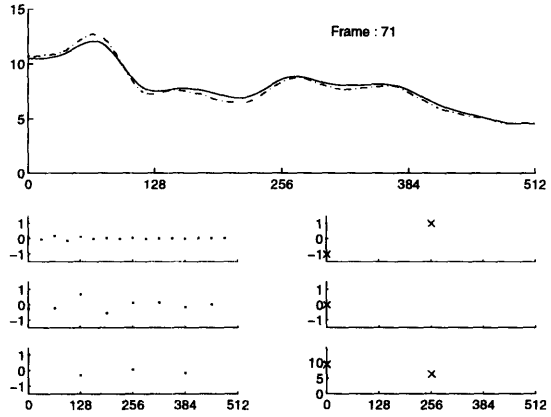

Frame : 66

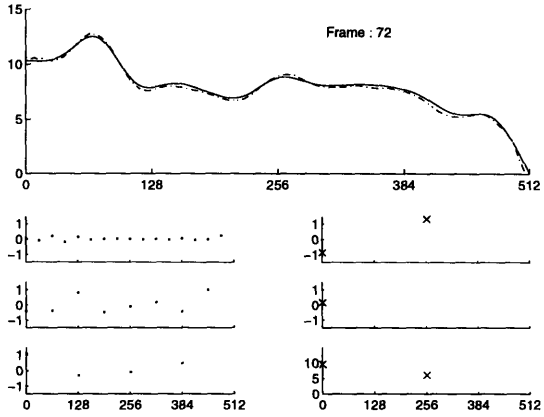

Frame : 67

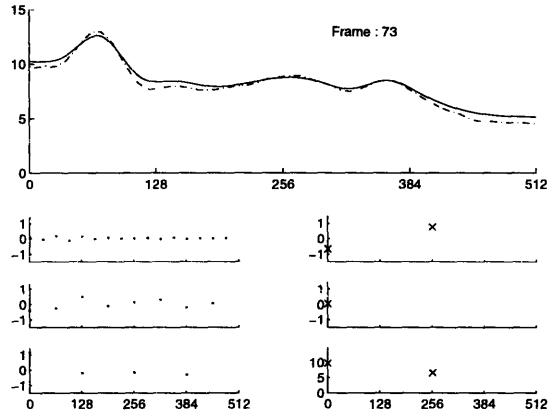

Frame : 68

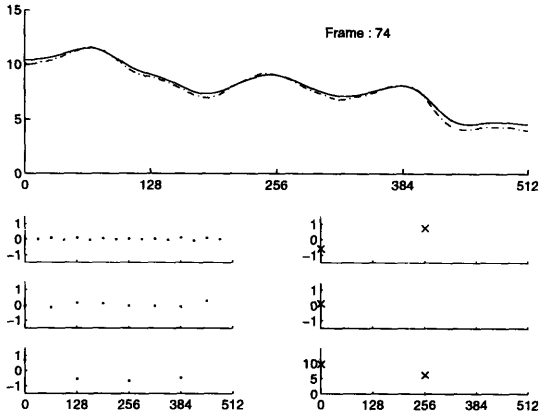

Frame : 69

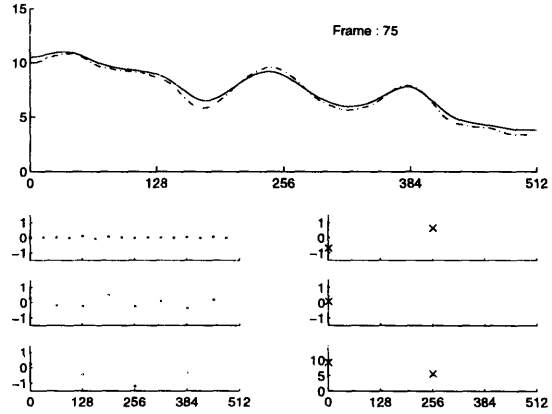

Frame : 70

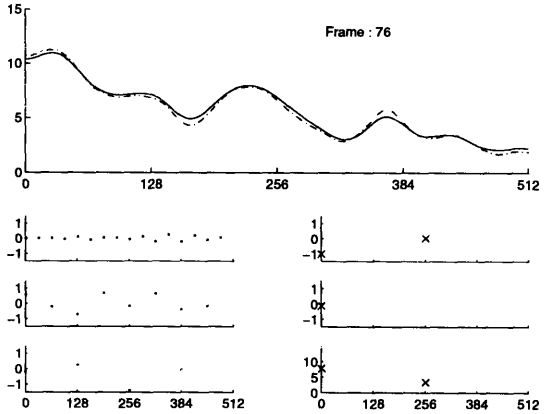

Frame : 71

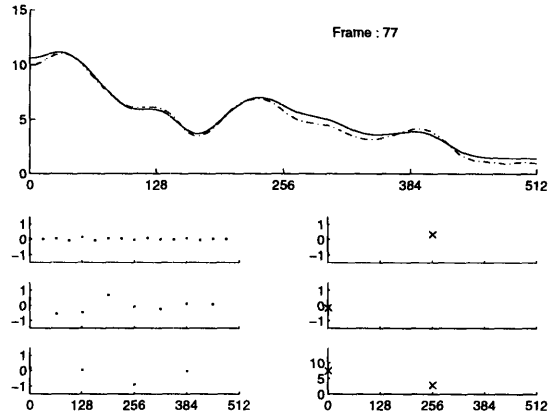

Frame : 72

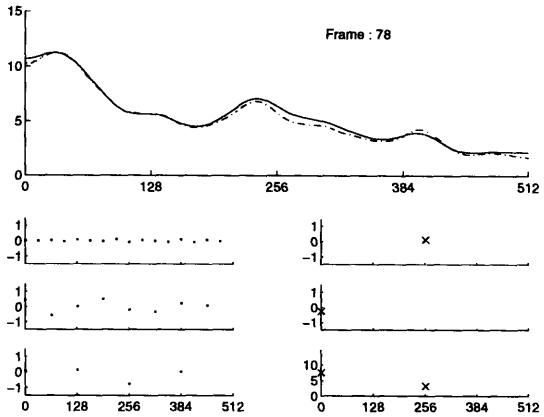

Frame : 73

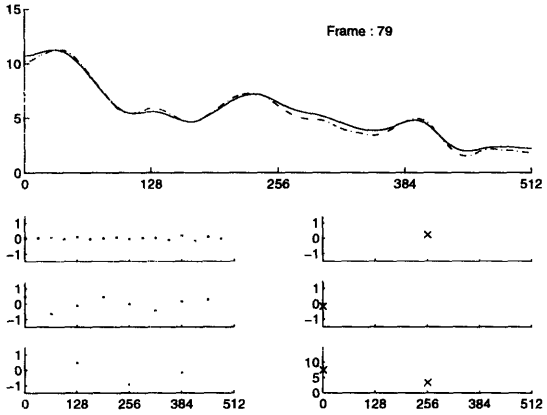

Frame : 74

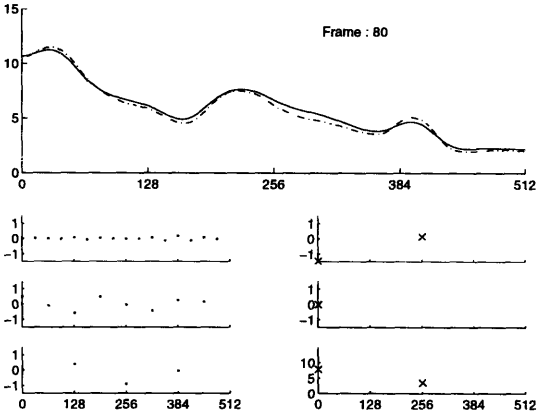

Frame : 75

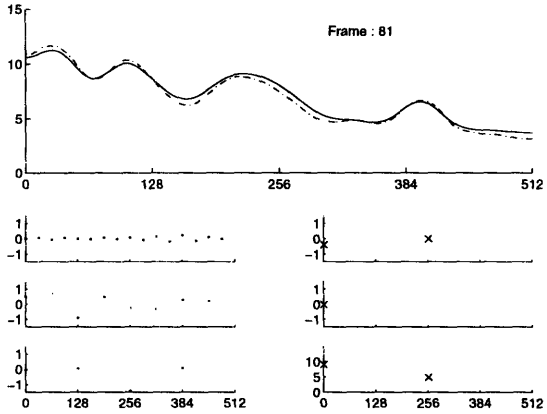

Frame : 76

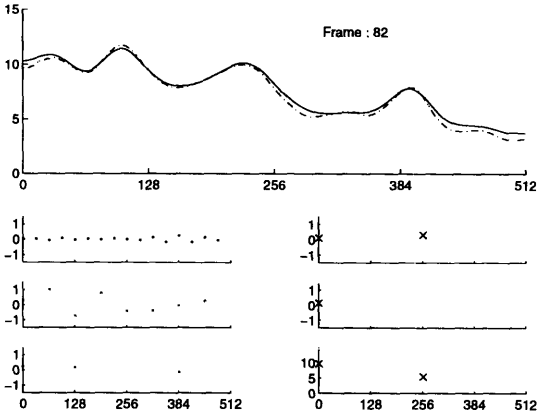

Frame : 77

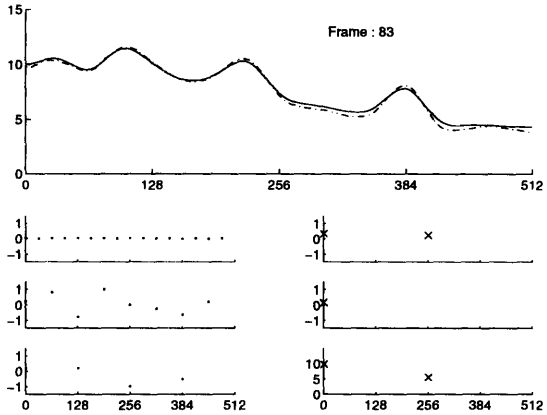

Frame : 78

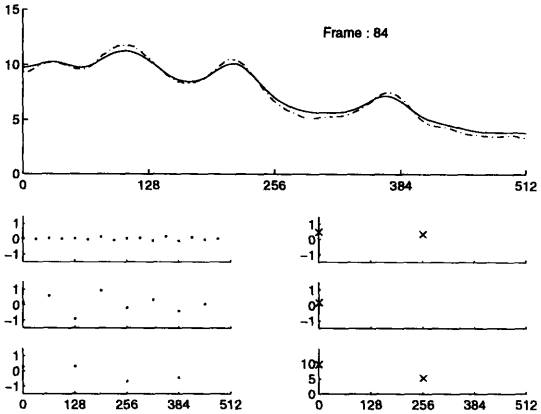

Frame : 79

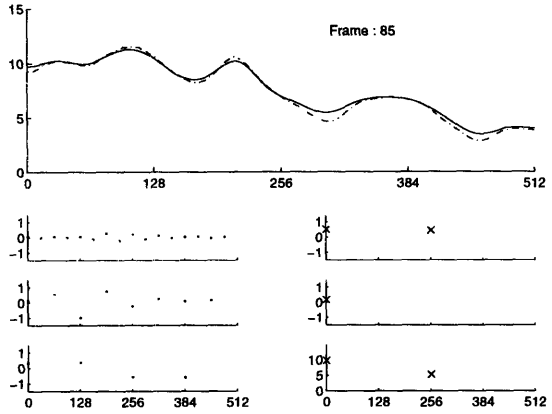

Frame : 80

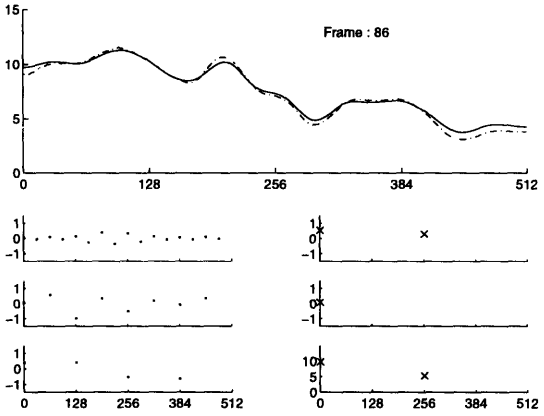

Frame : 81

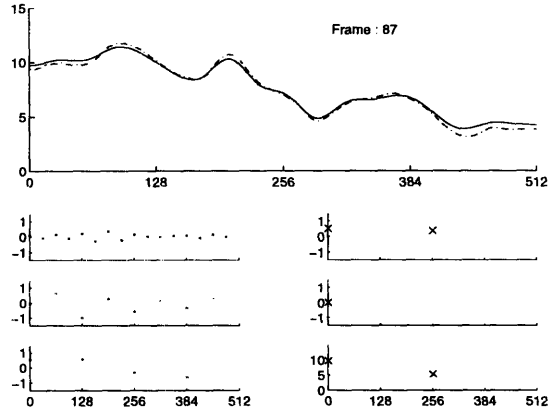

Frame : 82

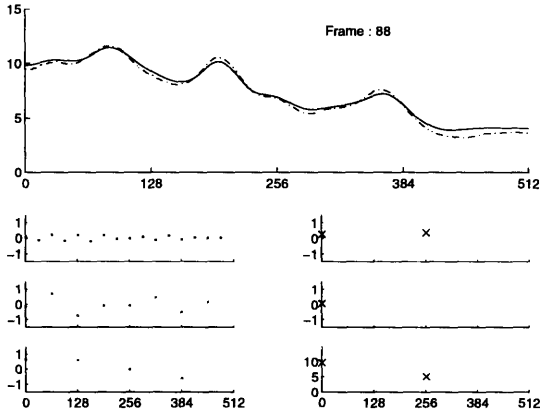

Frame : 83

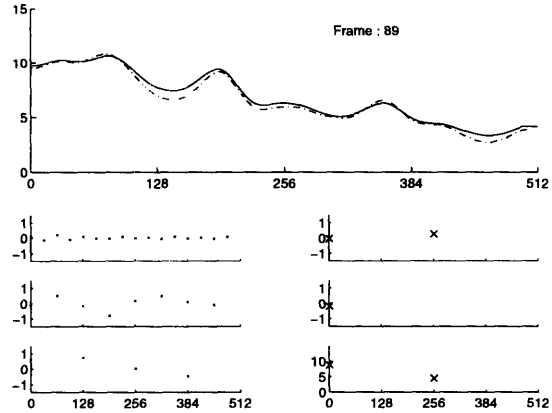

Frame : 84

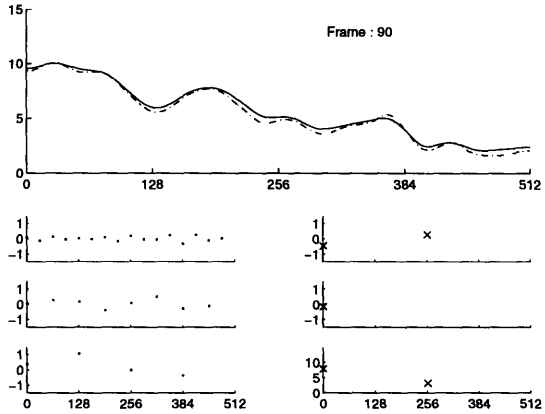

Frame : 85


Frame : 86


Frame : 87


Frame : 88


Frame : 89


Frame : 90

Bibliography

[1] B. S. Atal and S. L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.*, 50(2):637–655, August 1971.

[2] B. S. Atal and J. R. Remde. A new model of LPC excitation for producing natural-sounding speech at low bit rates. In *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Paris, France, 1982.

[3] T. Barnwell. Subband coder design using recursive quadrature mirror filters and optimum adpcm coders. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-30:751–765, October 1982.

[4] J.-H Chen and A. Gersho. Real-time vector apc speech coding at 2800 b/s with adaptive postfiltering. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, page 51.3.1, Dallas, 1987.

[5] C. K. Chui. *An Introduction to Wavelets*, volume 1 of *Wavelet Analysis and Its Applications*. Academic Press, 1992.

[6] C. K. Chui. *Wavelets: A Tutorial in Theory and Applications*, volume 2 of *Wavelet Analysis and Its Applications*. Academic Press, 1992.

[7] R. R. Coifman, Y. Meyer, and V. Wickerhauser. *Wavelets and Their Applications*, chapter Wavelet Analysis and Signal Processing. Jones and Bartlett, 1992.

[8] I. Daubechies. Orthonormal bases of complactly supported wavelets. *Comm. Pure and Appl. Math.*, 41:909–996, 1988.

[9] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, Penn., 1992.

[10] G. Fant. Vocal tract wall effects, losses, and resonance bandwidths. In *STL-QPSR*, 2-3, pages 28–51, Stockholm, Sweden, 1972. Royal Institute of Technology - Speech Transmission Labo ratory.

[11] J. L. Flanagan. Difference limen for the intensity of a vowel sound. *Journal of the Acoust Soc of America*, 27(6):1223–1225, November 1955.

[12] J. L. Flanagan. Difference limen for vowel formant frequency. *Journal of the Acoust Soc of America*, 27(6):613–617, November 1955.

[13] J. L. Flanagan. Difference limen for formant amplitude. *Journal of Speech and Hearing Disorders*, 22(2):205–212, June 1957.

[14] J. L. Flanagan and R. M. Golden. Phase vocoder. *Bell Systems Technical Journal*, 45:1493–1509, 1966.

[15] A. Grossman and J. Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM*, 15:723–736, January 1984.

[16] A. Grossman, J. Morlet, and T. Paul. Transforms associated to square integrable group representations, i. *J. Math. Phys.*, 26:2473–2479, 1985.

[17] J. N. Homes. The JSRU channel vocoder. *IEEE Proc.*, 127:53–60, 1980.

[18] D. H. Klatt and L. C. Klatt. Analysis, synthesis, and perception of voice quality variationsamong female and male talkers. *Journal of the Acoust Soc of America*, 87(2):820–857, February 1990.

[19] S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-11:674–693, July 1989.

[20] Stephane Mallat and W. L. Hwang. Singularity detection and processing with wavelets. *IEEE Trans. on Information Theory*, 38(2):617–643, March 1992.

[21] Stephane Mallat and Sifen Zhong. Characterization of signals from multiscale edges. Technical Report 592, Courant Institute of Mathematical Sciences, New York University, New York, New York, November 1991.

[22] R. J. McAulay and T. F. Quatieri. Speech analysis-synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-34:744–754, 1986.

[23] R. J. McAulay and T. F. Quatieri. Low-rate speech coding based on the sinusoidal model. In Sadaoki Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*, pages 165–207. Marcel Dekker, Inc., 1992.

[24] Y. Meyer. Ondelettes et functions spliles. In *Sem. Equations aux Derivees Partielles*, Paris, France, December 1986. Ecole Ploytechnique.

[25] J. Morlet, G. Arens, I. Fourgeau, and D. Giard. Wave propagation and sampling theory. *Geophysics*, 47:203–206, 1982.

[26] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing.* Prentice-Hall, Englewood Cliffs, N.J., 1975.

[27] D. B. Paul. The spectral envelope estimation vocoder. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-29:786–794, 1981.

[28] J. O. Pickles. *An Introduction to the Physiology of Hearing.* Academic Press, 1988.

[29] T. F. Quatieri and R. J. McAulay. Speech transformations bases on a sinusiodal representation. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-34:1449–1464, 1986.

[30] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, Englewoods Cliffs, New Jersey, 1978.

[31] Olivier Rioul and Martin Vetterli. Wavelets and signal processing. *IEEE SP Magazine*, pages 14–38, October 1991.

[32] M.J.T. Smith and T.P. III Barnwell. Exact reconstruction techniques for tree-structured subband coders. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-34(3):434–441, June 1986.

[33] M.J.T. Smith and S.L. Eddins. Analysis/synthesis techniques for subband inage coding. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-38(8):1446–1456, August 1990.

[34] K. Stevens. Lectures on Speech Communications, March 1993. Course at MIT.

[35] P.P. Vaidyanathan. *Multirate Systems and Filter Banks.* Prentice Hall Signal Processing Series. Prentice Hall, 1993.

[36] G. Wornell and A. V. Oppenheim. Estimation of fractal signals from noisy measurements using wavelets. *IEEE Trans. on Signal Processing*, 40(3):611–623, March 1992.

[37] Gregory Wornell. Synthesis, analysis, and processing of fractal signals. Technical Report 566, Research Laboratory of Electronics, MIT, Cambridge, MA, October 1991.