IDENTIFICATION OF TRANSIENT SPEECH USING WAVELET TRANSFORMS

by

Daniel Motlotle Rasetshwane

BS, University of Pittsburgh, 2002

Submitted to the Graduate Faculty of

School of Engineering in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

SCHOOL OF ENGINEERING

This thesis was presented

by

Daniel Motlotle Rasetshwane

It was defended on

April 4, 2005

and approved by

Patrick Loughlin, Professor, Electrical and Computer Engineering

Amro A. El-Jaroudi, Associate Professor, Electrical and Computer Engineering

John D. Durrant, Professor, Department of Communications Science and Disorders

J. Robert Boston, Professor, Electrical and Computer Engineering
Thesis Director

IDENTIFICATION OF TRANSIENT SPEECH USING WAVELET TRANSFORMS

Daniel Motlotle Rasetshwane, MS

University of Pittsburgh, 2005

It is generally believed that abrupt stimulus changes, which in speech may be time-varying frequency edges associated with consonants, transitions between consonants and vowels and transitions within vowels are critical to the perception of speech by humans and for speech recognition by machines. Noise affects speech transitions more than it affects quasi-steady-state speech. I believe that identifying and selectively amplifying speech transitions may enhance the intelligibility of speech in noisy conditions. The purpose of this study is to evaluate the use of wavelet transforms to identify speech transitions. Using wavelet transforms may be computationally efficient and allow for real-time applications. The discrete wavelet transform (DWT), stationary wavelet transform (SWT) and wavelet packets (WP) are evaluated. Wavelet analysis is combined with variable frame rate processing to improve the identification process. Variable frame rate can identify time segments when speech feature vectors are changing rapidly and when they are relatively stationary. Energy profiles for words, which show the energy in each node of a speech signal decomposed using wavelets, are used to identify nodes that include predominately transient information and nodes that include predominately quasi-steady-state information, and these are used to synthesize transient and quasi-steady-state speech components. These speech components are estimates of the tonal and nontonal speech components, which Yoo et al identified using time-varying band-pass filters. Comparison of spectra, a listening test and mean-squared-errors between the transient components synthesized using wavelets and Yoo's nontonal components indicated that wavelet packets identified the best estimates of Yoo's components. An algorithm that incorporates variable frame rate analysis into wavelet packet analysis is proposed. The development of this algorithm involves the processes of choosing a wavelet function and a decomposition level to be used. The algorithm itself has 4 steps: wavelet packet decomposition; classification of terminal nodes; incorporation of variable frame rate processing; synthesis of speech components. Combining wavelet analysis with variable frame rate analysis provides the best estimates of Yoo's speech components.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank my committee, Dr. J. Robert Boston, Dr. Ching-Chung Li, Dr. John Durrant, Dr Patrick Loughlin, and Dr Amro El-Jaroudi, for committing their time and for their advice and recommendations.

I would like to thank my advisor, Dr. Boston, for giving me an opportunity to pursue graduate studies and research. Dr Boston has provided his invaluable knowledge, guidance and motivation. I will forever be grateful.

To Sungyub Yoo and Paul Tantibundhit, thanks guys for paving the way. I learnt a lot from you.

Lastly, I would like to thank my family for their love, encouragement and patience

1.0 INTRODUCTION

Listening to someone speak in a noisy environment, such as a cocktail party, requires some effort and tends to be exhausting. Speaking louder, which is equivalent to amplifying speech by multiplying it by a constant, does not help much as it does not increase the intelligibility of speech in noisy conditions. In this study, we investigate method to improve the intelligibility of speech in noisy environment. Perhaps understanding what the human auditory system looks for in speech may give us clues as to what parts of speech we need to emphasize to enhance the intelligibility of speech.

For one who wishes to study perception of speech, the first task is obvious enough: it is to find the cues - the physical stimuli - that control the perception [26]. Since the invention of the spectrogram at AT&T Bell Laboratories, hundreds of articles on acoustical cues that influences the perceived phoneme have been published. A few of these articles, which influenced the current study, are cited here. In no way am I claiming that these articles are the earliest or the most influential in the research area of speech perception.

Potter et al, in a study of the transitions between stop consonants and vowels, using spectrograms, found that there are different movements of the second formant of the start of a vowel for stops with different place of articulation [41]. Joos also noted that formant transitions are characteristically different for various stop-consonant-vowel syllables [18].

Liberman characterized the formant transitions between stop-consonants-plus-vowel syllables, and concluded that (1) the second formant transition can be an important cue for distinguishing among either the voiceless stops /p, t, k/ or the voiced stops /b, d, g/ and (2) the perception of the different consonants depends on the direction and size of the formant transition and on the vowel [25]. In the same study, Liberman determined that the same transitions of the second formant observed for stop consonants can be used to distinguish the place of articulation of nasal consonants /m, n, n/. Characteristics of the spectrum during the release of the consonant as well as formant transition between consonants and vowels are important clue for identifying the place of articulation [15].

Third formants of vowels as compared to the first two formants typically carry much lower energy and have little or no effect on the phonetic identity of vowels [27]. This has led to fewer studies on the effect of third formant transition on perception. A study by Liberman found that, when frequencies of the first and second formants and the transition into these formants for the vowels /ae/ and /i/ are fixed, the transition of the third formant influenced the perceived place of articulation for voiced stop consonant /b, d, g/ [26].

These studies tied the place of articulation of stop consonants to the patterns in transitions of formants observed on spectrograms. It is to be noted though that spectrographic pattern for a particular phoneme typically looks very different in different contexts. For example, Liberman noted that /d/ in the syllable /di/ has a transition that rises into the second formant of /i/, while /d/ in /du/ has a transition that falls into the second formant of /u/ [28]. We should also note that the most important cues are sometimes among the least prominent parts of the acoustic signal [26]. The studies cited above also accentuate the importance of formant transition as acoustic cues for identifying and distinguishing some phonemes. Although these studies were conducted in noise-free environments, we believe that the same acoustic cues may be important for identifying and differentiating phonemes in noisy environments.

Steady-state formant activity is associated with vowels; in fact the perception of vowels depends primarily on the formant frequencies [28]. On the other hand, formant transitions are probably associated with consonants, transitions between consonants and vowels and transitions within some vowels. Compared to steady-state formant activity, formant transitions are short-lived and have very low energy, making them more susceptible to noise.

Researchers in the speech community have incorporated the importance of speech transitions into speech processing applications. Variable frame rate speech processing has been shown, by several authors, to improve the performance of automated speech recognition systems [40], [56], [23] [24]. Brown and Algazi identified spectral transitions in speech using the Karhunen-Loeve transform, with the intention of using them for sub-word segmentation and automatic speech recognition [4]. Quatieri and Dunn developed a speech enhancement method motivated by the sensitivity of the auditory system to spectral change [42]. Yoo et al intended to isolated transition information in speech, with the goal of using this information for speech enhancement [54]. Their method, which motivated the current study, is described below.

Yoo et al applied three time-varying band-pass filters (TVBF), based on a formant tracking algorithm by Rao and Kumaresan, to extract quasi-steady-state energy from highpass filtered speech [44], [54]. The algorithm applied multiple dynamic tracking filters (DTF), adaptive all-zero filters (AZF), and linear prediction in spectral domain (LPSD) to estimate the frequency modulation (FM) information and the amplitude modulation (AM) information. Each TVBF was implemented as an FIR filter of order 150 with the center frequencies determined by the FM information, and the bandwidth estimated using the AM information.

The output of each time-varying band pass filter was considered to be an estimate of the corresponding formant. The sum of the outputs of the three filters was defined as the *tonal component* of the speech. Yoo et al estimated the *nontonal component* of the speech signal by subtracting the tonal component from the highpass filtered speech signal. They considered the tonal component to contain most of the steady-state information of the input speech signal and the non-tonal component to contain most of the transient information of the input speech signal.

The speech signals were preprocessed by highpass filtering at 700 Hz to remove most of the energy associated with the first formant. Without highpass filtering, the adaptation of the TVBF was dominated by low-frequency energy. Removing this low-frequency energy made the algorithm more effective in extracting quasi-steady-state energy. The highpass filtered speech signals were as intelligible as the original speech signals, as shown by psychoacoustic studies of growth of intelligibility as a function of speech amplitude.

Yoo et al illustrated their decomposition of the word ‘pike’ (phonetically represented by /paIk/) spoken by a female [54]. Their results are reproduced in Fig 1.1, which shows the waveforms and corresponding spectrograms for the original and highpass filtered speech, and the tonal and nontonal components. The tonal component included most of the steady-state formant-activity associated with the vowel /aI/, from approximately 0.07 to 0.17 sec. The nontonal component captured the energy associated with the noise burst release accompanying the articulatory release of /p/, from approximately 0.01 sec to 0.07 sec, and the articulatory release of /k/ at around 0.38 sec. The tonal component included 87 % of the energy of the highpass filtered speech but it was unintelligible. The nontonal component included only 13 % of the energy of the highpass filtered speech but was almost as intelligible as the highpass filtered speech.

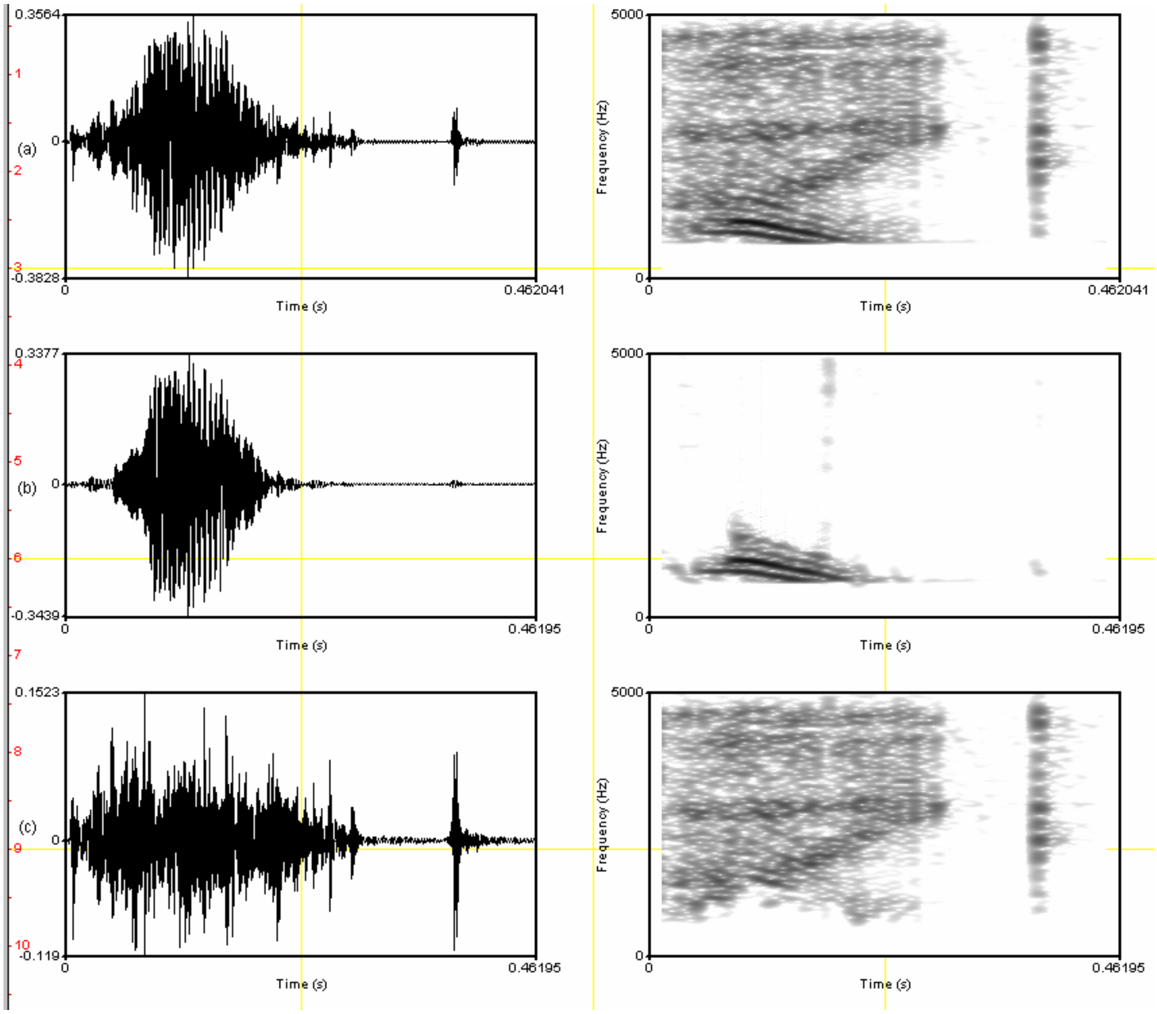

Figure 1.1: Waveform of speech (left column) and spectrograms (right column) for (a) highpass filtered speech, (b) tonal component and (c) nontonal component.

To determine the relative intelligibility of the highpass filtered, tonal and nontonal components compared to the original speech, Yoo et al determined psychometric functions to show the growth of intelligibility as signal amplitude increased. 300 phonetically-balanced consonant-vowel-consonant (CVC) words obtained from the NU-6 word lists were processed using their algorithm. They presented test words in quiet background, through headphones, to 5 volunteer subjects with normal hearing, who sat in a sound-attenuated booth. The subjects repeated the words they heard and the number of errors was recorded.

Their results showed that the mean of energy of the tonal component was 82 % of the energy of the highpass filtered speech and 18 % of the energy in the original speech. The mean of energy of the nontonal component was 18 % of the energy of the highpass filtered speech and 2 % of the energy in the original speech. These results are presented in Table 1.1, with standard deviations in parenthesis.

Table 1.1 Mean of energy in tonal and nontonal components of monosyllabic words relative to the energy in the highpass filtered speech and in the original speech.

	Tonal	Nontonal
% of highpass filtered speech	82 % (6.7)	18 % (6.7)
% of original speech	12 % (5.5)	2 % (0.9)

The maximum word recognition rates for the original, highpass filtered, tonal and nontonal components, determined by Yoo et al, are presented in Table 1.2, with standard deviations in parenthesis. Statistical analyses of the maximum word recognition rates showed that the tonal component had a significantly lower maximum recognition rate than other components. The maximum word recognition rate of the nontonal was slightly lower than that of the original and highpass filtered speech. The original and highpass filtered speech had similar maximum word recognition rates. The fact that the nontonal component, which emphasizes formant transitions, had a maximum recognition rate that is almost twice that of the tonal asserts the importance of formant transitions as important cues for identifying and distinguishing phonemes.

Table 1.2: Maximum recognition rates for original and highpass filtered speech, and for tonal and nontonal components.

	Max. recognition rate
original	98.7 % (3.0)
highpass filtered	96.5 % (2.1)
tonal	45.1 % (19.3)
nontonal	84.9 % (14.4)

The algorithm of Yoo et al appears to be effective in extracting quasi-steady-state energy from speech, leaving a speech component that emphasizes transitions. They suggested that selective amplification of the nontonal component might enhance the intelligibility of speech in noisy conditions. However, the algorithm is computationally intensive and unsuitable for real-time applications. Wavelet analysis provides a method of time-frequency analysis that can be implemented in real-time. The purpose of this study is to determine whether a wavelet-based analysis can be used to identify the nontonal speech component described by Yoo et al. Identifying speech transitions using wavelets may reduce the computation time and allow for real-time applications of the proposed speech enhancement technique.

In this study, wavelet analysis of the highpass filtered speech, tonal and nontonal components of Yoo are carried out. Wavelet coefficients of the highpass filtered speech are then compared to those of the tonal and nontonal components to determine whether specific coefficients are associated with the tonal or nontonal. Through the analysis, speech components that are similar to the tonal and nontonal of Yoo components are identified. Although the identification process will use the components of Yoo, it is expected to shed light on the extent to which wavelet analysis can be applied to the identification of speech transitions.

Wavelet transforms have been used by several investigators in the speech research community for automatic speech recognition [45] [13], pitch detection [20] [46] [6], speech coding and compression [34] [50] [38], speech denoising and enhancement [1] [51] [29] [14] and other processes. Wavelet analysis, because of its multiresolution properties, can detect voiced stops, since stops have a sudden burst of high frequency [13].

Another method of identifying speech transitions is provided by variable frame rate (VFR) processing, which identifies time segments when speech feature vectors are changing rapidly. Variable frame rate techniques have been used by several investigators in speech recognition studies [40], [56], [23] [24]. These studies were primarily concerned with reducing the amount of data to be processed and improving recognition rates. Time segments of the speech signal in which the speech feature vectors are changing rapidly may be associated with transient speech, while time segments in which the speech feature vectors are slowly changing may be associated with quasi-steady-state speech. An investigation to determine whether incorporating variable frame rate processing can improve the identification of speech transitions is also carried out.

This thesis is arranged as follows. Chapter 2 gives a summary of the relevant literature. This chapter begins with a summary of wavelet theory and a review of the use of wavelets in speech processing. A review of variable frame rate techniques follows, including discussions of linear predictive coding (LPC) and Mel-frequency cepstral coefficients (MFCC). Chapter 2 concludes with a brief discussion of Yoo's formant tracking algorithm. Chapter 3 describes the methods used and results obtained when the discrete wavelet transform, stationary wavelet transform and wavelet packets were evaluated for use in identifying transient and quasi-steady-state speech components. Chapter 4 presents an algorithm for identifying transient and quasi-steady-state speech components that incorporates wavelet packet analysis and variable frame rate processing. Results obtained with this algorithm are described. Chapter 5 discusses the results and limitations, possible improvement, and possible uses of the speech transient identification techniques.

2.0 BACKGROUND

The basic theory of wavelet transforms discussed here covers the continuous wavelet transform, multiresolution analysis, the discrete wavelet transform, the overcomplete wavelet transform, and signal decomposition and reconstruction using filter banks. In the discussion, the continuous wavelet, the discrete wavelet and the discrete scaling functions and their properties will be described.

Variable frame rate processing, linear predictive coding (LPC), Mel-frequency cepstral coefficients (MFCC) and the formant tracking algorithm are also discussed. The discussion of LPC and MFCC will focus on how these feature vectors are computed from speech and how they are applied to the variable frame rate process.

2.1 WAVELET THEORY

The use of wavelets in signal processing applications is continually increasing. This use is partly due to the ability of wavelet transforms to present a time-frequency (or time-scale) representation of signals that is better than that offered by Short-time Fourier transform (STFT). Unlike the STFT, the wavelet transform uses a variable-width window (wide at low frequencies and narrow at high frequencies) which enables it to “zoom in” on very short duration high frequency phenomena like transients in signals [7].

This section reviews the basic theory of wavelets. The discussion is based on [5] [7] [8] [17] [31] [32] [35] [47] [48] and [49]. The continuous wavelet transform (CWT), multiresolution analysis (MRA), the discrete wavelet transform (DWT), the overcomplete wavelet transform (OCWT), and filter banks are discussed. Wavelet properties that influence the type of wavelet basis functions that is appropriate for a particular application will be examined, and some of the uses of wavelets in speech processing will be reviewed.

2.1.1 The Continuous Wavelet Transform

A function ψ(t) $\in$ L^2(R) is a *continuous wavelet* if the set of functions

$$\psi_{b,a}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{2.1}$$

is an orthonormal basis in the Hilbert space L^2(R), where a and b are real.

The set of functions $\psi_{b,a}(t)$ are generated by translating and dilating the function ψ(t). Parameter (*a*) is a scaling parameter. Varying it changes the center frequency and the bandwidth of ψ(t). The time and frequency resolution of the wavelet transform, discussed below, also depend on a. Small values of the scaling parameter (*a*) provide good time localization and poor frequency resolution, and large values of the scaling parameter provide good frequency resolution and poor time resolution. The time delay parameter

(b) produces a translation in time (movement along the time axis). Dividing ψ by $\sqrt{a}$ insures that all members of the set $\{\psi_{b,a}(t)\}$ have unity Euclidean norm (L^2-norm) i.e. $\|\psi_{b,a}\|_2 = \|\psi\|_2 = 1$ for all integer a and b. The function $\psi(t)$ from which the set of functions $\psi_{b,a}(t)$ are generated is called the *mother* or *analyzing wavelet.*

The function ψ(t) has to satisfy the following properties to be a wavelet:

1. ψ(t) integrates over time to zero and it's Fourier transform $\Psi(\omega)$ equals to zero at ω = 0 [35]

$$\Psi(\omega = 0) = \int_{-\infty}^{\infty} \psi(t) dt = 0 . \tag{2.2}$$

2. ψ(t) has finite energy, i.e. most of the energy of ψ(t) has to be confined to a finite duration

$$\int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty . \tag{2.3}$$

3. ψ(t) satisfies the *admissibility condition*, [35] i.e.

$$\int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{\omega} d\omega = C_\psi < \infty \tag{2.4}$$

The admissibility condition ensures perfect reconstruction of a signal from its wavelet representation and will be discussed later in this section.

The wavelet function ψ(t) may be complex. In fact, a complex wavelet function is required to analyze the phase information of signals [32].

The *continuous wavelet transform* (CWT) $W_x(b,a)$ of a continuous-time signal x(t) is defined by [35]

$$W_x(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-b}{a} \right) dt \tag{2.5}$$

where a and b are real. The CWT is the inner product of x(t) and the complex conjugate of the translated and scaled version of the wavelet, ψ(t), i.e. $W_x(b,a) = \langle x(t), \psi^*{}_{b,a}(t) \rangle$. Eq. 2.5, shows that the wavelet transform $W_x(b,a)$ of a one dimensional signal x(t) is two dimensional. The CWT can be expressed as a convolution by [47]

$$W_x(b,a) = \langle x(t), \psi^*{}_{b,a}(t) \rangle = x(t) * \psi^*{}_{b,a}(-t). \tag{2.6}$$

The CWT expressed as a convolution may be interpreted as the output of an infinite bank of linear filters described by the impulse responses $\psi_{b,a}(t)$ over the continuous range of scales *a* [47].

To recover x(t) from $W_x(b,a)$, the mother wavelet ψ(t) has to satisfy the admissibility condition given in Eq. 2.4. If the admissibility condition is satisfied, x(t) can be perfectly reconstructed from $W_x(b,a)$ as

$$x(t) = \frac{1}{C_{\psi^*}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{a^2} W_x(b,a) \psi^*{}_{b,a}(t) da db . \tag{2.7}$$

The constant C_ψ is the admissibility constant and is defined in Eq. (2.4).

The CWT is covered here for completeness. It was not evaluated for use in identifying transient and quasi-steady-state speech, since the CWT is mainly useful for characterization of signals (analysis) [17]. Since computers are usually used to evaluate wavelet transform, the CWT cannot be evaluated directly in most applications. The discrete version is needed. For some signals, the coordinates (a, b) may cover the entire time-scale plane, giving a redundant representation of x (t). The calculation of the CWT is also not efficient because the CWT is defined continuously over the time-scale plane [47].

2.1.2 Multiresolution Analysis and Scaling function

In this section, the scaling function *φ*(t) will be introduced via a multiresolution analysis. The relationship between the scaling function *φ*(t) and the wavelet function *ψ*(t) will be discussed. This discussion follows the description given by Vaidyanathan et al [48]. In the following discussion, L^2 refers to the space of square-integrable signals.

Multiresolution analysis involves the approximation of functions in a sequence of nested linear vector spaces $\{V_k\}$ in L^2 that satisfy the following 6 properties:

1. Ladder property: $\ldots V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \ldots$
2. $\bigcap_{j=-\infty}^{\infty} V_j = \{0\}$.

3. Closure of $\bigcup_{j=-\infty}^{\infty} V_j$ is equal to L^2

4. Scaling property: $x(t) \in V_j$ if and only if $x(2t) \in V_{j+1}$. Because this implies that "$x(t) \in V_0$ if and only if $x(2^{-j}t) \in V_j$", all the spaces V_j are scaled versions of the space V_0. For j>0, V_j is a coarser space than V_0.

5. Translation invariance: If $x(t) \in V_0$, then $x(t-k) \in V_0$; i.e. the space V_0 is invariant to translation by integers. The scaling property implies that V_j is invariant to translation by $2^{-j}k$.

6. Special Orthonormal basis: A function $\phi(t) \in V_0$ exists such that the integer shifted version $\{\phi(t-k)\}$ forms an orthonormal basis for V_0. Using the scaling property means that $\left\{2^{-\frac{j}{2}}\phi\left(2^{-j}t-k\right)\right\}$ is an orthonormal basis of V_j. The function $\phi(t)$ is called the *scaling function* of *multiresolution analysis*.

The scaling function $\phi_{j,k}(t) = 2^{-\frac{j}{2}}\phi\left(2^{-j}t-k\right)$ spans the space V_j. To better describe and parameterize signals in this space, a function that spans the difference between the spaces spanned by various scales of the scaling function is needed. Wavelets are these functions.

The space W_j spanned by the wavelet function has the following properties [49];

1. $\{\psi(t-k)\}$ is an orthonormal basis of W_0, given by the

   orthogonal complement of V_0 in V_1, i.e. $V_1 = V_0 \oplus W_0$, where V_0 is the initial space spanned by $\varphi(t)$.

2. If $\psi(t) \in$ W$_0$ exists, then $\psi_{j,k}(t) = 2^{-\frac{j}{2}}\psi(2^{-j}t - k)$ is an

   orthonormal basis of the space W$_j$. W$_j$ is the orthogonal complement of V$_j$ in V$_{j+1}$, i.e. $V_{m+1} = V_m \oplus W_m = V_0 \oplus W_0 \oplus W_1 \oplus \cdots \oplus W_m$.

3. $L^2 = V_0 \oplus W_0 \oplus W_1 \oplus \ldots$

Using the scaling function and the wavelet function, a set of functions that span all of L^2 can be constructed. A function x(t) $\in$ L^2 can be written as a series expansion in terms of these two functions as [5]

$$x(t) = \sum_{k=-\infty}^{\infty} c(j,k)\phi_{J,k}(t) + \sum_{j=0}^{\infty}\sum_{k=-\infty}^{\infty} d(j,k)\psi_{j,k}(t). \tag{2.8}$$

Here J is the coarsest scale. In the above expression, the first summation gives an approximation to the function x(t) and the second summation adds the details. The coefficients c(*j*,*k*) and d(*j*,*k*) are the *discrete scaling coefficients* and the *discrete wavelet coefficients* of x(t) respectively [5]

2.1.3 The Discrete Wavelet Transform

The discrete wavelet transform (DWT) is obtained in general by sampling the corresponding continuous wavelet transform [47]. The discussion of this section is based on [47].

To discretize the CWT, an analyzing wavelet function that generates an orthonormal (or biorthonormal) basis for the space of interest is required. An analyzing wavelet function with this property allows the use of finite impulse response filters (FIR) in the DWT implementation. There are many possible discretizations of the CWT, but the most common DWT uses a dyadic sampling lattice. Figure 2.1 shows the time-scale cells corresponding to dyadic sampling. Dyadic sampling and restricting the analyzing wavelets to ones that generates orthonormal bases allows the use of an efficient algorithm known as the *Mallat algorithm* or *fast wavelet transform* in the DWT implementation [31]. The Mallat algorithm will be discussed in the next section.

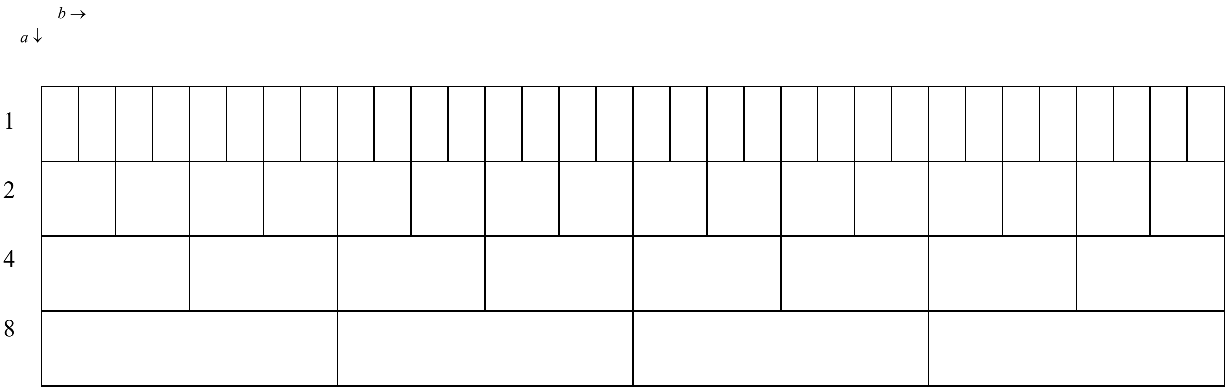

Figure 2.1: Time-scale cells corresponding to dyadic sampling.

Sampling the CWT using a dyadic sampling lattice, the discrete wavelet is given by

$$\psi_{j,k}(t) = 2^{-\frac{j}{2}}\psi\left(2^{-j}t - k\right) \tag{2.9}$$

where *j* and *k* take on integer values only. Parameter *j* and *k* are related to parameters *a* and *b* of the continuous wavelet by $a = 2^j$, and $k = 2^{-j}b$.

2.1.4 Signal Decomposition and Reconstruction using Filter Banks

The discussion of this section will follow the description given by [5]. Eq. 2.8 can be expanded as

$$x(t) = \sum_k c(j,k) 2^{-\frac{j}{2}} \phi\left(2^{-j} t - k\right) + \sum_k d(j,k) 2^{-\frac{j}{2}} \psi\left(2^{-j} t - k\right) \qquad (2.10)$$

In this and subsequent equations, scale j+1 is coarser that scale j.

If the wavelet function is orthonormal to the scaling function, the level j coefficients c(j,k) and d(j,k) can be obtained as:

$$c(j,k) = \left\langle x(t), \phi_{j,k} \right\rangle = \int x(t) 2^{-\frac{j}{2}} \phi\left(2^{-j} t - k\right) dt \qquad (2.11)$$

$$d(j,k) = \left\langle x(t), \psi_{j,k} \right\rangle = \int x(t) 2^{-\frac{j}{2}} \psi\left(2^{-j} t - k\right) dt \,. \qquad (2.12)$$

The level j+1 scaling and detail coefficients can be obtained from the level j scaling coefficients as [5]

$$c(j+1,k) = \sum_m \tilde{h}(m-2k) c(j,m) \qquad (2.13)$$

$$d(j+1,k) = \sum_m \tilde{g}(m-2k) c(j,m) \qquad (2.14)$$

Using these equations, level j+1 scaling and wavelet coefficients can be obtained from the level j scaling coefficients by filtering with finite impulse response (FIR) filters $\tilde{h}(n)$ and $\tilde{g}(n)$, then downsampling the result. This technique is known as the *Mallat decomposition algorithm* and is illustrated in Figure 2.2 [31]. The partial binary tree of Figure 2.2 is sometimes referred to as a Mallat tree.

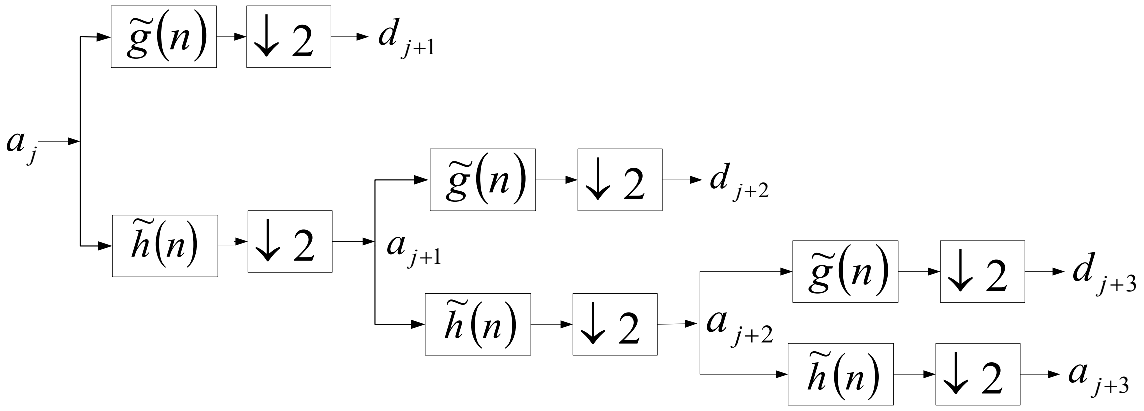

Figure 2.2: A three-stage Mallat signal decomposition scheme

In the decomposition scheme, the first stage splits the spectrum into two equal bands: one highpass and the other lowpass. In the second stage, a pair of filters splits the lowpass spectrum into lower lowpass and bandpass spectra. This splitting results in a logarithmic set of bandwidth shown in Figure 2.3.

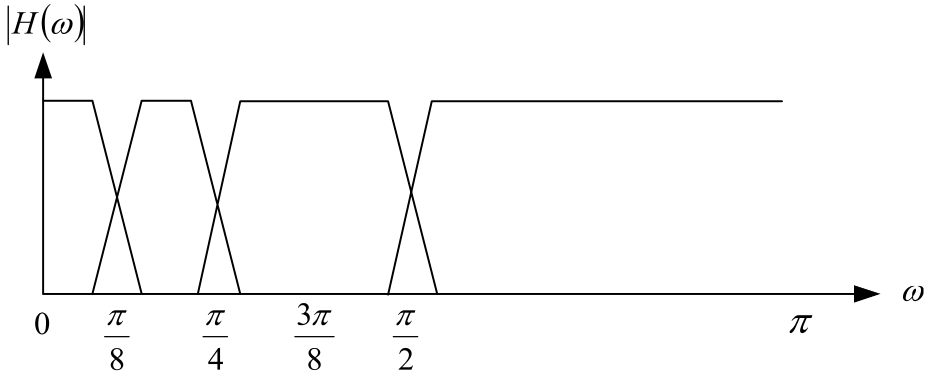

Figure 2.3: Frequency response for a level 3 discrete wavelet transform decomposition

Level *j* scaling coefficients can be reconstructed from the level *j*+1 wavelet and scaling coefficients by

$$c(j,k) = \sum_m c(j+1,m)h(k-2m) + \sum_m d(j+1,m)g(k-2m). \quad (2.15)$$

In words, the level *j* scaling coefficients are obtained from the level *j*+1 scaling and wavelet coefficients by upsampling the level *j*+1 wavelet and scaling coefficients, filtering the outputs from the upsamplers using filters *h*(n) and *g*(n), and then adding the filter outputs. The signal reconstruction scheme is illustrated in Figure 2.4.

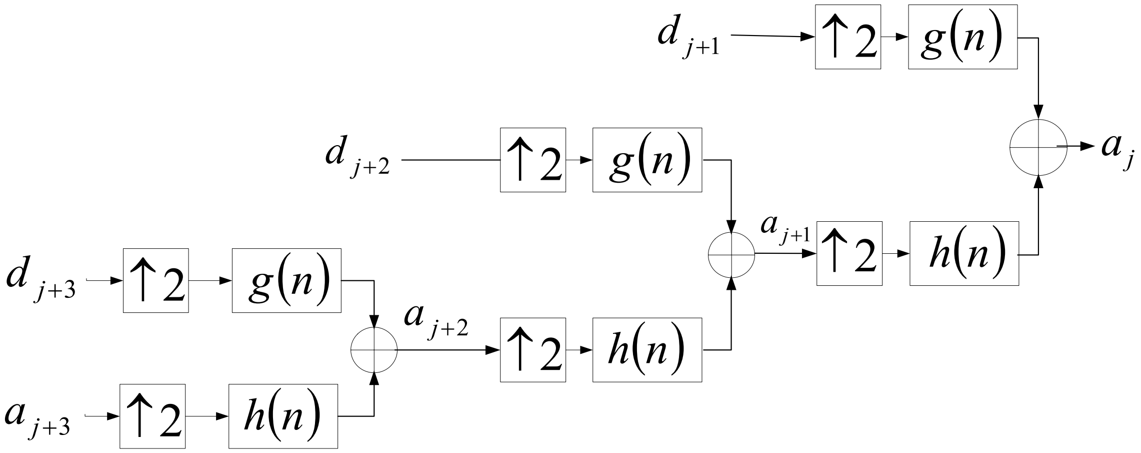

Figure 2.4: A three-stage Mallat signal reconstruction scheme

Filters $\tilde{h}(n)$ and *h*(n) are low-pass whereas filters $\tilde{g}(n)$ and *g*(n) are high-pass. The impulse responses of these filters satisfy the following properties [31];

1. $\tilde{h}(n)$ = *h*(-n) and $\tilde{g}(n)$ = *g*(-n).

2. g(n) = $(-1)^{1-n}$h(1-n) i.e. H and G are *quadrature mirror filters.*
3. |H(0)| = 1 and *h*(n) = O(n^{-2}) at infinity, i.e. the asymptotic upper bound of *h*(n) at infinity is n^{-2}.
4. $|H(\omega)|^2 + |H(\omega+\pi)|^2 = 1$.

2.1.5 The Overcomplete Wavelet Transform

Because the discrete wavelet transform uses a dyadic sampling grid and generate an orthonormal basis, it is computationally efficient and has reasonable storage requirements (an N sample signal decomposed at a maximum scale S produces $\sum_{s=1}^{S} 2^{-s} N + 2^{-S} N$ samples when using the DWT versus SN samples when using the CWT). The efficiency of the DWT is achieved with potential loss of performance benefits compared to the CWT. Compared to the CWT, the DWT is more susceptible to noise and has restrictions on the analyzing wavelet used. A compromise can be reached by using the overcomplete wavelet transform (OCWT).

Nason and Silverman described the *Stationary Wavelet Transform* (SWT)*,* which like the OCWT defined by Liew is similar to the DWT but omits decimation [39]. With the SWT, the level *j*+1 scaling and wavelet coefficients are computed from the level *j* scaling coefficients by convolving the latter with modified version of filter *h*(n) and *g*(n). The filters are modified by inserting a zero between every adjacent pair of elements of the filters *h*(n) and *g*(n). Teolis defined his OCWT by sampling, on the time-scale plane, the corresponding CWT [47]. However the sampling lattice for the OCWT is not dyadic.

Teolis defined a semilog sampling grid for the OCWT whereby the scale samples were exponentially spaced and the time samples where uniformly spaced [47]. Mallat et al computed the OCWT by computing the DWT and omitting the decimations [30].

Specifically, the OCWT is defined by an analyzing wavelet, the corresponding CWT, and a discrete time-scale sampling set [47]. A condition put on the sampling set is that it should produce an OCWT representation that spans the Hilbert space [47]. If this condition is met, the OCWT representation is invertible and an inverse transform exists [47].

The advantages of the OCWT over the DWT include [47];

(1) Robustness to imprecision in representation of coefficients, for example, quantization effects.

(2) Freedom to select an analyzing wavelet since the OCWT does not require an analyzing wavelet that generates an orthonormal basis.

(3) Robustness to noise that arises from the overcompleteness of the representation.

2.1.6 Wavelet Packets

The DWT results in a logarithmic frequency resolution. High frequencies have wide bandwidth whereas low frequencies have narrow bandwidth [5]. The logarithmic frequency resolution of the DWT is not appropriate for some signals. Wavelet Packets allow for the segmentation of the higher frequencies into narrower bands. An entropy measure can also be incorporated into the wavelet packet system to achieve an adaptive wavelet packet system (adapted to particular signal or class of signals [5]). This section discusses the full wavelet packet decomposition, following [5]. The coarsest level will be designated by the highest numerical level, rather than level 0 as in [5].

2.1.6.1 Full Wavelet Packet Decomposition

In the DWT decomposition, to obtain the next level coefficients, scaling coefficients (lowpass branch in the binary tree) of the current level are split by filtering and downsampling. With the wavelet packet decomposition, the wavelet coefficients (highpass branch in binary tree) are also split by filtering and downsampling. The splitting of both the low and high frequency spectra results in a full binary tree shown in Figure 2.5 and a completely evenly spaced frequency resolution as illustrated in Figure 2.6. (In the DWT analysis, the high frequency band was not split into smaller bands.) In the structure of Figure 2.5, each subspace, also referred to as a node, is indexed by its depth and the number of subspaces below it at the same depth. The original signal is designated depth zero.

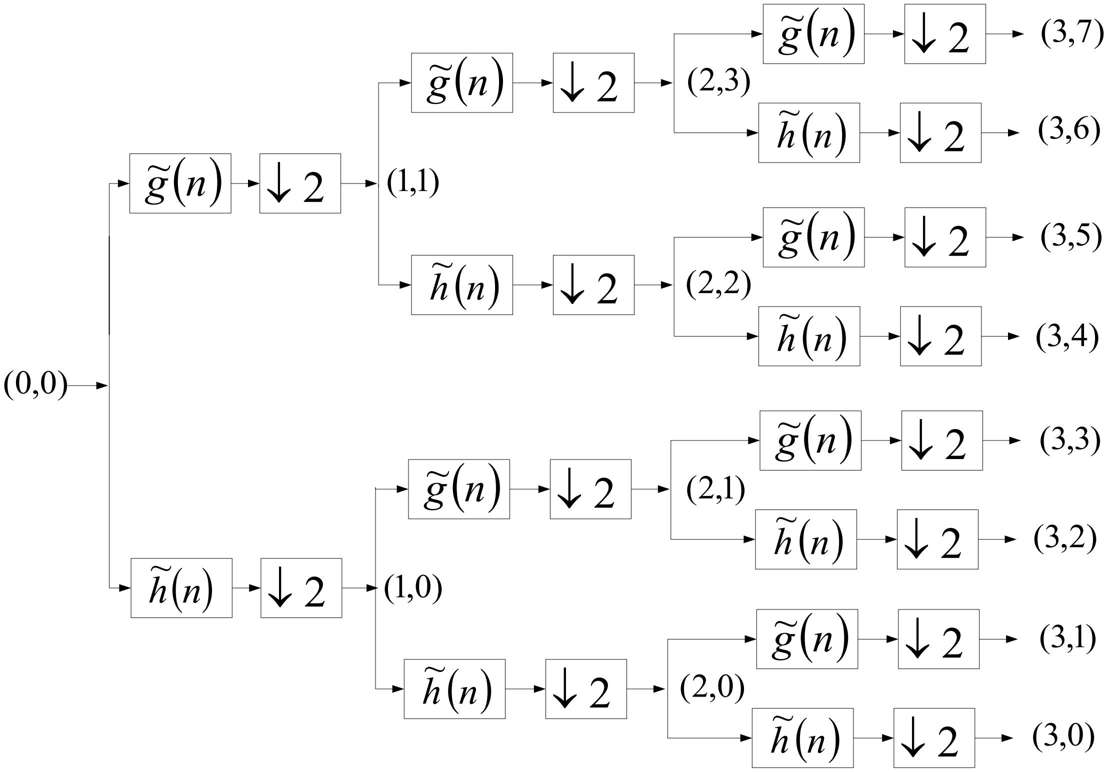

Figure 2.5: Three-stage full wavelet packet decomposition scheme

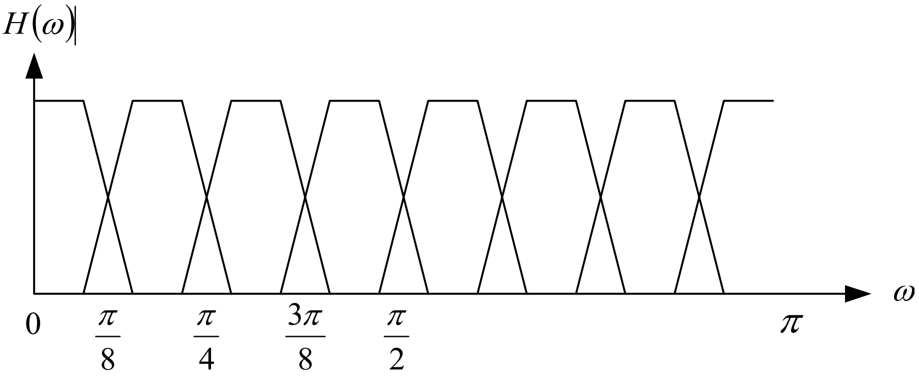

Figure 2.6: Frequency response for a level 3 wavelet packets decomposition

An alternative tree labeling scheme is shown in Figure 2.7 for a wavelet packet decomposition of depth 4. In this scheme, the nodes are labeled using counting numbers with index 0, corresponding to the original signal, as the root of the tree.

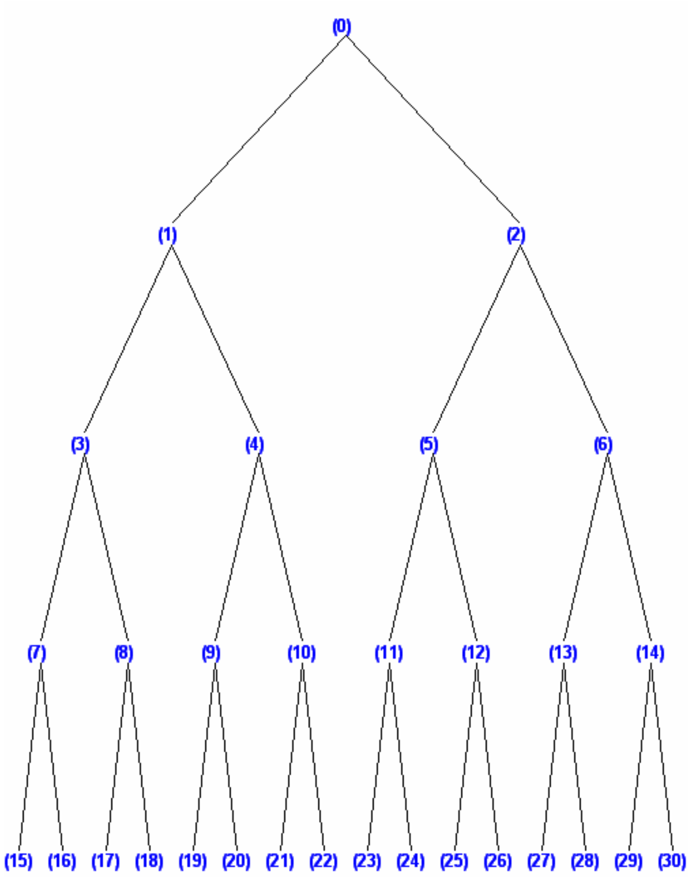

Figure 2.7: Alternate wavelet packet tree labeling.

The wavelet packet reconstruction scheme is achieved by upsampling, filtering with appropriate filters and adding coefficients. This scheme is shown in Figure 2.8. This WP reconstruction tree structure is labeled the same as the WP decomposition structure.

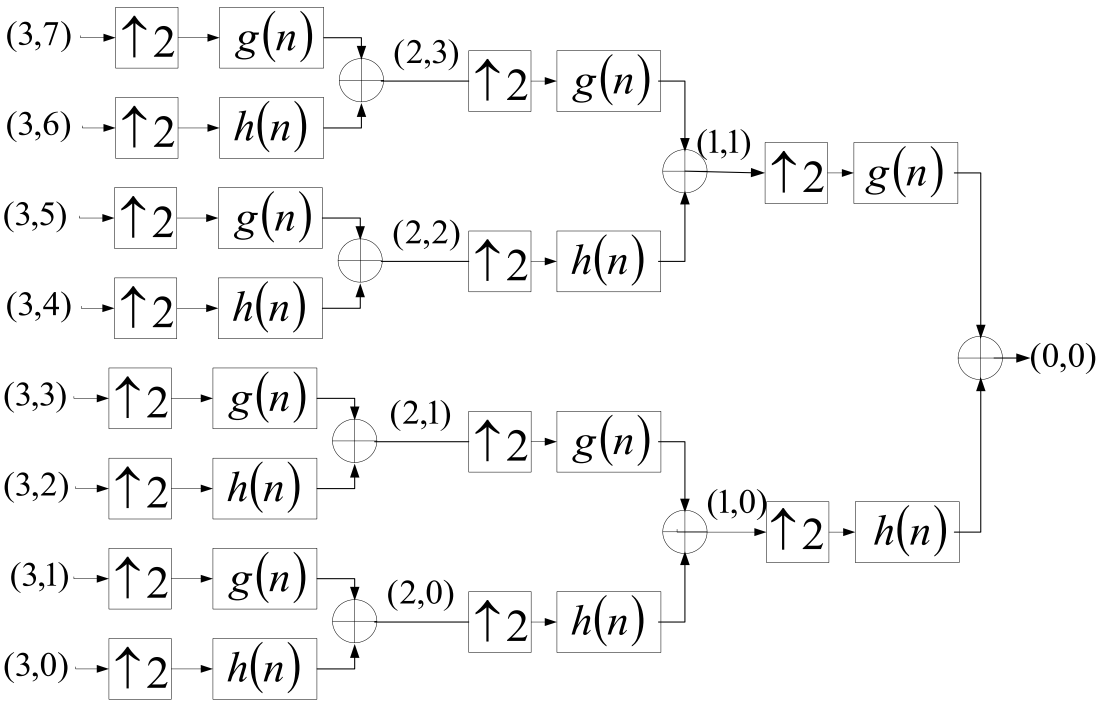

Figure 2.8: Three-stage full wavelet packet reconstruction scheme

As in the DWT scheme, $\tilde{h}(n)$ and *h*(n) are lowpass filters whereas $\tilde{g}(n)$ and *g*(n) are highpass filters. Additional properties that were presented for these filters in the DWT scheme (section 2.1.3) also hold here.

2.1.7 Choosing a Wavelet Function

Since the formulation of the Haar wavelet in the early twentieth century, many other wavelets have been proposed. The paper ‘Where do wavelets come from?-a personal point of view’ by Daubechies presents a good historical perspective on wavelets [8]. This paper, among others, discusses the works of Morlet, Grossmann, Meyer, Mallat and Lemarié that led to the development of wavelet bases and the wavelet transforms.

A well chosen wavelet basis will result in most wavelet coefficients being close to zero [32]. The ability of the wavelet analysis to produce a large number of non-significant wavelet coefficients depends on the *regularity* of the analyzed signal x(t), and the number of *vanishing moments* and *support size* of ψ(t). Mallat related the number of vanishing moments and the support size to the wavelet coefficients amplitudes [32].

Vanishing Moments

ψ(t) has p vanishing moments if

$$\int_{-\infty}^{\infty} t^k \psi(t) dt = 0 \quad for \quad 0 \le k < p \,. \tag{2.16}$$

If x(t) is regular and ψ(t) has enough vanishing moments, then the wavelets coefficients $d(j,k) = \langle x(t), \psi_{j,k} \rangle$ are small at fine scale.

Size of Support

If x(t) has an isolated singularity (a point at which the derivative does not exist although it exists everywhere else) at t_0 and if t_0 is inside the support of $\psi_{j,k}(t)$, then $d(j,k) = \langle x(t), \psi_{j,k} \rangle$ may have large amplitudes. If ψ(t) has a compact support of size K, there are K wavelets $\psi_{j,k}(t)$ at each scale 2^j whose support includes t_0. The number of large amplitude coefficients may be minimized by reducing the support size of ψ(t).

If $\psi(t)$ has p vanishing moments, then its support size is at least 2p – 1 [32]. A reduction in the support size of $\psi(t)$ unfortunately means a reduction in the number of vanishing moments of $\psi(t)$. There is a trade off in the choice of $\psi(t)$. A high number of vanishing moments is preferred if the analyzed signal x(t) has few singularities. If the number of singularities of x(t) is large, a $\psi(t)$ with a short support size is a better choice.

Examples of wavelets basis

This subsection presents some properties of three wavelet families, Daubechies, Symlets and Morlet. Daubechies and Symlets wavelets were evaluated for use in decomposing speech.

Daubechies and Symlets wavelets are orthogonal wavelets that have the highest number of vanishing moments for a given support width. In the naming convention, db*i* or sym*i*, *i* is an integer that denotes the order, e.g. db8 is an order 8 Daubechies wavelet and sym7 is an order 7 Symlets wavelet [7]. An order *i* wavelet has *i* vanishing moments, a support width of 2*i*-1 and a filter of length 2*i*. These wavelets are suitable for use with both the continuous wavelet transform and the discrete wavelet transform. The difference between these two wavelet functions is that Daubechies wavelets are far from symmetry while Symlets wavelets are nearly symmetric. As an example, Figure 2.9 (a) and (b) show wavelet (psi) and scaling functions (phi) for order 4 Daubechies and Symlets wavelets.

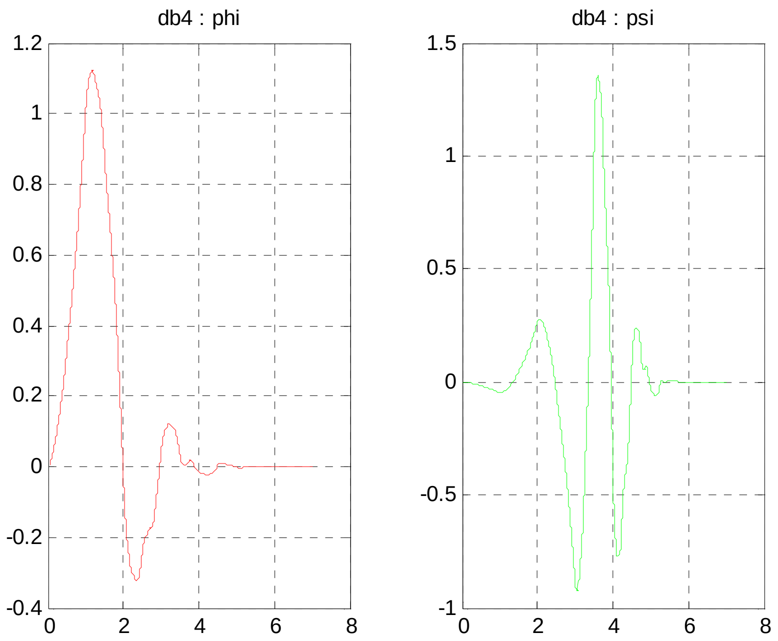

Figure 2.9: Order 4 Daubechies scaling (phi) and wavelet (psi) functions.

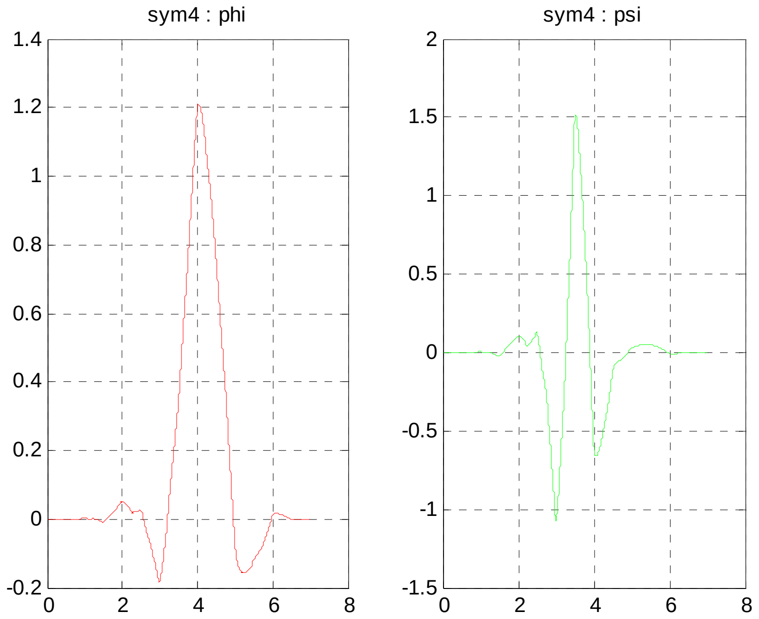

Figure 2.10: Order 4 Symlets scaling (phi) and wavelet (psi) functions.

The Morlet wavelet was formulated by J. Morlet for a study of seismic data [37]. The Morlet wavelets function is a complex wavelet which, because of the lack of existence of a corresponding scaling function, can only be used for CWT analysis. Although the Morlet wavelet has infinite support, its effective support is in the range [-4 4]. The Morlet wavelet function, which is a modulated Gaussian function, is given by

$$\psi(t)=e^{j\omega_0 t}e^{\frac{t^2}{2}} \tag{2.17}$$

Figure 2.11 shows the real part of the Morlet wavelet function with ω_0 = 5, ($\psi(t)=e^{\frac{t^2}{2}}\cos(5t)$). The real part of the Morlet wavelet is a cosine modulated Gaussian function.

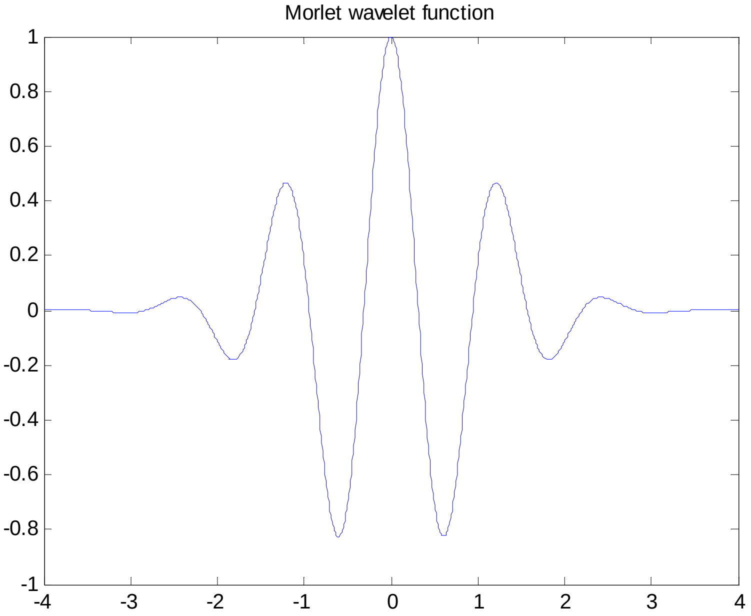

Figure 2.11: Morlet wavelet function

2.2 USE OF WAVELETS IN SPEECH PROCESSING

Recently, wavelet transforms have found widespread use in various fields of speech processing. Among the many applications, wavelets have been used in automatic speech recognition, pitch detection, speech coding and compression, and speech denoising and enhancement. This subsection will review some of the work in applying wavelets to speech processing.

Ris, Fontaine and Leich presented a method to represent relevant information of a signal with a minimum number of parameters [45]. They proposed a pre-processing algorithm that produces acoustical vectors at a variable frame rate. Signal analysis and segmentation of Ris et al was based on the Malvar wavelets [33]. They computed a Malvar cepstrum from the Malvar wavelet coefficients and used it as input to a Hidden Markov model (HMM) based speech recognizer. Before the Malvar wavelet coefficients were presented to the HMM recognizer, segmentation based on an entropy measure was performed to produce a variable frame rate coded feature vector. The segmentation produced short segments for transient and unvoiced speech and long segments for voiced speech. In an isolated word speech recognition task, the performance of the Ris et al method was comparable to that of an LPC cepstrum recognizer when segmentation was not used. With segmentation in the Ris method, the LPC cepstrum recognizer performed better than the Ris method.

Farooq and Datta used a Mel filter-like admissible wavelet packet (WP) structure instead of the popular Mel-frequency cepstral coefficients (MFCC) to partition the frequency axis into bands similar to those of the Mel-scale for speech recognition [13]. Instead of using the logarithm of the amplitude Fourier transform coefficients as input to the filter banks, they used WP coefficients. Just as in the MFCC computation, Farooq et al computed the discrete cosine transform of the output of the filter banks. In a speech recognition test, they observed that the features derived from WP performed better than MFCC features for unvoiced fricatives and voiced stops, and MFCC features outperformed WP features for voiced fricatives and vowels. According to Farooq et al, the reason for this was that the STFT (which uses cosines and sines) which is used in the MFCC computation is more efficient for the extraction of periodic structure from a signal. Also wavelet packets have multiresolution properties that enable them to capture stops because stops have a sudden burst of high frequency.

Kadambe and Boudreaux-Bartels developed a noise-robust event-detection pitch detector that was based on the dyadic wavelet transform [20]. Their pitch detector was suitable for both low-pitched and high-pitched speakers. The dyadic wavelet transform was applied to detect the glottal closure (defined as an event), and the time interval between two such events was the estimate of the pitch period. They demonstrated that their pitch detector was superior to classical pitch detectors that utilize autocorrelation and cepstrum methods to estimate pitch period. More recent wavelet-based pitch detectors have followed the work of Kadambe and Boudreaux-Bartels.

Shelby et al used the pitch detection method of Kadambe and Boudreaux-Bartels to detect pitch period in tone languages [46]. Jing and Changchun incorporated an autocorrelation function into the pitch detector of Kadambe and Boudreaux-Bartels [20].

Chen and Wang improved the pitch detector of Kadambe et al [20] by developing a wavelet-based method for extracting pitch information from noisy speech [6]. They applied a modified spatial correlation function to improve the performance of the pitch detector in a noisy environment. To further increase the performance of their pitch detector, an aliasing compensation algorithm was used to eliminate the aliasing distortion caused by the downsampling and the upsampling performed in the computation of DWT coefficients. Through simulations, they showed that their pitch detection method gave better results in noisy conditions than other time, spectral and wavelet domain pitch detectors.

Mandridake and Najim described a scheme for speech compression that employed discrete wavelet transform and vector quantization (VQ) [34]. In their coding system which they called discrete wavelet vector transform quantization (DWVTQ), a speech signal was transformed to wavelet coefficients corresponding to different frequency bands which were then quantized separately. Their method used product code structure for each frequency band. Mandridake et al took account of both the statistics of the wavelet coefficients and the fact that the ear is less sensitive to high frequencies in their bit assignment for the vector codes. Results showed that their method outperformed the discrete wavelet scalar transform quantization (DWSTQ) method; it was more efficient, and showed improved optimal bit allocation in comparison to uniform bit allocation.

Xiaodong, Yongming and Hongyi presented a speech compression method based on the wavelet packet transform [50]. The signals were compressed in domains with different time-frequency resolutions according to their energy distributions in those domains, i.e. a signal whose energy was more concentrated in a domain with high time resolution was compressed in the time domain, while a frequency domain signal was compressed in the frequency domain. They showed that their method was simple to implement and effective for compressing audio and speech at bit rates as low as 2 kbps.

Najih et al evaluated the wavelet compression technique on speech signals [38]. They evaluated a number of wavelet filters to determine the most suitable filters for providing low bit rate and low computation complexity. Their speech compression technique employed five procedures: ‘one-dimensional wavelet decomposition’; ‘thresholding’; ‘quantization’; ‘Huffman coding’; ‘reconstruction using several wavelet filters’. Najih et al evaluated their method using peak signal to noise ratio (PSNR), signal to noise ratio (SNR) and normalized root mean squared error (NRMSE). Their results showed that the Daubechies-10 wavelet filter gave higher SNR and better speech quality than other filters. They achieved a compression ratio of 4.31 times with satisfactory quality of decoded speech signals.

Farooq and Datta proposed a pre-processing stage based on wavelet denoising for extracting robust MFCC features in the presence of additive white Gaussian noise [14]. They found that MFCC features extracted after denoising were less affected by Gaussian noise and improved recognition by 2 to 28 % for signal-to-noise ratios in the range 20 to 0 dB.

Barros et al developed a system for enhancement of the speech signal with highest energy from a linear convolute mixture of n statistically independent sound sources recorded by m microphones, where m<n [2]. In their system, adaptive auditory filter banks, pitch tracking, and the concept of independent component analysis were used. Wavelets were used in the process of extracting the speech fundamental frequency and as a bank of adaptive bandpass filters. They constructed a bandpass filter, using wavelets, centered around the central frequency given at each time instant by a quantity they termed the driver. The driver was defined as the frequency value corresponding to the maximum value of the speech spectrogram at each time instant in a given frequency range.

Their filter banks where centered at the fundamental frequency and its harmonics, thus mimicking the nonlinear scaling of the cochlea. They used a modified Gabor function. Where they had access to the original signal, Barros et al used objective quality measures to evaluate their system, and their results showed good performance. For the cases where there was no access to the original signal, they measured subjective quality by the MOS scale, which is a five-point scale providing the options Excellent, Good, Fair, Poor, and Bad. Using this scale, the enhanced speech was generally regarded as good when compared to the mixed speech signal, which was generally regarded as poor.

Yao and Zhang investigated the bionic wavelet transform (BWT) for speech signal processing in cochlear implants [51]. The BWT is a modification of a wavelet transform that incorporates the active cochlear mechanism into the transform, resulting in a nonlinear adaptive time-frequency analysis. When they compared speech material processed with the BWT to that processed with the WT, they concluded that application of the BWT in cochlear implants has a number of advantages, including improved recognition rates for both vowels and consonants, reduction in the number of channels in the cochlear implant, reduction in the average stimulation duration for words, better noise tolerance and higher speech intelligibility rates.

Bahoura and Rouat proposed a wavelet speech enhancement scheme that is based on the Teager energy operator [1]. The Teager energy operator is a nonlinear operator that is capable of extracting signal energy based on mechanical and physical considerations [22]. Their speech enhancement process was a wavelet thresholding method where the discriminative threshold in various scales was time adapted to the speech waveform. They compared their speech enhancement results with those obtained using an algorithm by Ephraim et al [12] and concluded that their scheme yields higher SNR. Unlike the speech enhancement method of Ephraim et al, the method of Bahoura et al did not require explicit estimation of the noise level or a priori knowledge of the signal-to-noise ratio (SNR).

Favero devised a method to compound two or more wavelets and used the compounded wavelet to compute the sampled CWT (SCWT) of a speech signal [16]. He used the compound-wavelet computed SCWT coefficients as input parameters for a speech recognition system. Favero found that using the compound wavelet decreases the number of coefficients input to a speech recognition system and improves recognition accuracy by about 15 per cent.

Kadambe and Srinivasan used adaptive wavelet coefficients as input parameters to a phoneme recognizer [21]. The wavelet was adapted to the analyzed speech signal by choosing the sampling points on the scale and time axes according to the speech signal. This adaptive sampling was achieved using conjugate gradient optimization and neural networks. The adaptive wavelet based phoneme recognizer produced results that were comparable to cepstral based phoneme recognizers.

2.3 VARIABLE FRAME RATE CODING OF SPEECH

Variable frame rate (VFR) techniques allow for the reduction of frames processed by a front-end automatic speech recognizer (ASR) and, importantly for this study, the identification of speech transients. To reduce the amount of data processed and improve recognition performance, the VFR technique varies the rate at which acoustic feature vectors are selected for input to an ASR system. A higher frame rate is used where the feature vectors change rapidly while a lower frame rate is used when feature vectors change slowly. Acoustic feature vectors evaluated for VFR coding of speech for this study are linear prediction code (LPC) and Mel-frequency cepstral coefficients (MFCC).

A description of LPC and MFCC is given below focusing on how these feature vectors are created from speech. This will be followed by a discussion of VFR.

2.3.1 Linear Prediction Analysis

Linear prediction analysis has found widespread use in speech processing, particularly speech recognition. This section gives a brief description of how the linear prediction parameters (code) are obtained from speech. A detailed explanation of linear prediction analysis may be found in [11] and [43].

2.3.1.1 Long-term Linear Prediction Analysis

The objective of Linear Prediction (LP) is to identify, for a given speech signal s(n), the parameters ($\hat{a}(i)$) of an all-pole speech characterization function given by;

$$H(z) = G\frac{1}{1-\sum_{i=1}^{M}\hat{a}(i)z^{-i}} \qquad (2.18)$$

with excitation sequence

$$u(n) = \begin{cases} \sum_{q=-\infty}^{\infty}\delta(n-qP) & voiced \\ noise & unvoiced \end{cases} \qquad (2.19)$$

In Eq. 2.18, H(z) is a filter that represents the vocal tract, G is the gain and M is the order of the LPC analysis. The all-pole nature of the LP characterization of speech means that the magnitude of the spectral dynamics of the speech is preserved while the phase characteristics are not. Typical values for the order of LPC analysis (M) are 8 to 16 [43]. Figure 2.12 shows a block diagram for the linear prediction model of speech synthesis.

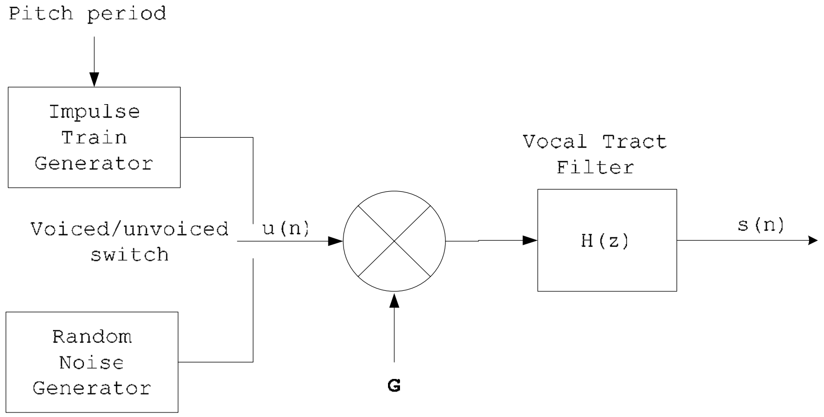

Figure 2.12: LP speech synthesis model

From Figure 2.12, the relation between u(n) and s(n) is

$$S(z) = GH(z)U(z)$$

$$S(z) = G\frac{1}{1-\sum_{i=1}^{M}\hat{a}(i)z^{-i}}U(z)$$

$$S(z) = \sum_{i=1}^{M}\hat{a}(i)S(z)z^{-i} + GU(z) \tag{2.20}$$

In the time domain, the relation is

$$s(n) = \sum_{i=1}^{M} \hat{a}(i) s(n-i) + Gu(n) \tag{2.21}$$

Except for the excitation sequence, s(n) can be *predicted* using a *linear* combination of its past values, hence the name *linear prediction*. The $\hat{a}(i)$'s form the prediction equation coefficients and their estimates are called the *linear prediction code* [11].

Linear Prediction Equations

The input and output to the block diagram of Figure 2.12 are known but the transfer function is unknown. The problem is to find H'(z) (estimate of the true frequency response) such that the mean squared error between the true speech s(n) and the estimated speech s'(n) is minimized. From Figure 2.13 (a) we realize that the $\hat{a}(i)$'s are nonlinearly related to H'(z), which makes the problem of determining the $\hat{a}(i)$'s a difficult one. The problem can be simplified by considering the inverse model shown in Figure 2.13 (b). In this model, the $\hat{a}(i)$'s are linearly related to $H'_{\text{inv}}(z)$ and

$$H'_{inv}(z) = \alpha(0) + \sum_{i=1}^{\infty} \alpha(i) z^{-i} \tag{2.22}$$

Imposing the constraints $\alpha(0)$ = 1 and $\alpha(i)$ = 0 for i>M, the problem now reduces to finding a finite impulse response (FIR) filter of length M+1 that minimizes the mean squared error between the true excitation sequence u(n) and the estimated excitation sequence u'(n). The LP parameters are then given by $\hat{a}(i)$ = -$\alpha(i)$, $\alpha(i)$ being the coefficients of the inverse filter $H'_{inv}(z)$.

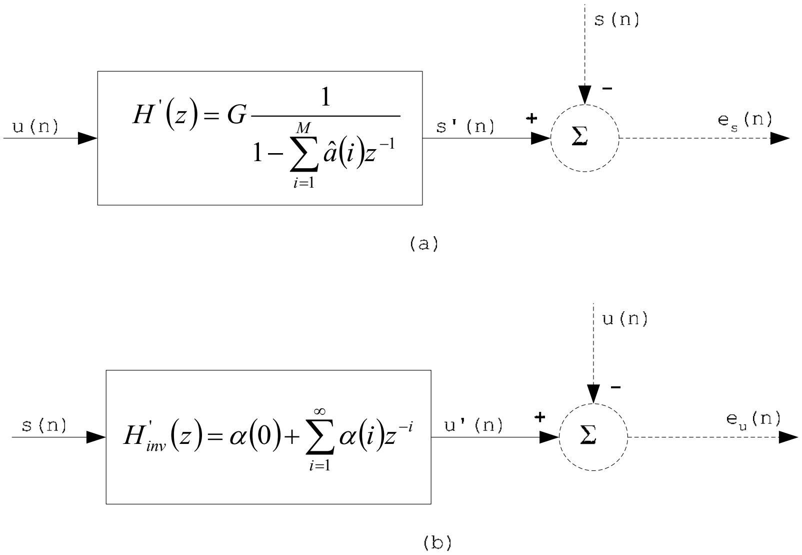
s(n)
-
u(n)
$H'(z) = G\frac{1}{1-\sum_{i=1}^{M}\hat{a}(i)z^{-1}}$
s'(n)
+
Σ
e_s(n)
(a)
u(n)
-
s(n)
$H'_{inv}(z) = \alpha(0) + \sum_{i=1}^{\infty}\alpha(i)z^{-i}$
u'(n)
+
Σ
e_u(n)
(b)

Figure 2.13: (a) Estimated model and (b) Inverse model

From Figure 2.13 (b), Eq. 2.22 and using $\alpha(0)$ = 1 we have the mean squared error

$$E_u = \sum_n e_u(n) = \sum_n \left[u'(n) - u(n)\right]^2$$

$$E_u = \sum_n \left[s(n) + \sum_{i=1}^{M} \alpha(i) s(n-i) - u(n) \right]^2 \qquad (2.23)$$

Differentiating Eq. (2.23) with respect to $\alpha(\eta)$ and setting the result to zero, we have

$$\frac{\partial E_u}{\partial \alpha(\eta)} = 2\sum_n \left[s(n) + \sum_{i=1}^{M} \alpha(i) s(n-i) - u(n) \right] s(n-\eta) = 0$$

$$\sum_n s(n)s(n-\eta)+\sum_{i=1}^{M}\alpha(i)\sum_n s(n-i)s(n-\eta)-\sum_n u(n)s(n-\eta)=0$$

$$\phi_{ss}(\eta)+\sum_{i=1}^{M}\alpha(i)\phi_{ss}(i-\eta)-\phi_{us}(\eta)=0 \tag{2.24}$$

where $\phi_{ss}(\eta)$ is the time autocorrelation of s(n) and $\phi_{us}(\eta)$ is the time cross-correlation of the sequences s(n) and u(n). The assumption that the sequences s(n) and u(n) are wide-sense stationary (WSS) has been made. If we also assume that the excitation sequence is a unity-variance orthogonal random process, i.e., $\phi_{ee}(\eta)=\delta(n)$, then $\phi_{us}(\eta)=C\delta(\eta) \quad \textit{for} \quad \eta \geq 0$, which is therefore zero for positive η [11].

Recalling that $\hat{a}(i)$ = - $\alpha(i)$, Eq. (2.24) becomes

$$\sum_{i=1}^{M}\hat{a}(i)\phi_{ss}(i-\eta)=\phi_{ss}(\eta) \tag{2.25}$$

The M Eqs of (2.25), sometimes called the *normal equations,* are used to compute the conventional LP parameters [11].

Since speech is considered quasi-steady state only over short time intervals, computing long-term LP parameters for a short speech segment of interest would give bad estimates. Short-term LP analysis resolves this problem by computing LP parameters in the interval of interest only.

2.3.1.2 Short-term Linear Prediction Analysis

There are two well known short-term LP techniques; the autocorrelation method and the covariance method.

Autocorrelation Method

The N sample *short-term autocorrelation function* for the signal s(n) is defined as [11]

$$\phi_{ss}(\eta;m)=\frac{1}{N}\sum_{n=-\infty}^{\infty}s(n)w(m-n)s(n-|\eta|)w(m-n+|\eta|) \tag{2.26}$$

where w(n) is a window function which is zero outside the interval of N points ending at m. Using the short-term autocorrelation function in the long-term normal equation gives

$$\sum_{i=1}^{M}\hat{a}(i;m)\phi_{ss}(\eta-i;m)=\phi_{ss}(\eta;m) \tag{2.27}$$

In matrix notation, Eq. (2.27) can be expressed as

$$R_{ss}(m)\hat{a}(m)=\phi_{ss}(m) \tag{2.28}$$

with $R_{ss}(m)$ having the form

$$R_{ss}(m)=\begin{bmatrix} \phi_{ss}(0) & \phi_{ss}(1) & \phi_{ss}(2) & \cdots & \phi_{ss}(M-1) \\ \phi_{ss}(1) & \phi_{ss}(0) & \phi_{ss}(1) & \cdots & \phi_{ss}(M-2) \\ \phi_{ss}(2) & \phi_{ss}(1) & \phi_{ss}(0) & \cdots & \phi_{ss}(M-3) \\ \vdots & \vdots & \vdots & & \vdots \\ \phi_{ss}(M-1) & \phi_{ss}(M-2) & \phi_{ss}(M-3) & \cdots & \phi_{ss}(0) \end{bmatrix} \tag{2.29}$$

The M-by-M matrix ($R_{ss}(m)$) of autocorrelation values known as the *short-term autocorrelation* matrix is a *Toeplitz* matrix and can be solved using the *Durbin algorithm* [11].

Covariance Method

The N sample *short-term covariance function* for the signal s(n) for time in the interval m-N+1 < n < m is defined as [11]

$$\varphi_{ss}(\alpha, \beta; m) = \frac{1}{N} \sum_{n=m-N+1}^{m} s(n-\alpha)s(n-\beta). \tag{2.30}$$

The covariance estimator of the LP parameters is obtained by using the short-term covariance function (Eq. 2.30) as an estimate of autocorrelation in the long-term normal equations (Eq. 2.25). This gives

$$\sum_{i=1}^{M} \hat{a}(i;m)\varphi_{ss}(i,v;m) = \varphi_{ss}(0,v,m). \tag{2.31}$$

In matrix notation, Eq. 2.31 can be expressed as

$$\Phi_{ss}(m)\hat{a}(m) = \varphi_{s}(m) \tag{2.32}$$

with $\Phi_{ss}(m)$ having the form

$$\Phi_{ss}(m) = \begin{bmatrix} \varphi_{ss}(1,1) & \varphi_{ss}(1,2) & \varphi_{ss}(1,3) & \cdots & \varphi_{ss}(1,M) \\ \varphi_{ss}(2,1) & \varphi_{ss}(2,2) & \varphi_{ss}(2,3) & \cdots & \varphi_{ss}(2,M) \\ \varphi_{ss}(3,1) & \varphi_{ss}(3,2) & \varphi_{ss}(3,3) & \cdots & \varphi_{ss}(3,M) \\ \vdots & \vdots & \vdots & & \\ \varphi_{ss}(M,1) & \varphi_{ss}(M,2) & \varphi_{ss}(M,3) & \cdots & \varphi_{ss}(M,M) \end{bmatrix} \tag{2.33}$$

The M-by-M matrix $\Phi_{ss}(m)$ of covariance values known as the *short-term covariance matrix* can be solved using the *Cholesky decomposition method.*

Note that the covariance method does not involve the use of a window function; it is computed over a range of points and uses the unweighted speech directly. Of the two methods for computing short-term LP parameters, the autocorrelation method has found the most extensive use in speech processing.

2.3.2 Mel-Frequency Cepstral Coefficients

Today, most automatic speech recognizers use Mel-frequency Cepstral coefficients (MFCC), which have proven to be effective and robust under various conditions [36]. MFCC capture and preserve significant acoustic information better than LPC [10]. MFCC have become the dominant features used for speech recognition and the following discussion of MFCC will follow the description of [29].

Figure 2.14 shows the process for creating MFCC features from a speech signal. The first step is to convert the speech into frames by applying a windowing function; frames are typically 20 to 30 ms in duration with a frame overlap of 2.5 to 10 ms. The window function (typically a Hamming window) removes edge effects at the start and end of the frame. A cepstral feature vector is generated for each frame.

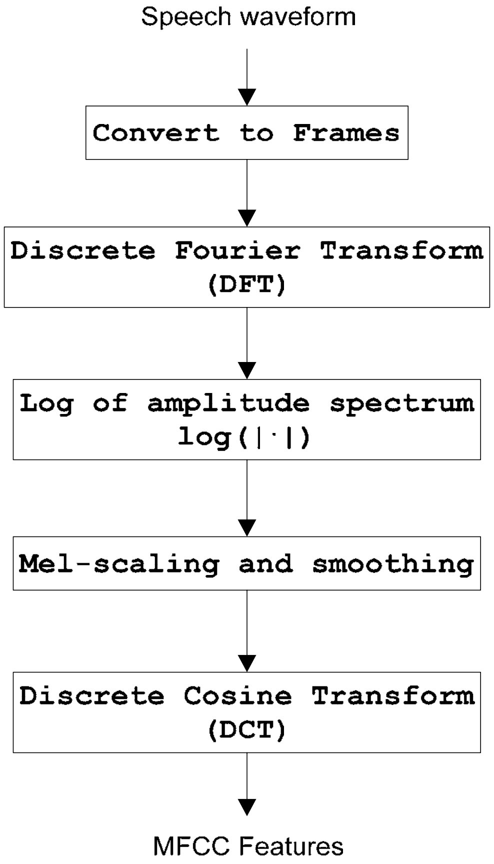
Speech waveform
Convert to Frames
Discrete Fourier Transform (DFT)
Log of amplitude spectrum log(|·|)
Mel-scaling and smoothing
Discrete Cosine Transform (DCT)
MFCC Features

Figure 2.14: Process to create MFCC features from speech

The next step is to compute the discrete Fourier transform (DFT) for each frame. Then the logarithm of the amplitude spectrum of the DFT is computed. Computing the amplitude of the DFT discards the phase information but retains the amplitude information which is regarded as the most important for speech perception [30]. In the next step, the Fourier spectrum is smoothed using filter-banks arranged on a *mel-scale.* The mel-scale emphasizes perceptually meaningful frequencies. This scale is approximately linear up to 1000 Hz and logarithmic thereafter. In the final step, the discrete cosine transform (DCT) is computed. The discrete cosine transform, which is used here as an approximation of the Karhunen-Loeve (KL) transform, has the effect of decorrelating the log filter-bank coefficients and compressing the spectral information into the lower-order coefficients.

LPC Parameter Conversion to Cepstral Coefficients

LPC cepstral coefficients, c_m, can be derived directly from LPC parameters by using the recursive formulas [43]

$$c_0 = \ln(G) \tag{2.34}$$

$$c_m = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k}, \quad 1 \le m \le M \tag{2.35}$$

$$c_m = \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k}, \quad m > M \tag{2.36}$$

where G is the gain term in the LPC model, a_m are the LPC coefficients, and M is the order of the LPC analysis. The cepstral coefficients have been shown to be a more robust and reliable feature set for speech recognition than the LPC coefficients [43]. Generally, a cepstral representation with N > M coefficients is used in speech recognition, where $N \approx \left(\frac{3}{2}\right)M$.

2.3.3 Variable Frame Rate Techniques

In most speech recognizers, a speech signal is windowed into frames (typically of 20 – 30 ms duration) with a certain fixed overlap between adjacent frames. Windowing is done with the assumption that speech is not stationary, but exhibits quasi-stationary properties over short segments of time. Each frame is then represented with feature vector parameters such as MFCC or LPC. These parameters are used in the pattern matching stage of the recognizer.

In the vowel parts of speech, parameters of successive frames may look much alike and computing parameters every 20 to 30 ms may be redundant. Variable frame rate techniques take advantage of this by picking more frames where parameters of successive frames are different and few where parameters are similar. This reduces the computational load of speech recognizers without performance loss.

A number of variable frame rate (VFR) analysis methods for speech recognition have been proposed for use in automatic speech recognizers [40], [56], [23].

Ponting and Peeling proposed a VFR technique where the Euclidean distance between the current frame and the last retained frame was used in the frame picking decision [40]. This method will be referred to as the *classical method.* In Ponting and Peeling's VFR technique, a frame was picked if the Euclidean distance between that frame and the last retained frame was greater than a set threshold.

Zhu and Alwan improved on the classical VFR technique by weighing the Euclidean distance with the log energy of the current frame [56]. They also proposed a new frame picking method where a frame was picked if the accumulated weighted

Euclidean distance was greater than a set threshold. Their method will be referred to as the *log-energy method.*

Le Cerf and Van Compernolle proposed a *derivative* VFR technique, where the Euclidean norm of the first derivatives of the feature vectors was used as the decision criteria for frame picking [23]. The derivative method VFR technique discards a frame if the Euclidean norm of that frame is less than a chosen threshold.

The fact that variable frame rate techniques pick more frames when there is rapid change and fewer frames elsewhere suggests that they can be used for identification of transients in speech. In the classical method, only two frames are considered in the decision-making process, and this does not represent completely the whole environment of the frame [23] [24]. The calculation of derivatives takes into account the whole environment of the frame and is able to measure the change in the signal better [23] [24]. For this reason the derivative method VFR technique was used for detection of transients in speech in this study.

2.4 DECOMPOSING SPEECH USING THE FORMANT TRACKING ALGORITHM

Yoo et al applied multiple time-varying band-pass filters, based on a formant tracking algorithm by Rao and Kumaresan, to track speech formants [44], [52], [53]. The formant tracking algorithm applied multiple dynamic tracking filters (DTF), adaptive all-zero filters (AZF), and linear prediction in spectral domain (LPSD) to estimate the frequency modulation (FM) information and the amplitude modulation (AM) information. The FM information was then used to determine the center frequencies of the DTF and to update the pole and zero locations of the DTF and the AZF. The AM information was used to estimate the bandwidth of the time-varying band-pass filters.

The output of each time-varying band pass filter was considered to be an estimate of the corresponding formant. The sum of the outputs of the filters was defined as the *tonal component* of the speech. Yoo et al estimated the *non-tonal component* of the speech signal by subtracting the tonal component from the original speech signal. Yoo et al considered the tonal component to contain most of the steady-state information of the input speech signal and the non-tonal component to contain most of the transient information of the input speech signal. A block diagram of the formant tracking speech decomposition scheme is shown in Figure 2.15.

In the present study, the tonal and nontonal speech components obtained from the formant tracking algorithm of Yoo et al will be used as reference signals to which the quasi-steady-state and transient speech components synthesized from wavelet representations will be compared.

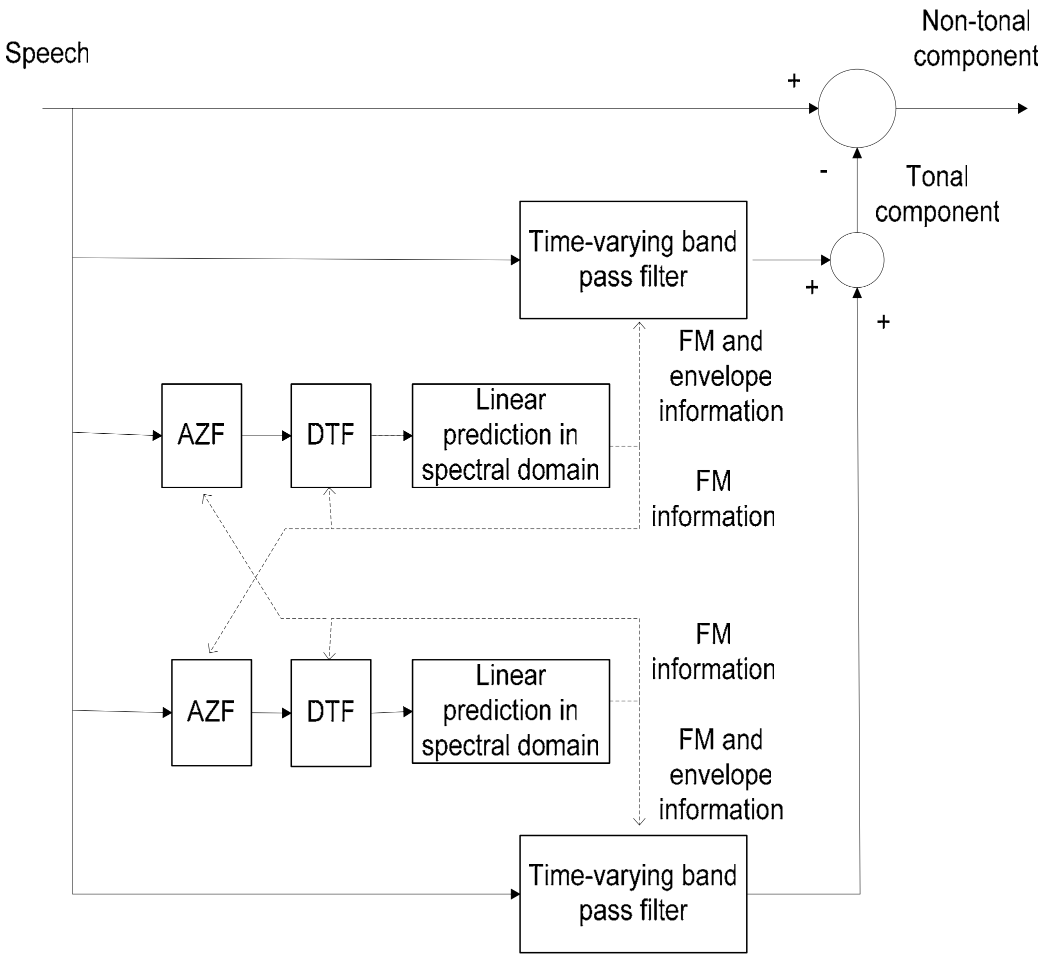
Speech
Non-tonal component
+
-
Tonal component
Time-varying band pass filter
+
+
FM and envelope information
AZF
DTF
Linear prediction in spectral domain
FM information
FM information
AZF
DTF
Linear prediction in spectral domain
FM and envelope information
Time-varying band pass filter

Figure 2.15: Block diagram of formant tracking speech decomposition [55].

3.0 WAVELET TRANSFORMS AND PACKETS TO IDENTIFY TRANSIENT SPEECH

The process of using of the discrete wavelet transform (DWT), stationary wavelet transform (SWT) and wavelet packet (WP) for identifying transient and quasi-steady-state speech components is described here. As stated earlier, these speech components are based on and compared to the nontonal and tonal speech components defined by Yoo [52], [53] [54].

The analysis algorithms were implemented using MATLAB software. The wavelet toolbox, the Voicebox speech processing toolbox developed at Imperial College [3], and WaveLab802 developed by Donoho D., Duncan M. R., Huo, X. and Levi, O. at Stanford University were particularly important tools. Speech samples were obtained from the audio CDROM that accompanies Contemporary Perspectives in Hearing Assessment, by Frank E. Musiek and William F. Rintelmann, Allyn and Bacon, 1999 (referred to as CDROM # 1). These speech signals were downsampled from 44100 Hz to 11025 Hz and highpass filtered at 700 Hz. Yoo's formant tracking algorithm worked better when the first formant was removed. The highpass filtered speech signals were as intelligible as the original speech signals, as shown by psychoacoustic studies of growth of intelligibility as a function of speech amplitude [54].

Wavelet analysis is equivalent to a bank of bandpass filters that divides the frequency axis into logarithmic bandwidths when the DWT and SWT are used or into equal bandwidths when wavelet packets are used. The wavelet level concept as used to refer to the number of decimations performed in the DWT and SWT analysis may be thought of as an index label of the filter banks and is associated with a particular frequency interval. The terminal node label of wavelet packets analysis may be thought of in the same way.

Daubechies and Symlets wavelets of different orders were evaluated to determine the wavelet basis to use for the decompositions. The db20 wavelet function, shown in Figure 3.1, was chosen because it has a median time support length (3.54 ms). Results obtained using the db20 where not very different from those obtained using other Daubechies wavelets and Symlets wavelets of comparable time support. Wavelets with short time support, like db1 (Haar), do not have good frequency localization, and wavelet with long time support resulted in long computation times.

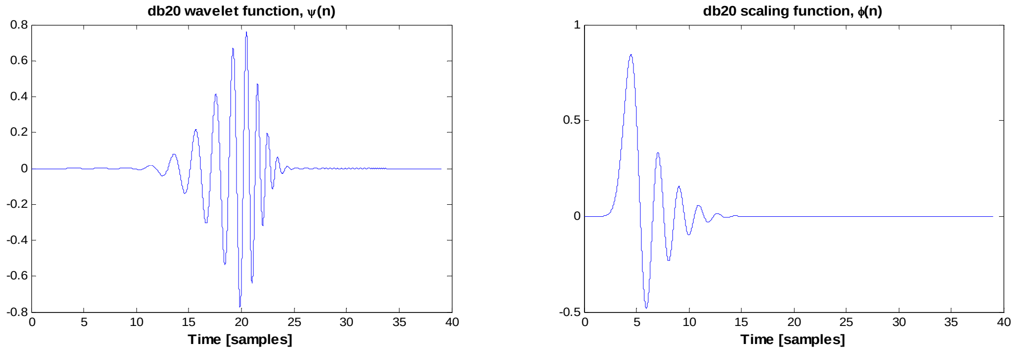

Figure 3.1: Wavelet and scaling functions for db20

3.1 METHOD FOR DISCRETE AND STATIONARY WAVELET TRANSFORMS

The DWT and SWT are similar in a number of ways. As a result, the procedures by which they were used for the identification of transient and quasi-steady-state speech components have several similarities and the methods used will be discussed together. Energy profiles, which are used to identify the dominant type of information (transient or quasi-steady-state information) of the wavelet coefficients at each level, will be defined first.

The highpass filtered, tonal and nontonal speech components were decomposed using a db20 wavelet function and a maximum decomposition level of 6. With this decomposition, the wavelet coefficients at level 5 and 6 and the scaling coefficients at level 6, which fall below the 700 Hz cutoff frequency, have very low energy. Using a decomposition level above 6 would not be beneficial since the wavelet coefficients at these higher levels would also have very low energy. Figure 3.2 shows, as a reference for the frequency intervals of each level, the filter frequency responses for levels 1 to 6. In this diagram, d*i*, *i* = 1, 2… 6 is the filter frequency response for the wavelet coefficient at level *i*, and a6 is the filter frequency response for the scaling coefficients at level 6.

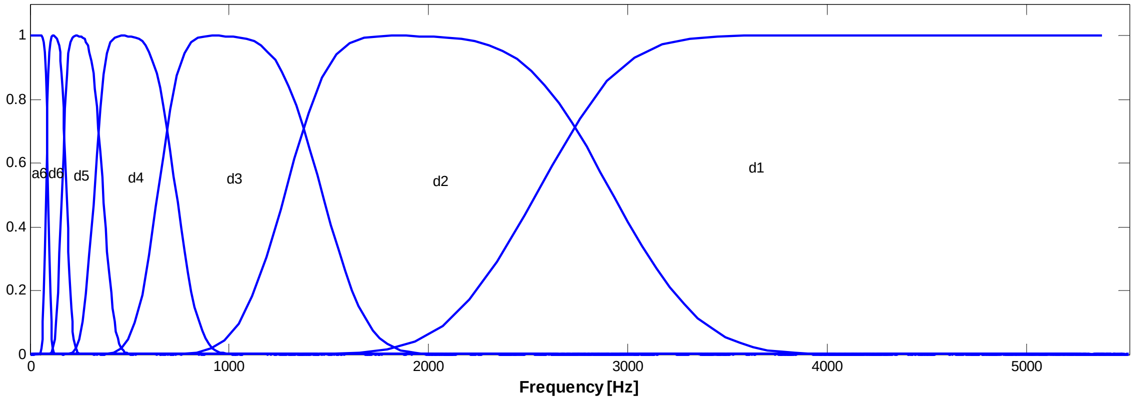

Figure 3.2: Filter frequency response at each level for a db20 wavelet function.

The energy distribution by level is used to identify wavelet levels which predominately include transient and quasi-steady-state information. This energy distribution will be refereed to as the *energy profile* for the word.

To identify wavelet levels with predominately transient or predominately quasi-steady-state information, the energy profile of the highpass filtered speech was compared to the energy profiles of Yoo's tonal and nontonal speech components. At a given level, if the energy of the wavelet coefficients of the highpass filtered speech was closer to the energy of the wavelet coefficients of the tonal speech, then the wavelet coefficients of the highpass filtered speech at that level are considered to have more quasi-steady-state information than transient information. On the other hand, if the energy of the wavelet coefficients of the highpass filtered speech was closer to the energy of the wavelet coefficients of Yoo's nontonal speech, then the wavelet coefficients of the highpass filtered speech at that level are considered to have more transient information.

Figure 3.3 shows, as an example, the energy profiles for the highpass filtered, nontonal and tonal speech components for the word ‘pike’ as spoken by a female obtained using the DWT and SWT. A db20 wavelet function was used for the level 6 decomposition. In this example, level 1, 2, 5 and 6 wavelet coefficients of the highpass filtered speech are considered to have transient information, since their energies are closer to energies of the wavelet coefficients of Yoo’s nontonal component at the same levels. Level 3 and 4 wavelet coefficients of the highpass filtered speech are considered to have quasi-steady-state information, since their energies are closer to the energies of the corresponding coefficients of the tonal component. Level 5 and 6 wavelet coefficients and scaling coefficients at level 6 had insignificant amounts of energy. A maximum decomposition level of 6 will be used in subsequent decomposition since any higher level will result in higher level wavelet coefficients of negligible energy.

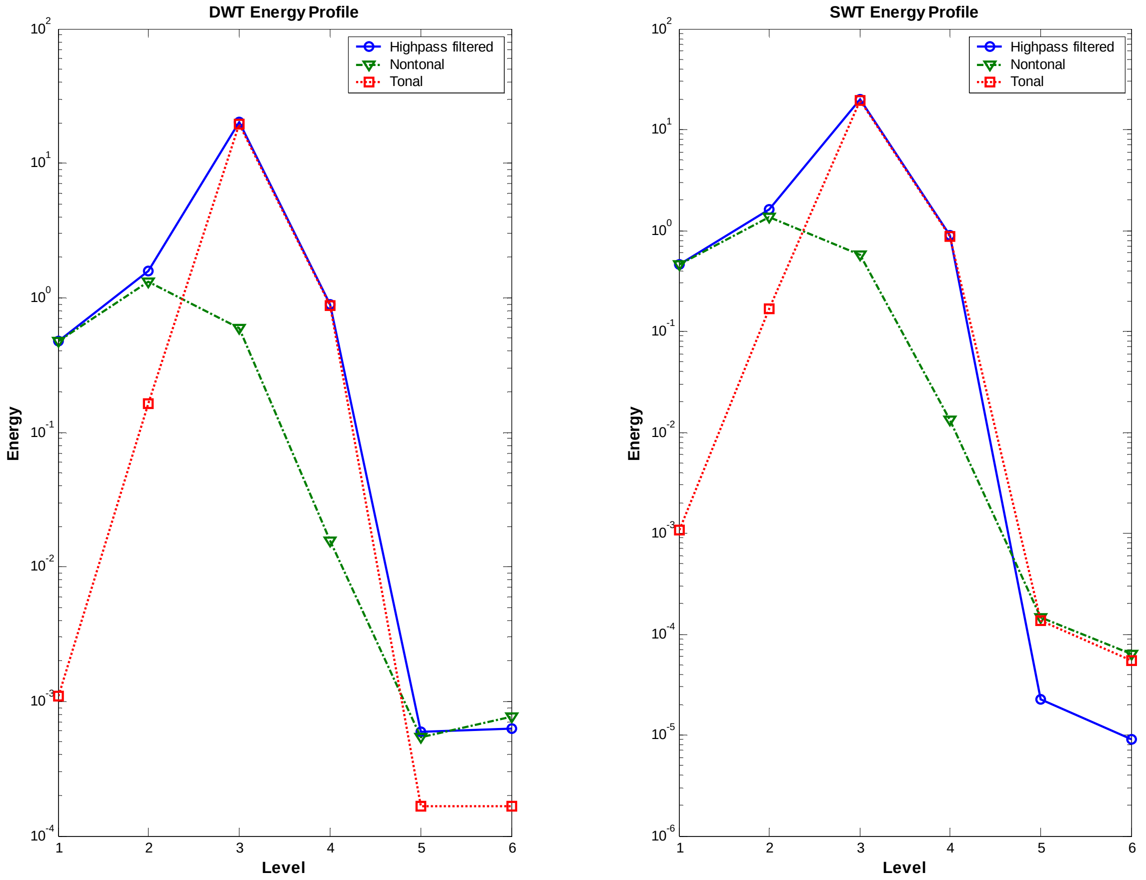

Figure 3.3: Energy profiles for the highpass filtered, nontonal and tonal speech components for the word 'pike' spoken by a female computed using the DWT and SWT.

In this example, level 1, 2, 5 and 6 wavelet coefficients of the highpass filtered speech are considered to have transient information, since their energies are closer to energies of the wavelet coefficients of Yoo's nontonal component at the same levels. Level 3 and 4 wavelet coefficients of the highpass filtered speech are considered to have quasi-steady-state information, since their energies are closer to the energies of the corresponding coefficients of the tonal component. Level 5 and 6 wavelet coefficients and scaling coefficients at level 6 had insignificant amounts of energy. A maximum decomposition level of 6 will be used in subsequent decomposition since any higher level will result in higher level wavelet coefficients of negligible energy.

After associating wavelet levels with either transient or quasi-steady-state speech, the inverse DWT and SWT are used to synthesize transient and quasi-steady-state speech components. A transient speech component is synthesized using wavelet levels that are identified to have transient information, and a quasi-steady-state speech component is synthesized using wavelet levels that are identified to have quasi-steady-state information. The synthesized speech components are compared, in the time- and frequency-domain, to Yoo's tonal and nontonal speech components. The spectra below 700 Hz and above 4 kHz, which relatively had low energy and did not contribute significantly to speech intelligibility, are ignored. An informal listening test was also used to compare the wavelet derived speech components to the speech components obtained using the algorithm of Yoo. The informal subjective listening test was conducted by the author listening to the speech components and making a judgment of how similar they sounded. These comparisons are a measure of how successful a wavelet transform was in identifying speech components that are a close estimate of the speech components of Yoo's algorithm.

As a means of comparing the transient component synthesized using wavelets to Yoo's nontonal component across many words, the estimation errors were computed for 18 words using the mean-squared-error (MSE) between the spectra of the two components.

3.2 RESULTS FOR DISCRETE AND STATIONARY WAVELET TRANSFORMS

This section presents examples of the results obtained when the DWT and SWT were used for decomposing speech into transient and quasi-steady-state components. For each type of wavelet transform, figures comparing the wavelet coefficients and energy profiles of the highpass filtered speech to those of Yoo’s tonal and nontonal speech are given. Following are figures comparing the transient and quasi-steady-state speech components, synthesized as described earlier, to Yoo’s nontonal and tonal speech components.

Discrete Wavelet Transform (DWT)

The DWT was explored for use in identifying transient and quasi-steady-state speech. Figure 3.4 shows, as an example, the DWT coefficients for the highpass filtered, tonal and nontonal speech components, and Figure 3.5 shows the energy profiles for these components. In Figure 3.4, in each column, level 0 is the original signal. For example, in the middle column, level 0 is Yoo’s nontonal component. Observing the energy profiles of Figure 3.5, the energy of the wavelet coefficients of the highpass filtered speech at level 1 is closer to the energy of the wavelet coefficients of Yoo’s nontonal speech component at level 1. These coefficients are considered to predominately include transient information. The wavelet coefficients of the highpass filtered speech at levels 2, 3, 4, 5 and 6 have energy that is closer to that of the wavelet coefficients of the tonal component than Yoo’s nontonal component at the same levels. Therefore these coefficients are considered to predominately include quasi-steady-state information.

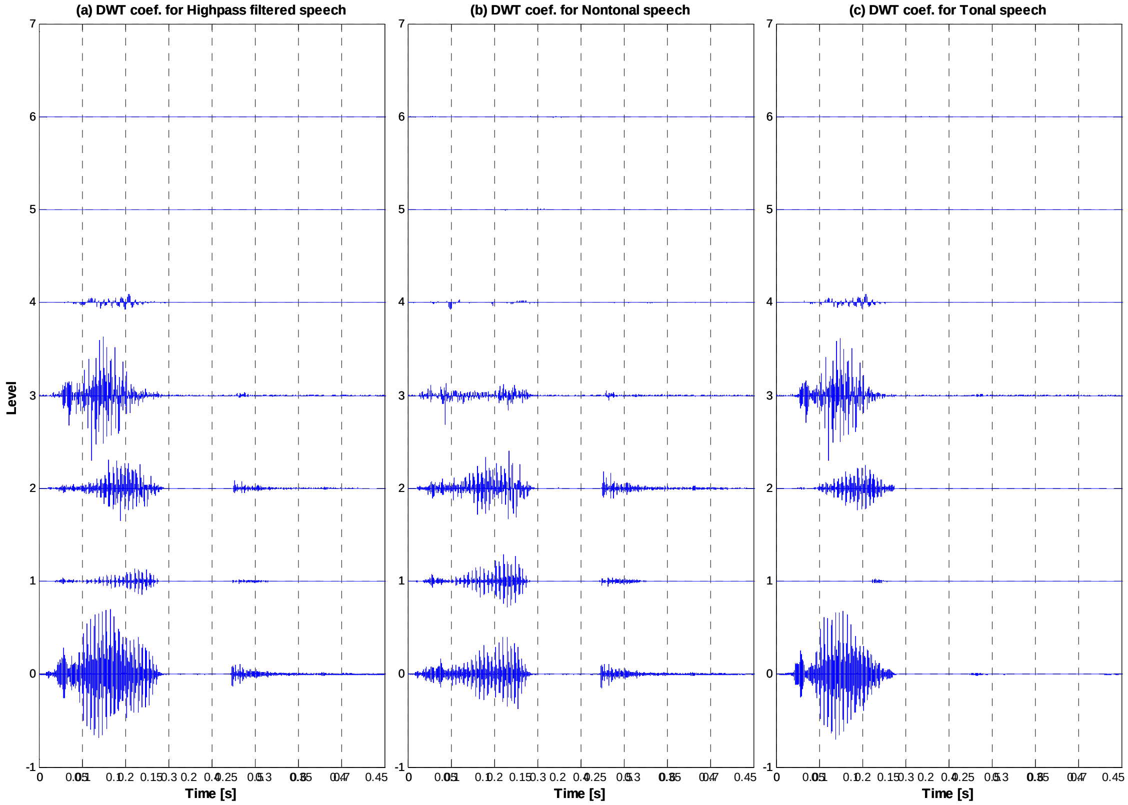

Figure 3.4: DWT coefficients for (a) highpass filtered speech, (b) nontonal speech and (c) tonal speech for the word ‘pike’ as spoken by a male.

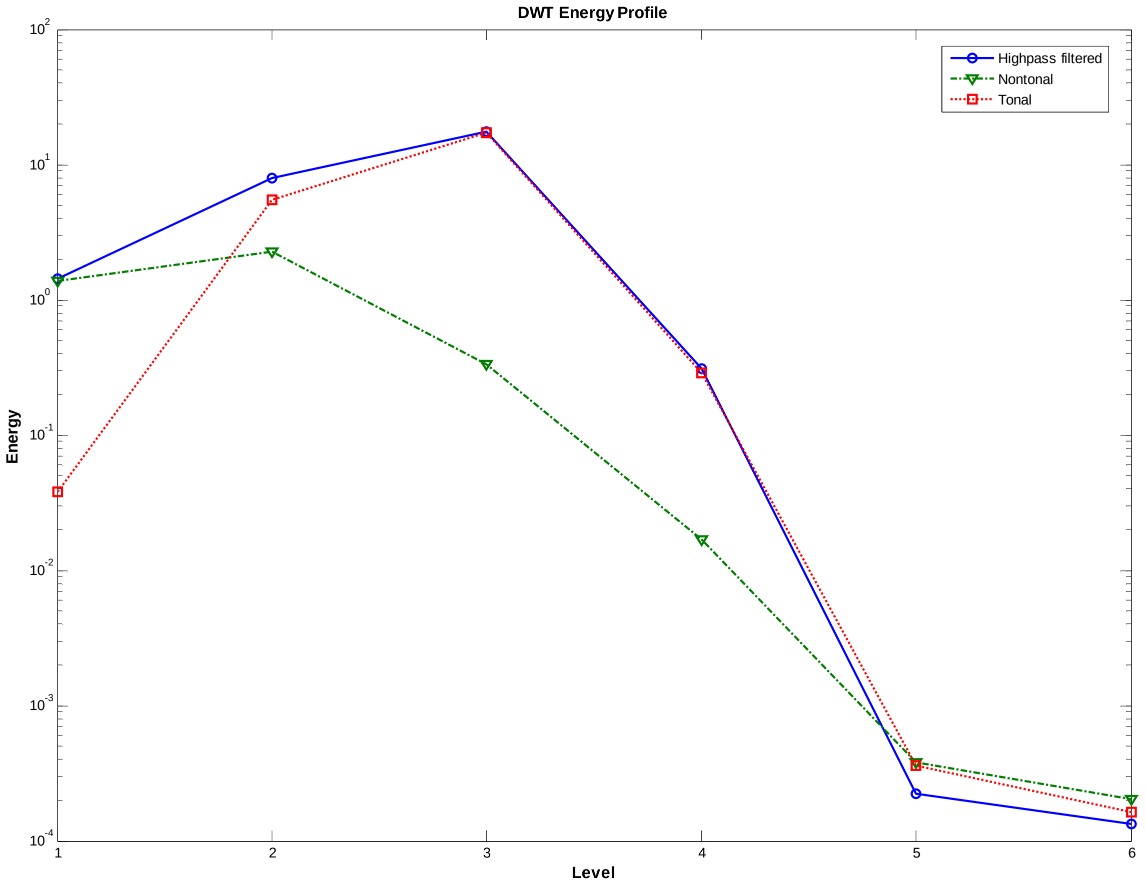

Figure 3.5: Energy profiles for the highpass filtered, nontonal and tonal speech components for the word ‘pike’ spoken by a male.

A transient speech component for the word ‘pike’ was synthesized from the level 1 wavelet coefficients of the highpass filtered speech. Figure 3.6 (c) and (c) show the DWT estimated transient component and the nontonal component, and Figure 3.7 (c) and (d) show their spectra. The spectrum of the transient component has little energy in the frequency interval of (0.7-1.5) kHz, where the spectrum of the nontonal component has significant energy. In the listening test, the transient speech component synthesized using the DWT was more whispered than and not as intelligible as the nontonal speech component.

The quasi-steady-state speech component for the word ‘pike’ was synthesized from the levels 2, 3, 4, 5 and 6 DWT coefficients of the highpass filtered speech. Figure 3.6 (a) and (b) show the quasi-steady-state component, estimated using the DWT and Yoo’s tonal component. The spectra of these two signals are shown in Figure 3.7 (a) and (b). The spectrum of the quasi-steady-state component includes frequencies present in the spectrum of Yoo’s tonal component and additional frequencies. The spectrum of the tonal component has some spectral peaks that where not observed in the spectrum of the quasi-steady-state component. In a listening test, the quasi-steady-state component synthesized using the DWT was more intelligible than Yoo’s tonal component.

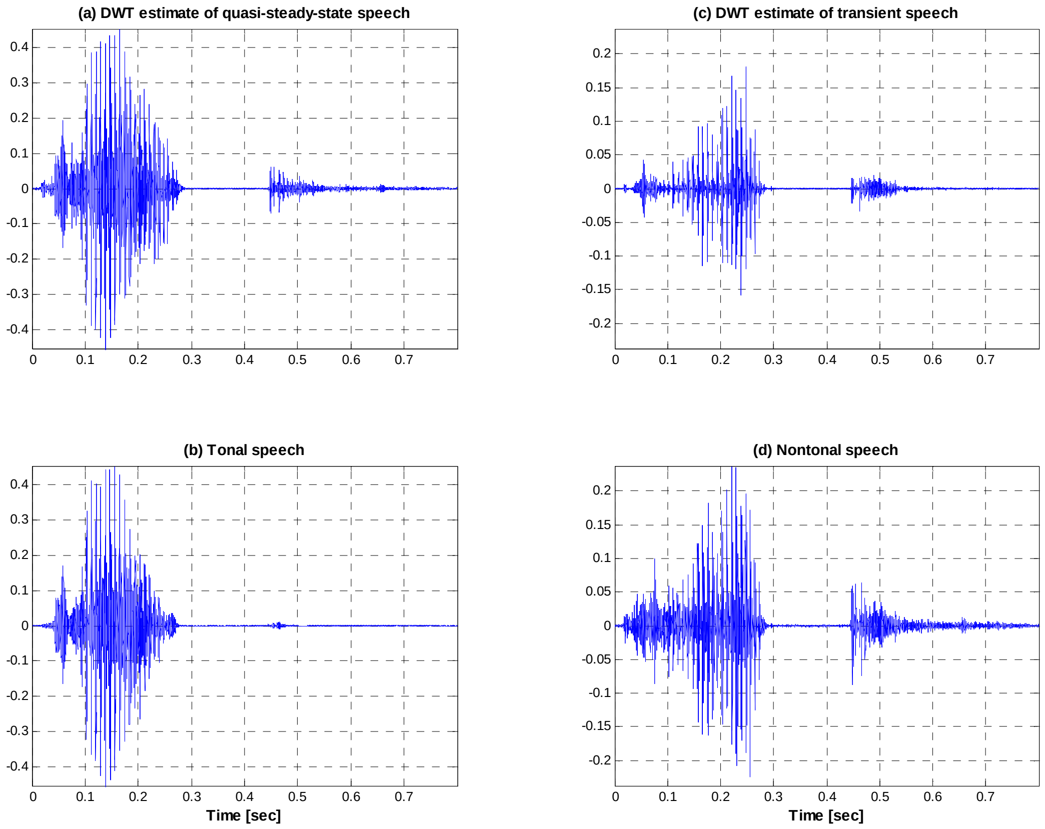

Figure 3.6: Time-domain plots of DWT estimate of quasi-steady-state and transient speech component, and of the tonal and nontonal speech components for the word ‘pike’ spoken by a male.

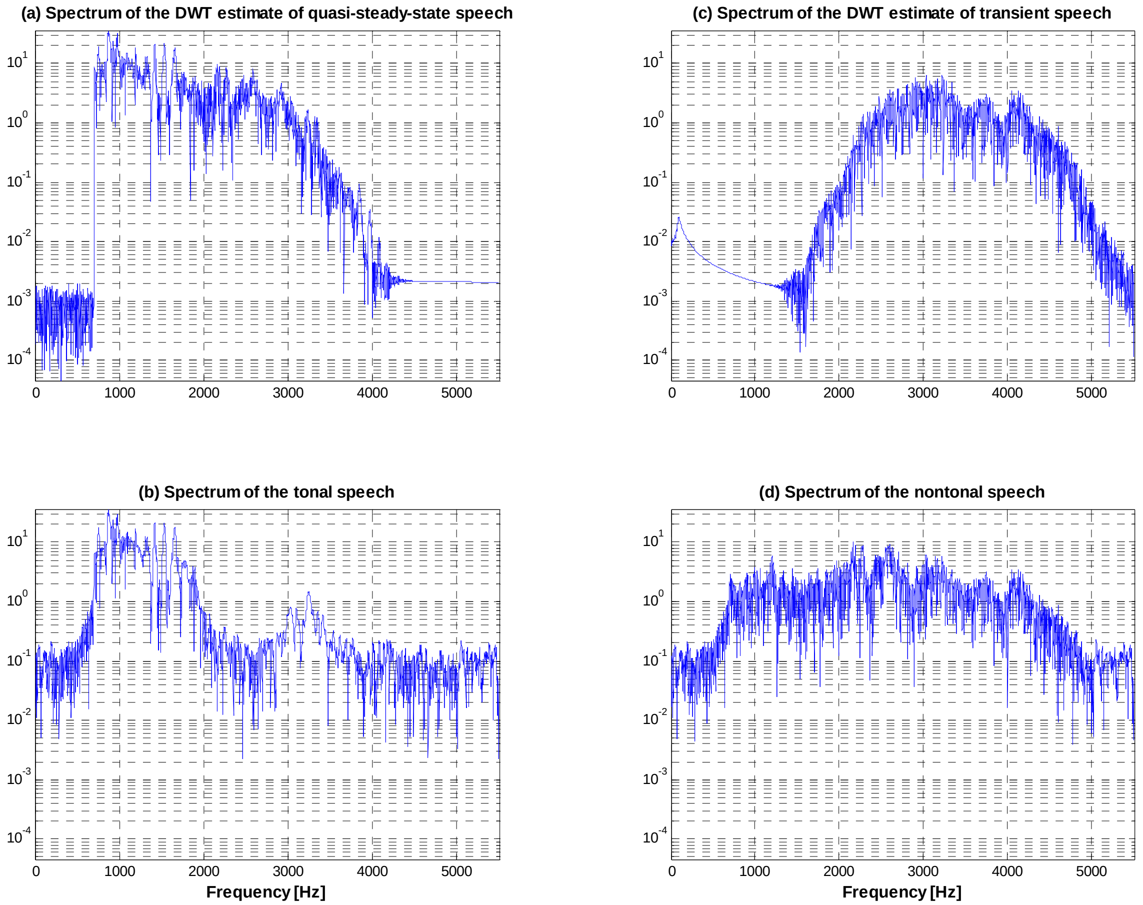

Figure 3.7: Frequency-domain plots of DWT estimate of quasi-steady-state and transient speech component, and of the tonal and nontonal speech components for the word ‘pike’ spoken by a male.

Figure 3.8 shows spectrograms for the quasi-steady-state and transient components synthesized using the DWT, and Yoo’s tonal and nontonal components for the word ‘pike’ spoken by a male. The spectrograms were computed using a 10 msec. Hamming window. The spectrograms show that, compared to Yoo’s tonal component, the quasi-steady-sate synthesized using the DWT is wideband and has some energy for t > 0.5 sec. The transient component synthesized using the DWT does not have energy for frequencies approximately less than 2 kHz. All these features were shown in the time-waveforms and spectra of the speech components, shown in Figures 3.6 and 3.7. The time-waveforms, though, also show that there is a difference in the characteristics of the release of the stop consonant /k/ (at approximately 0.45 sec.) observed for the transient component as compared to the nontonal component. These differences are not shown by the spectrograms, and for these reasons, the time-waveforms and spectra will be used for the remainder of this thesis, instead of spectrograms.

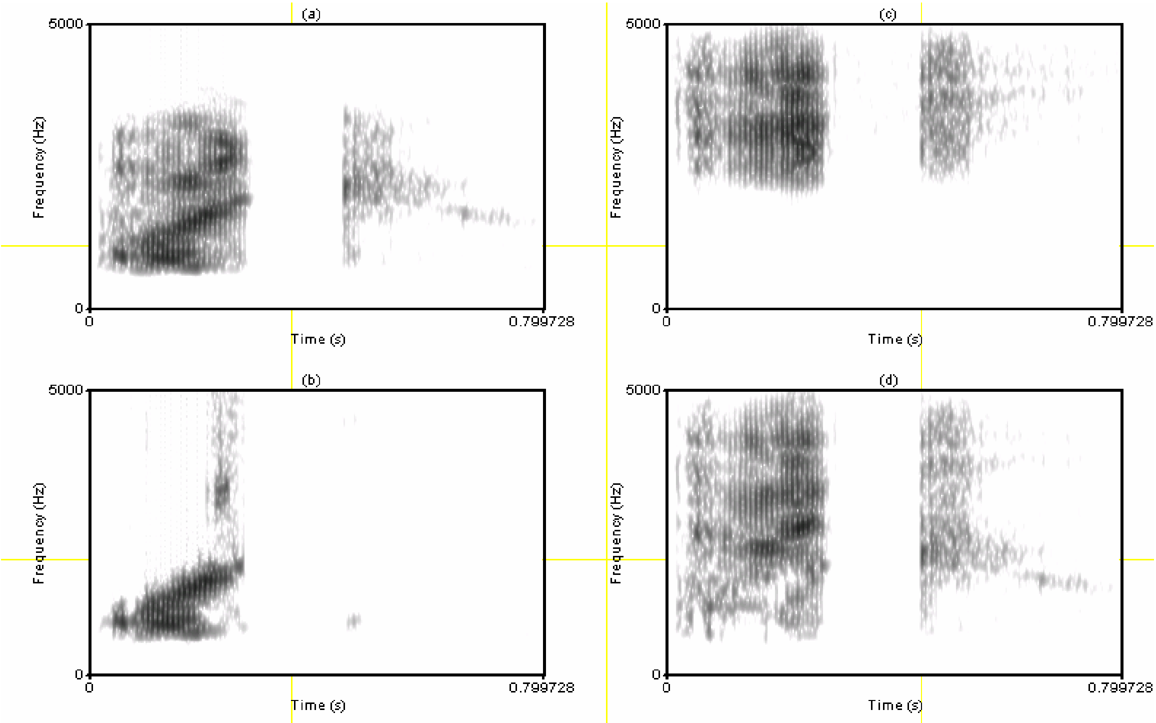

Figure 3.8: Spectrograms of (a) quasi-steady-state, (b) tonal, (c) transient and (d) nontonal speech components for the word ‘pike’ spoken by a male.

Level classifications for 18 words obtained using the DWT computed energy profiles are shown in Table A1 in the appendix. For most words, the wavelet coefficients at level 1, which constitute the upper half of the signal spectrum, were considered to have transient information. Level 3 wavelet coefficients, whose spectrum has its energy concentrated in the (700 1500) Hz frequency range, were identified as having quasi-steady-state information. The other levels were mixed. In general, transient components synthesized using the DWT were more whispered and less intelligible than Yoo's nontonal components, and quasi-steady-state components were more intelligible than Yoo's tonal components.

Stationary Wavelet Transform (SWT)

The use of the SWT to synthesize transient and quasi-steady-state components was explored using level 6 decomposition. As an example, Figure 3.9 shows the SWT coefficients for the highpass filtered, nontonal and tonal speech components for the word 'pike', spoken by a male, and Figure 3.10 shows their energy profiles. From the energy profiles shown, levels 1 and 5 were identified as having transient information and levels 2 and 3 were considered to have more quasi-steady-state information. Levels 5 and 6, which have very low energy, were classified as quasi-steady-state even though their energies were equally close to energies of Yoo's tonal and nontonal components at the same levels. This ambiguity will be resolved in Chapter 4.

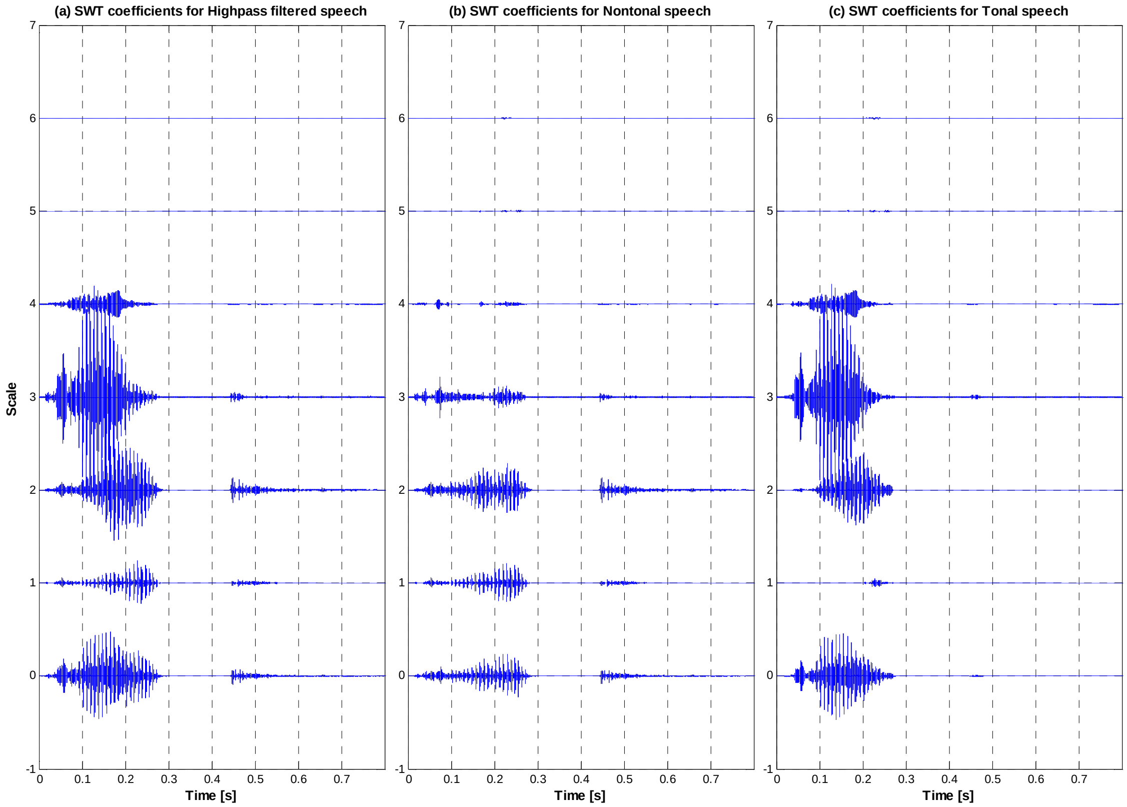

Figure 3.9: SWT coefficients for; (a) the highpass filtered speech, (b) the nontonal component and (c) the tonal component for the word ‘pike’ spoken by a male.

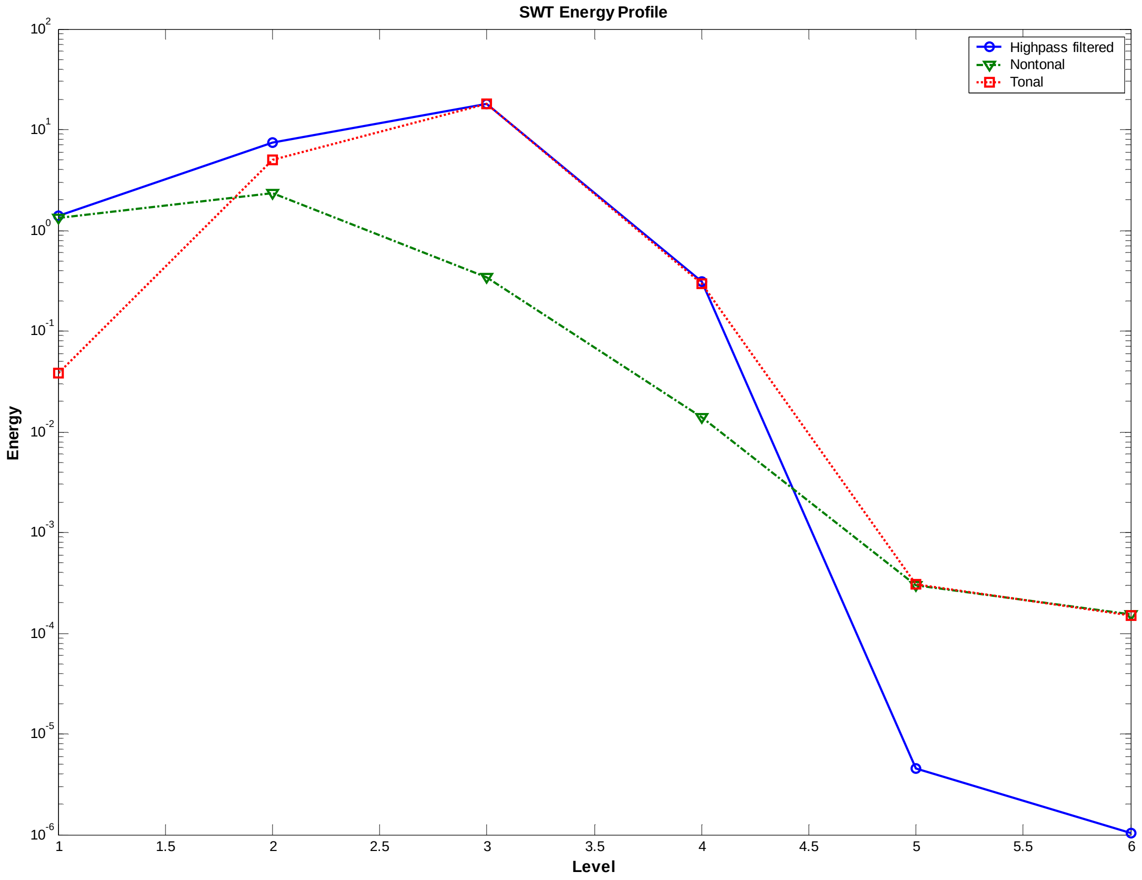

Figure 3.10: Energy profiles for the highpass filtered, nontonal and tonal speech components for the word ‘pike’ spoken by a male.

Transient and quasi-steady-state speech components for the word ‘pike’ spoken by a male were synthesized from the levels 1 and 5, and levels 2, 3, 4 and 6 SWT coefficients, respectively. Figure 3.11 compares the transient and quasi-steady-state speech components synthesized using the SWT to Yoo’s nontonal and tonal speech components respectively. Figure 3.11 compares the spectra of these speech components.

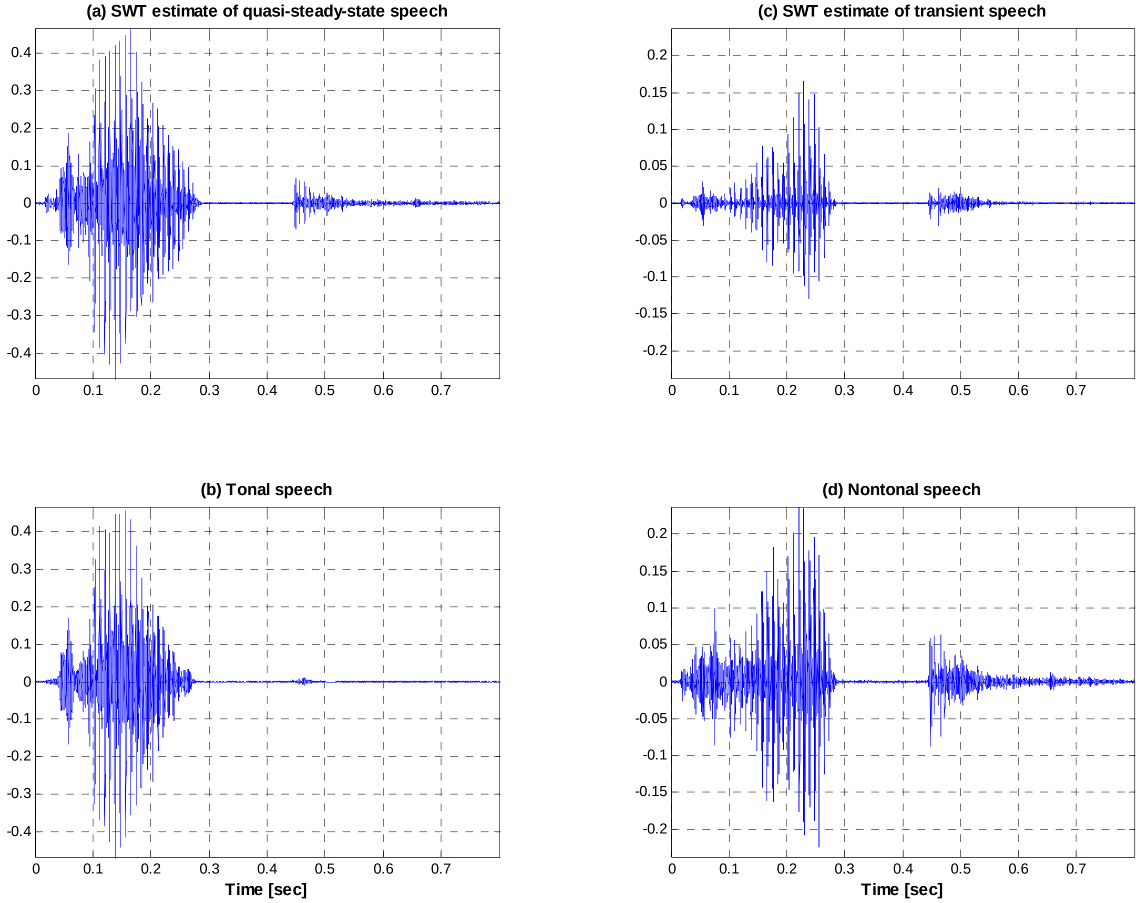

Figure 3.11: SWT estimated speech components, and the tonal and nontonal speech components of the word ‘pike’ spoken by a male.

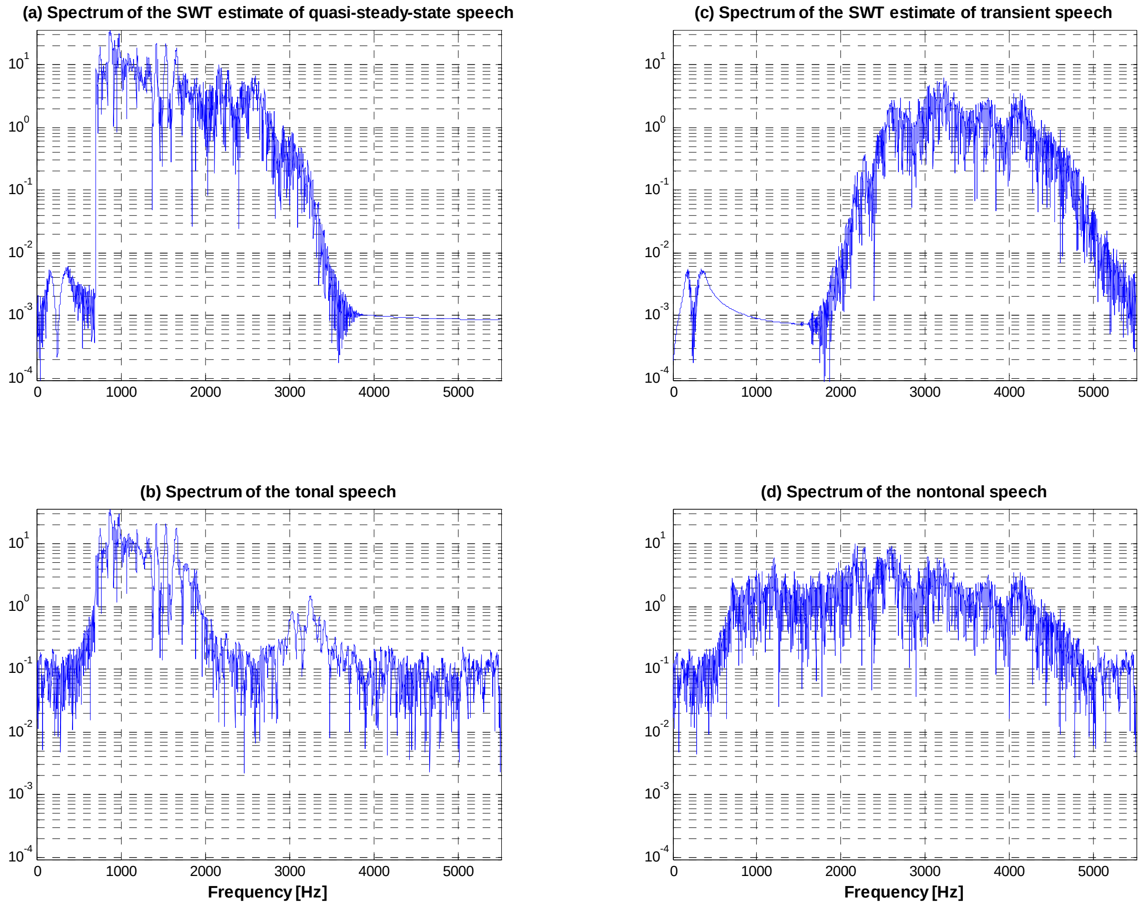

Figure 3.12: Spectra of SWT estimated speech components, and of the tonal and nontonal speech components of the word ‘pike’ spoken by a male.

The spectrum of the transient component had a narrower bandwidth than the spectrum of the nontonal component, while the spectrum of the quasi-steady-state component had a bandwidth wider than that of the tonal component. As in the DWT case, the transient component synthesized using the SWT was more whispered than the nontonal component and the quasi-steady-state component was more intelligible than the tonal component.

Table A2 in the appendix includes level classifications for 18 words obtained using the energy profiles computed using the SWT. Like the DWT, for most words, the wavelet coefficients at level 1, which constitute the upper half of the signal spectrum, were considered to have transient information. Level 3 wavelet coefficients, whose spectrum has energy concentrated in the (700 1500) Hz frequency range, were identified as having quasi-steady-state information.

In general, as observed with the DWT, transient components synthesized using the SWT were more whispered and less intelligible than the nontonal speech components, and quasi-steady-state components were more intelligible than the tonal components.

3.3 METHOD FOR WAVELET PACKETS

The DWT and SWT divide the signal spectrum into frequency bands that are narrow in the lower frequencies and wide in the higher frequencies. This limits how wavelet coefficients in the upper half of the signal spectrum are classified. Wavelet packets divide the signal spectrum into frequency bands that are evenly spaced and have equal bandwidth and will be explored for use in identifying transient and quasi-steady-state speech.

MATLAB software used to implement the wavelet packet based algorithm uses a ‘natural order’ index, which does not correspond to increasing frequency, to label nodes.

A wavelet packet tree for a decomposition depth of 4 generated using the ‘natural order’ index labeling of MATLAB was presented in Figure 2.7. For ease of reference, the terminal nodes in subsequent figures are rearranged to show increasing frequency from left-to-right. Tables 3.1, 3.2, and 3.3 show the frequency ordered nodes that correspond to natural order for decomposition levels of 0 to 4, 5 and 6 respectively.

Table 3.1: Frequency ordered terminal nodes for depths 0 to 4.

<table>
<tr><td rowspan="5">Decomposition depth</td><td>0</td><td colspan="16"><b>0</b>
$f0$</td></tr>
<tr><td>1</td><td colspan="8"><b>1</b>
f_1</td><td colspan="8"><b>2</b>
f_2</td></tr>
<tr><td>2</td><td colspan="4"><b>3</b>
f_3</td><td colspan="4"><b>4</b>
f_4</td><td colspan="4"><b>6</b>
f_5</td><td colspan="4"><b>5</b>
f_6</td></tr>
<tr><td>3</td><td colspan="2"><b>7</b>
f_7</td><td colspan="2"><b>8</b>
f_8</td><td colspan="2"><b>10</b>
f_9</td><td colspan="2"><b>9</b>
f_{10}</td><td colspan="2"><b>13</b>
f_{11}</td><td colspan="2"><b>14</b>
f_{12}</td><td colspan="2"><b>12</b>
f_{13}</td><td colspan="2"><b>11</b>
f_{14}</td></tr>
<tr><td>4</td><td><b>15</b>
f_{15}</td><td><b>16</b>
f_{16}</td><td><b>18</b>
f_{17}</td><td><b>17</b>
f_{18}</td><td><b>21</b>
f_{19}</td><td><b>22</b>
f_{20}</td><td><b>20</b>
f_{21}</td><td><b>19</b>
f_{22}</td><td><b>27</b>
f_{23}</td><td><b>28</b>
f_{24}</td><td><b>30</b>
f_{25}</td><td><b>29</b>
f_{26}</td><td><b>25</b>
f_{27}</td><td><b>26</b>
f_{28}</td><td><b>24</b>
$f_{2}9$</td><td><b>23</b>
f_{30}</td></tr>
<tr><td colspan="2"></td><td colspan="16">Frequency →</td></tr>
</table>

Table 3.2: Frequency ordered terminal nodes for level 3 and 5.

Decomposition depth	3	Lower half	**7** f_7				**8** f_8				**10** f_9				**9** f_{10}			
	5		**31** f_{31}	**32** f_{32}	**34** f_{33}	**33** f_{34}	**37** f_{35}	**38** f_{36}	**36** f_{37}	**35** f_{38}	**43** f_{39}	**44** f_{40}	**46** f_{41}	**45** f_{42}	**41** f_{43}	**42** f_{44}	**40** f_{45}	**39** f_{46}
	3	Upper half	**13** f_{11}				**14** f_{12}				**12** f_{13}				**11** f_{14}			
	5		**55** f_{47}	**56** f_{48}	**58** f_{49}	**57** f_{50}	**61** f_{51}	**62** f_{52}	**60** f_{53}	**59** f_{54}	**51** f_{55}	**52** f_{56}	**54** f_{57}	**53** f_{58}	**49** f_{59}	**50** f_{60}	**48** f_{61}	**47** f_{62}
			Frequency →															

Table 3.3: Frequency ordered terminal nodes for level 3 and 6.

Decomposition Level	3	1st quarter	**7** f_7								**8** f_8							
	6		**31** f_{63}	**32** f_{64}	**34** f_{65}	**33** f_{66}	**37** f_{67}	**38** f_{68}	**36** f_{69}	**35** f_{70}	**43** f_{71}	**44** f_{72}	**46** f_{73}	**45** f_{74}	**41** f_{75}	**42** f_{76}	**40** f_{77}	**39** f_{78}
	3	2nd quarter	**10** f_9								**9** f_{10}							
	6		**55** f_{79}	**56** f_{80}	**58** f_{81}	**57** f_{82}	**61** f_{83}	**62** f_{84}	**60** f_{85}	**59** f_{86}	**51** f_{87}	**52** f_{88}	**54** f_{89}	**53** f_{90}	**49** f_{91}	**50** f_{92}	**48** f_{93}	**47** f_{94}
	3	3rd quarter	**13** f_{11}								**14** f_{12}							
	6		**31** f_{95}	**32** f_{96}	**34** f_{97}	**33** f_{98}	**37** f_{99}	**38** f_{100}	**36** f_{101}	**35** f_{102}	**43** f_{103}	**44** f_{104}	**46** f_{105}	**45** f_{106}	**41** f_{107}	**42** f_{108}	**40** f_{109}	**39** f_{110}
	3	4th quarter	**12** f_{13}								**11** f_{14}							
	6		**31** f_{111}	**32** f_{112}	**34** f_{113}	**33** f_{114}	**37** f_{115}	**38** f_{116}	**36** f_{117}	**35** f_{118}	**43** f_{119}	**44** f_{120}	**46** f_{121}	**45** f_{122}	**41** f_{123}	**42** f_{124}	**40** f_{125}	**39** f_{126}
			Frequency →															

The distribution by terminal nodes of the signal energy for decomposed speech depends on the specific word, the preprocessing applied to the speech signal, and the gender of the speaker. Figure 3.13 shows examples of this energy distribution for the word ‘nice’ as spoken by a male and a female speaker. A db20 wavelet function was used for the depth 3 decomposition. This energy distribution by wavelet node will be referred to as the energy profile of the word. The energy profiles obtained using the DWT and SWT presents information similar to the energy profile obtained using wavelet packets in that both give information on how the energy of a speech signal is distributed into frequency intervals.

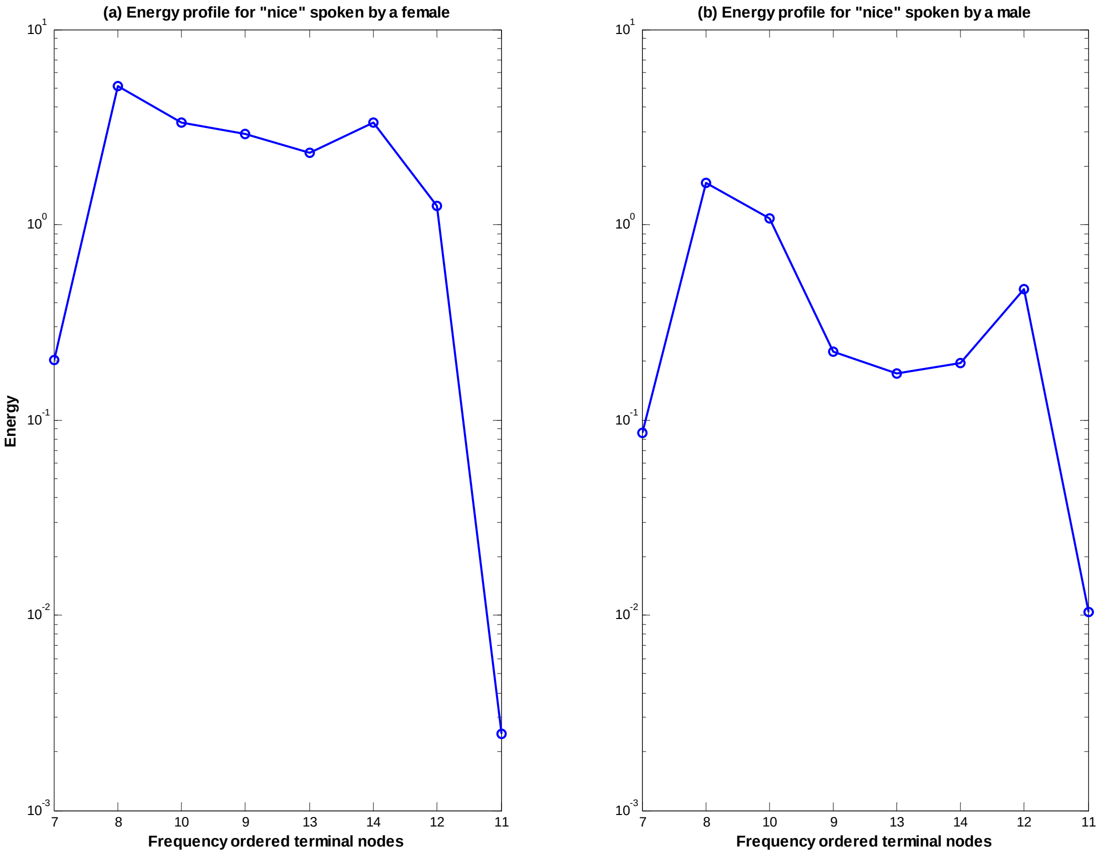

Figure 3.13: Energy distribution by node for the word ‘nice’ as spoken by a female and a male.

As in the DWT and SWT case, energy profiles are used to classify terminal nodes of the highpass filtered speech as having either more transient information or more quasi-steady-state information. Nodes with mostly transient information will be referred to as *transient nodes*, and nodes with mostly quasi-steady-state information will be referred to as *quasi-steady-state nodes*.

An example of the node classification is shown in Figure 3.14. In this figure, energy profiles for the highpass filtered, tonal and nontonal speech components are shown for the word ‘pike’ spoken by a female. A db20 wavelet function was used with decomposition level of 4. Even though a lower level was used, WP divided the frequency spectrum into 16 bands whereas the DWT and SWT divided the spectrum into only 6 bands. It can be observed that at node 18, the energy of the highpass filtered speech node is very close to that of the corresponding node of the tonal speech component, hence node 18 is considered to have predominately quasi-steady-state information. At node 27, the energy of the highpass filtered speech is closer to that of the nontonal speech component than the tonal component. As a result, this node is classified as a transient node. For this word, transient nodes are nodes {15, 21, 22, 20, 19, 27, 28, 30, 29, 25, 26 and 24}, and quasi-steady-state nodes are nodes {16, 18, 17 and 23}.

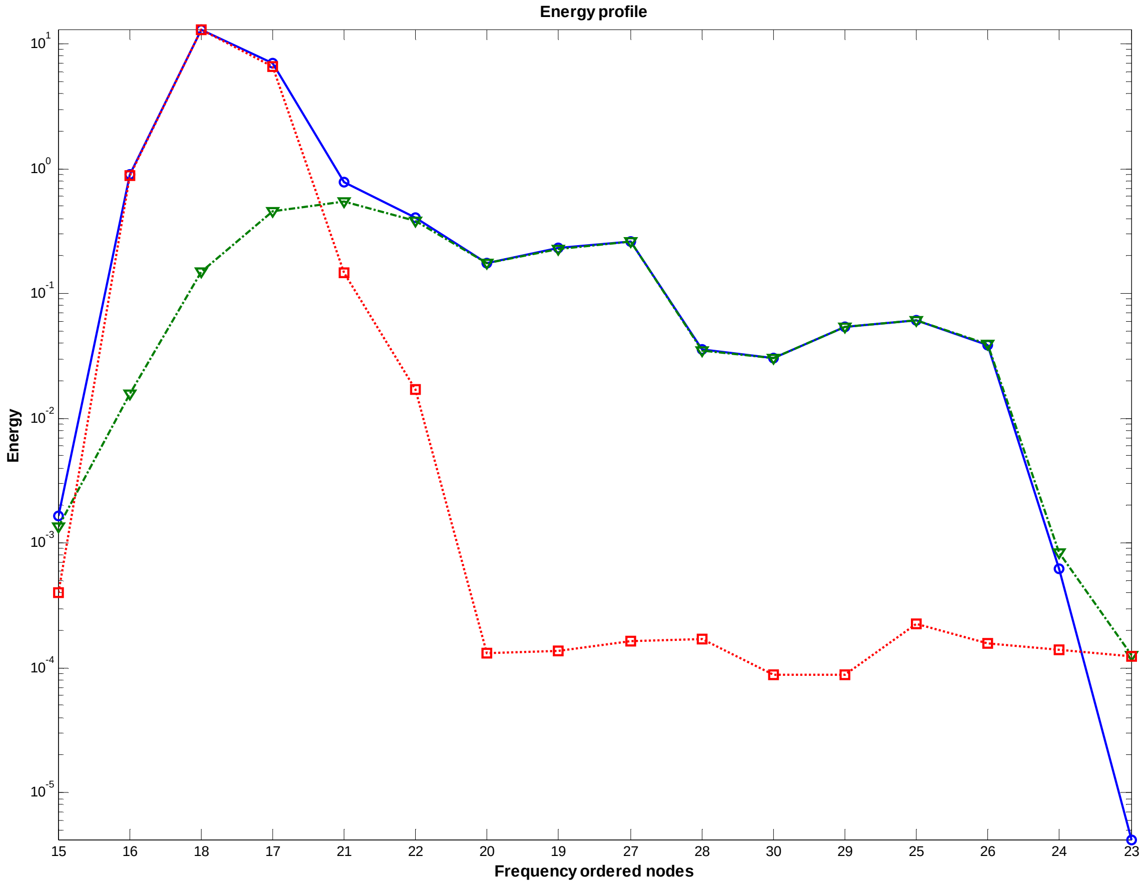

Figure 3.14: Node classification for the word ‘pike’ spoken by a female.

The inverse wavelet packet transform (IWPT), which was discussed in Chapter 2, was used to synthesize transient and quasi-steady-state speech components from the wavelet packet representation. To synthesize the transient speech component, wavelet coefficients of transient nodes were used, with wavelet coefficients of quasi-steady-state nodes set to zero. To synthesize the quasi-steady-state speech component, wavelet coefficients of quasi-steady-state node were used, with wavelet coefficients of transient nodes set to zero.

To evaluate how closely the estimates of the transient and quasi-steady-state speech components synthesized using wavelet packets approximated Yoo's nontonal and tonal components, the former were compared, in the time- and frequency-domain, to the latter. A listening test was also used to compare the wavelet derived speech components to the speech components obtained using the algorithm of Yoo. As before, the listening test was conducted by the author listening to the speech components and then making a judgment of how similar they were.

3.4 RESULTS FOR WAVELET PACKETS

In this subsection, example results using wavelet packets to identify transient and quasi-steady-state speech are presented through an example that illustrates the node classification and the speech component synthesis processes for the word 'pike' spoken by a male. A db20 wavelet function was used for the depth 4 decomposition. Figure 3.15 shows the energy profiles for the highpass filtered, tonal and nontonal speech components. Using these energy profiles, transient nodes were identified as node {15, 20, 19, 27, 28, 30, 29, 25, 26 and 23}, and quasi-steady-state node were identified as nodes {16, 18, 17, 21, 22 and 24}.

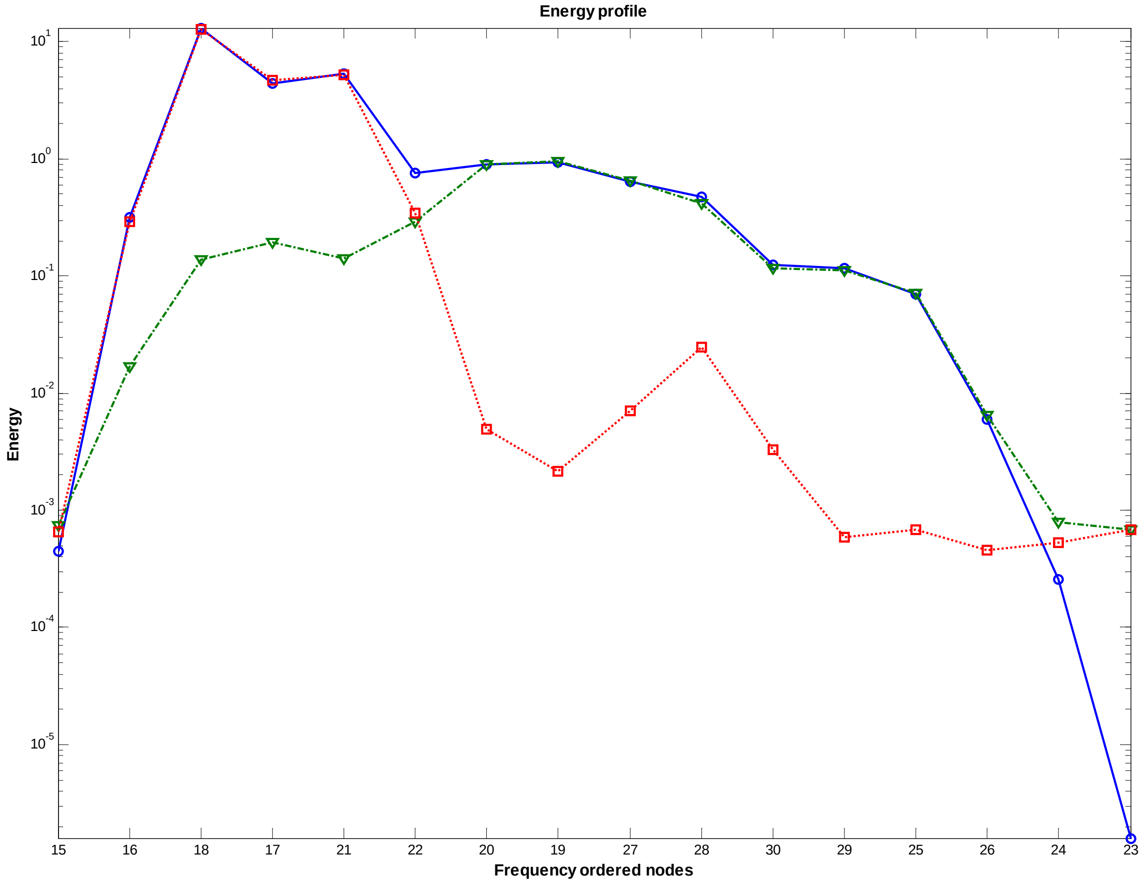

Figure 3.15: Energy profiles for the highpass filtered, tonal and nontonal components of the word ‘pike’ spoken by a male.

As an example of the synthesis process, Figure 3.16 compares the transient and quasi-steady-state components synthesized using wavelet packets to the nontonal and tonal components, respectively. Figure 3.17 compares the spectra of these speech components. The spectrum of the quasi-steady-state component synthesized using wavelet packets, like the spectrum of Yoo’s tonal component, has its energy concentrated in the frequency ranges of (700, 1800) Hz and (2600, 4000) Hz. Although the transient component has a narrower spectrum than the nontonal component, the shallow peaks observed in the transient component (around 2190, 2600, 3220, 3740 and 4140 Hz) match those observed in the nontonal component. In a listening test, the quasi-steady-state speech component was a close estimate of the tonal speech component although slightly more intelligible. The transient speech component was also a close estimate of the nontonal speech component, although slightly more whispered.

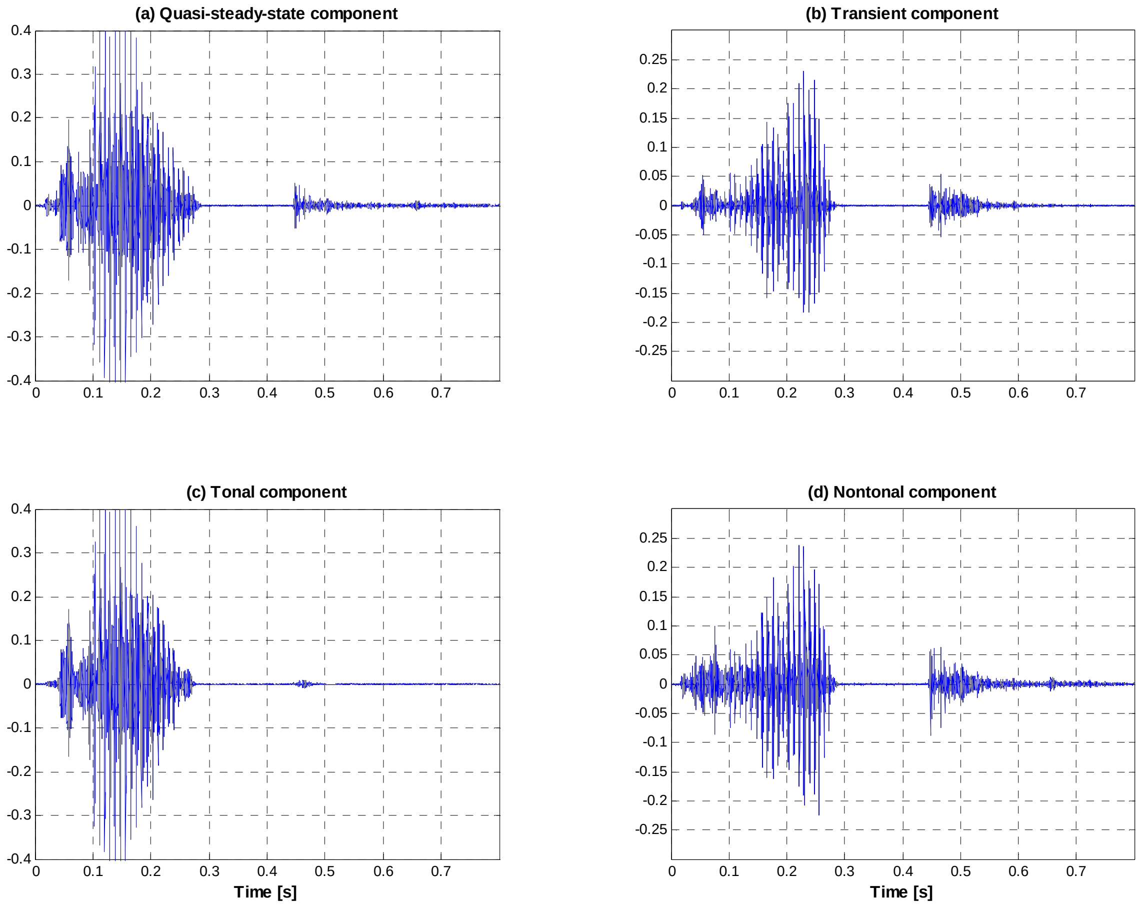

Figure 3.16: Wavelet packet synthesized speech components, and the tonal and nontonal speech components of the word 'pike' spoken by a male.

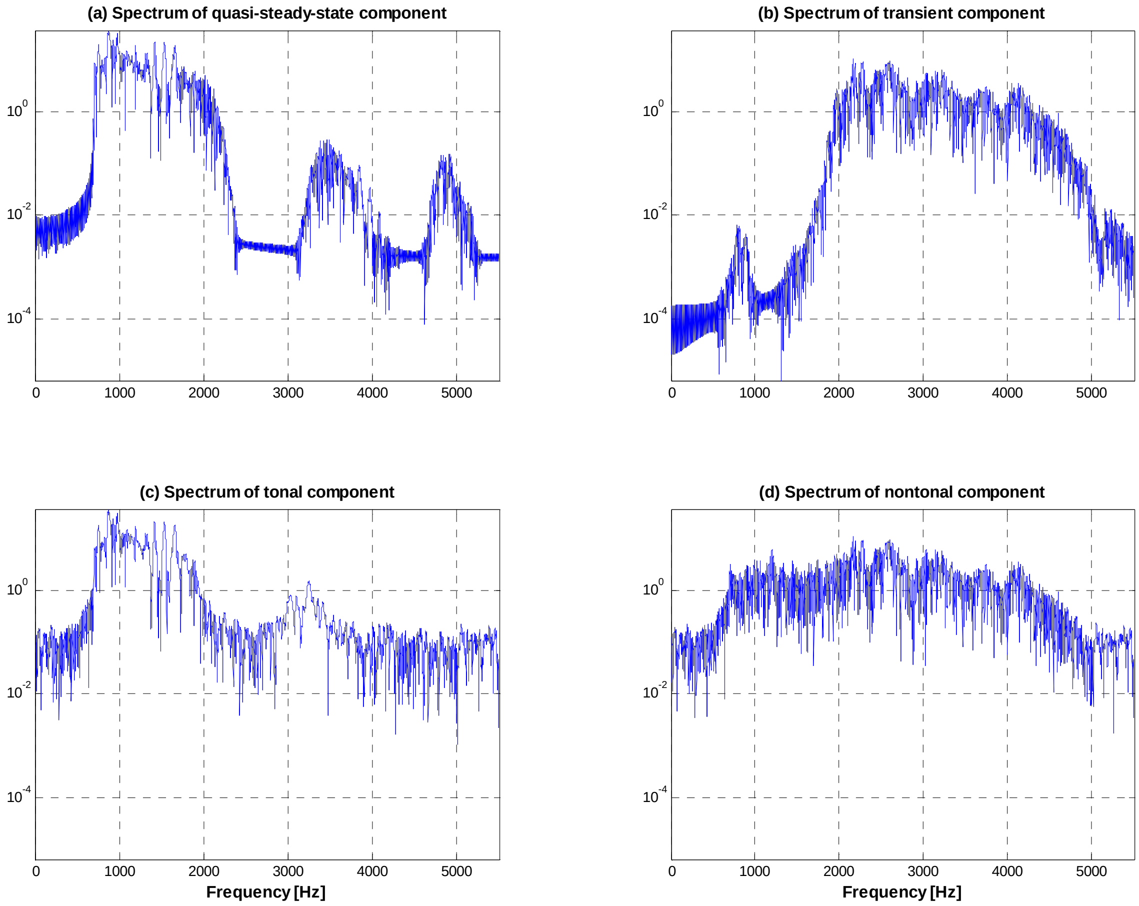

Figure 3.17: Spectra of wavelet packet estimated speech components, and of the tonal and nontonal speech components of the word 'pike' spoken by a male.

Table A3 in the appendix shows the node classification obtained for 18 words using a level 4 wavelet packet decomposition. For most words, one of nodes {16, 18 and 17}, which includes the signal spectrum from 700 Hz to 1800 Hz, was identified as a quasi-steady-state node. Most nodes from the set of nodes {20, 19, 27, 28, 30 and 29} were considered as transient nodes. Nodes {15 and 16} had zero energy because of the highpass filtering which was performed, and nodes {25, 26, 24 and 23} had insignificant amounts of energy. Nodes {21 and 22} were mixed.

Transient components synthesized using the wavelet packets were slightly more whispered than the nontonal speech components, and quasi-steady-state components were slightly more intelligible than the tonal components.

The estimation errors for 18 words, as given by the MSE between the spectra of the transient component synthesized using wavelet packets and the nontonal component, are given in Table 3.4. Table 3.4 also includes estimation errors incurred when the speech components were synthesized using the DWT and SWT. The subscript m and f denotes whether the word was spoken by a male or a female. The estimation errors incurred when wavelet packets were used to synthesize the transient components, as compared to when the DWT and SWT were used, are substantially smaller.

In general, the speech components synthesized using wavelet packets were better estimates of the tonal and nontonal speech components than the speech components synthesized using the DWT and SWT. This is evident from the spectral comparisons, the MSE measurement summarized in Table 3.4, and the listening tests.

Table 3.4: Estimation errors for transient speech components for 18 words synthesized using wavelet packets (2nd column), the SWT (3rd column) and DWT (right column).

Word	**MSE for WP**	**MSE for SWT**	**MSE for DWT**
pike$_m$	0.6808	2.3479	2.3011
pike$_f$	0.4755	0.4191	0.4768
calm$_m$	0.0388	0.0199	0.0335
calm$_f$	2.9531	3.0055	2.9932
nice$_m$	3.0623	6.4804	6.4760
nice$_f$	0.2678	0.6444	0.6169
keg$_m$	6.8761	19.7000	19.6998
keg$_f$	7.9848	9.0271	9.1006
fail$_m$	19.8731	18.1025	19.3032
fail$_f$	0.0945	0.2257	0.2481
dead$_m$	1.2299	3.1481	3.1488
chief$_f$	15.0378	27.7720	28.9644
live$_m$	0.9396	2.3942	2.2602
merge$_f$	10.2598	48.4013	30.1761
juice$_f$	2.3053	3.8451	3.8850
armchair$_f$	21.0267	24.3349	27.2667
headlight$_m$	3.3756	4.8069	4.7091
headlight$_f$	0.0814	0.07680	0.0814
Mean	**5.364606**	**9.708433**	**8.985606**

Comparing the level 6 DWT and SWT decompositions to the depth 4 wavelet packet decomposition, the wavelet packet decomposition is able to divide the level 1 signal spectrum into 8 frequency bands, and the level 2 signal spectrum into 4 frequency bands. This division of the signal spectrum allows for a more efficient classification of which frequency bands (as given by node) have more transient or more quasi-steady-state information.

For example, for the word ‘nice’ spoken by a male, in the upper half of the signal spectrum, wavelet packets analysis associated nodes {28, 30, 29, 25, 26 and 23} with the quasi-steady-state component and node {27 and 24} with the transient component. On the other hand, because of the inability of the DWT and SWT to divide the upper half of the signal spectrum, the entire upper half of the spectrum was associated with the transient component.

The spectra of the transient and quasi-steady-state components of the word ‘nice’ spoken by a male synthesized using the 3 wavelet transforms are compared to the spectra of the tonal and nontonal components in Figure 3.18. Despite the regions of low energy that are present in the spectra of the wavelet packet synthesized speech components but absent in the spectra of the tonal and nontonal components, the speech components synthesized using wavelet packets, as compared to those synthesized using the DWT and SWT, provide much better estimates of the tonal and nontonal speech components.

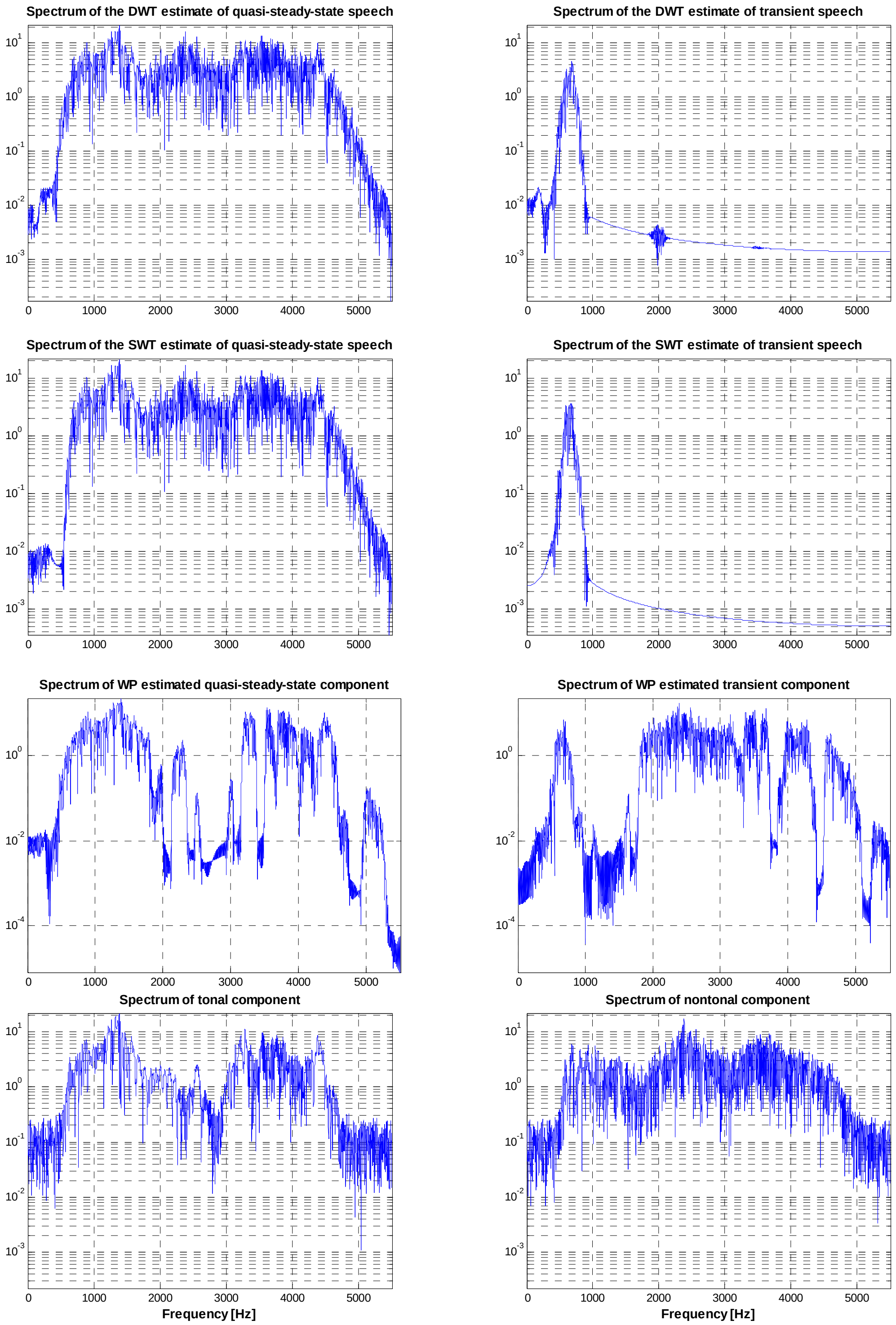

Figure 3.18: Spectra of speech components for the word ‘nice’ spoken by a male synthesized using the DWT (1st row), SWT (2nd row), WP (3rd row) and Yoo’s algorithm.

4.0 A WAVELET PACKETS BASED ALGORITHM FOR IDENTIFYING TRANSIENT SPEECH

The methods used in Chapter 3 to identify transient and quasi-steady-state speech use energy, a global measure, and wavelets to identify these speech components. Those approaches required a given node to be classified as either quasi-steady-state or transient for the entire duration of the speech signal. Integrating variable frame rate processing into the method may provide a mechanism to associate coefficients of a given node with either the quasi-steady-state or transient component at different times depending on whether the speech is relatively stationary or transitive.

This chapter describes an algorithm to identify transient and quasi-steady-state speech components. It combines the variable frame rate process with wavelet packets analysis. The processes of choosing a wavelet function to use for the decomposition, choosing a decomposition level, classifying terminal nodes of a decomposed speech signal, incorporation of the VFR process into the wavelet analysis, and synthesis of transient and quasi-steady-state speech are described. The design and selection criteria are described with the algorithm. Results of studies to evaluate the different criteria are presented in results.

4.1 METHOD

The wavelet-packet based algorithm to identify transient and quasi-steady-state speech components involves 4 steps:

1) **Wavelet Packet decomposition of speech:** The speech signal is decomposed using a wavelet function and a decomposition level that were selected in the development of the algorithm.
2) **Classification of terminal nodes:** Energy profiles are used to classify terminal nodes of the decomposed highpass filtered speech signal as having predominately transient information; predominately quasi-steady-state information; both type of information (ambiguous).
3) **Incorporation of variable frame rate processing and synthesis of speech components:** Variable frame rate is applied to ambiguous nodes to identify time segments that are predominately transient or predominately quasi-steady-state, and ambiguous nodes during these time segments are associated with transient or quasi-steady-state components accordingly.
4) **Synthesis of Speech Components:** Transient and quasi-steady-state speech components are synthesized.

4.1.1 Wavelet Packet decomposition of speech

The goal in using wavelet packets was to obtain a division of the frequency spectrum with frequency bands that are equal in bandwidth, have equal peak amplitudes, no side-lobes and smooth frequency responses. Figure 4.1 shows filter frequency responses that divide the frequency spectrum into bands with these properties. The purpose of the first step of the algorithm was to identify a wavelet function that provides a division of the frequency spectrum that is as close as possible to this goal.

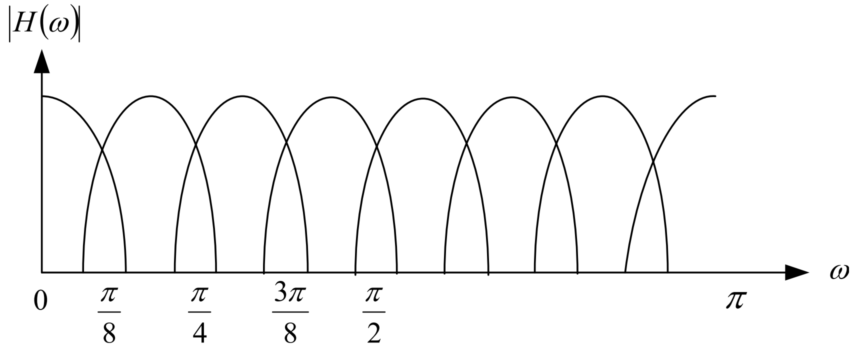

Figure 4.1: Evenly spaced equal bandwidth frequency splitting.

For actual wavelet functions, the filter frequency responses have unequal peak amplitudes, bandwidths, and side-lopes. Figure 4.2 (a) shows the filter frequency responses for a db4 wavelet.

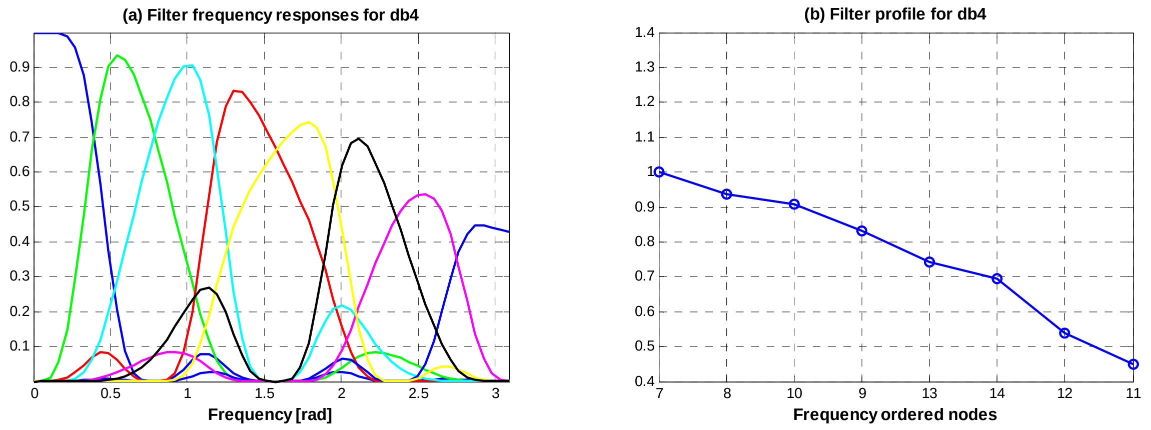

Figure 4.2: (a) Filter frequency responses and (b) filter profile for a db4 wavelet function. The frequency responses have side lobes, unequal bandwidth and peak amplitudes.

If the peaks filter frequency responses shown in Figure 4.2 are connected, a function that will be referred to as the *filter profile* is obtained. The filter profile may be interpreted as a function that shows the uniformity of the filter amplitudes. The filter profile for the db4 wavelet function, which has a downward slope, is shown if Figure 4.2 (b). If a wavelet function having the properties shown in Figure 4.1 is used to decompose a linear swept-frequency signal (chirp) with instantaneous frequencies of 0 Hz and half the sampling rate occurring between t = 0 and t = t_{max}, then barring end effects, each frequency band would have the same energy. If a db4 wavelet function instead, which has a downward sloping filter profile, is used, then low frequencies are emphasized.

The size of side-lobes in the filter frequency responses was also a consideration. The wavelet function to be used for the decomposition should have narrow bands, small side-lobes and a flat filter profile.

As an example, Figure 4.3 shows the filter frequency responses and filter profiles for db12 and db20 wavelet functions for a decomposition of depth 3. $\psi_i(\omega)$ denotes the filter frequency response for terminal node i, with the nodes labeled using the natural order. The profile for the db12 wavelet function has an upward slope, while the profile for the db20 wavelet function is flatter. A wavelet function with a filter profile similar to that of db20 would be preferred over one with a filter profile similar to that of db4 or db12, since this profile is a good estimate of the desired profile.

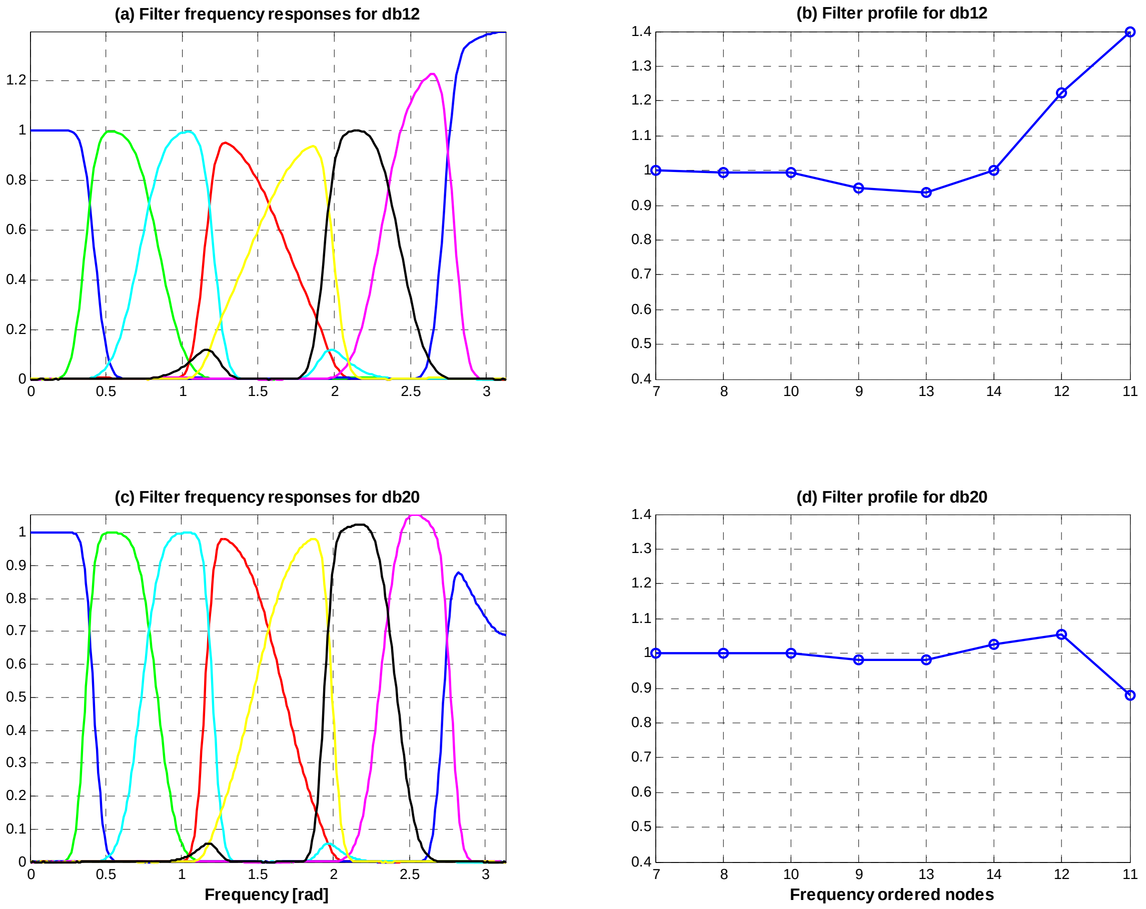

Figure 4.3: Filter frequency responses and filter profiles for db12 (top) and db20 (bottom) wavelet functions.

Daubechies and Symlets wavelets were considered, but since results observed using the two wavelets families were very similar, only Daubechies wavelets were evaluated in detail to identify the wavelet function that approximated the desired properties most closely.

4.1.2 Classification of Terminal Nodes

Energy profiles for the highpass filtered speech, and the tonal and nontonal speech components were used to classify terminal nodes of the highpass filtered speech as having either more transient information or more quasi-steady-state information, as described below. Nodes with mostly transient information will be referred to as *transient nodes*, and nodes with mostly quasi-steady-state information will be referred to as *quasi-steady-state nodes*. There are instances where a given terminal node is not predominately either type. These nodes will be refereed to as *ambiguous nodes.* Specific procedures to identify these nodes are explained below.

To classify terminal nodes of a highpass filtered speech signal, the energy profile of the highpass filtered speech is compared node-by-node to the energy profiles of the tonal and nontonal speech components. A terminal node from the highpass filtered speech is classified as transient if its energy is close to the energy in the corresponding node of the nontonal component and greater that a threshold difference δ from the energy in the corresponding node of the tonal component. A terminal node of the highpass filtered speech is classified as quasi-steady-state if its energy is close to the energy of the corresponding node of the tonal component and δ greater than the energy of the corresponding node of the nontonal component. If the energy of a terminal node of the highpass filtered speech is within δ dB of the energies of both Yoo's tonal and nontonal components, that node is considered to have mixed information and is identified as an ambiguous node. The threshold, δ, is referred to as the *ambiguity threshold.* The node grouping formula can be summarized as follows;

For a given terminal node with label f_i,

$$if \quad \left|E_{hp}(f_i)-E_{nt}(f_i)\right| < \left|E_{hp}(f_i)-E_t(f_i)\right| \quad AND \quad \left|E_{hp}(f_i)-E_t(f_i)\right| > \delta \qquad node = transient$$

$$if \quad \left|E_{hp}(f_i)-E_t(f_i)\right| < \left|E_{hp}(f_i)-E_{nt}(f_i)\right| \quad AND \quad \left|E_{hp}(f_i)-E_{nt}(f_i)\right| > \delta \quad node = steadystate$$

$$else \qquad\qquad node = ambiguous$$

$E_{hp}(f_i)$, $E_{nt}(f_i)$ and $E_t(f_i)$ are the energies of the highpass filtered, nontonal, and tonal speech, respectively, for the node labeled f_i. A threshold value of δ = 0, results in no ambiguous nodes, while a threshold value of δ = ∞ results in all nodes being classified as ambiguous. This method of node classification is similar to the method used in Chapter 3, with the addition of ambiguous nodes.

The effect of decomposition level on the energy in ambiguous nodes was investigated. We assume that it would be desirable to have as little energy in ambiguous nodes as possible. To reduce the proportion of energy in ambiguous nodes, the decomposition level was increased from the initial decomposition level of 3 to 4, on the basis that the children of the nodes classified as ambiguous nodes at level 3 might not be classified as ambiguous nodes at level 4. If ambiguous nodes still existed at level 4, the decomposition was increased to 5, and then 6 if ambiguities still existed at level 5. The different decomposition levels are compared with respect to the energy in ambiguous nodes to determine the best level to use.

An example of the node classification using δ = 7 dB, is illustrated in Figure 4.4. Energy profiles for the highpass filtered, tonal and nontonal speech components are shown for the word ‘pike’ spoken by a female. A db20 wavelet function was used with decomposition level of 4. Consider node 17. The energy in node 17 of the highpass filtered speech is very close to that of node 17 of the tonal speech, while the energy of node 17 of the nontonal component is more than δ = 7 dB smaller. Therefore this node is classified as a quasi-steady-state node. Node 27 is classified as a transient node because the energy of node 27 of the highpass filtered speech is closer to the energy of node 27 of the nontonal speech than node 27 of the tonal speech. At node 21, the energies of both the tonal and nontonal nodes are within 7 dB of the highpass filtered speech. This node is classified as ambiguous.

The overall node classification is shown in the bar beneath the energy profiles plot. Transient nodes are nodes {22, 20, 19, 27, 28, 30, 29, 25 and 26}, quasi-steady-state nodes are nodes {16, 18 and 17}, and ambiguous nodes are nodes {15, 21, 24 and 23}. In this bar, transient, quasi-steady-state, and ambiguous nodes are indicated by;

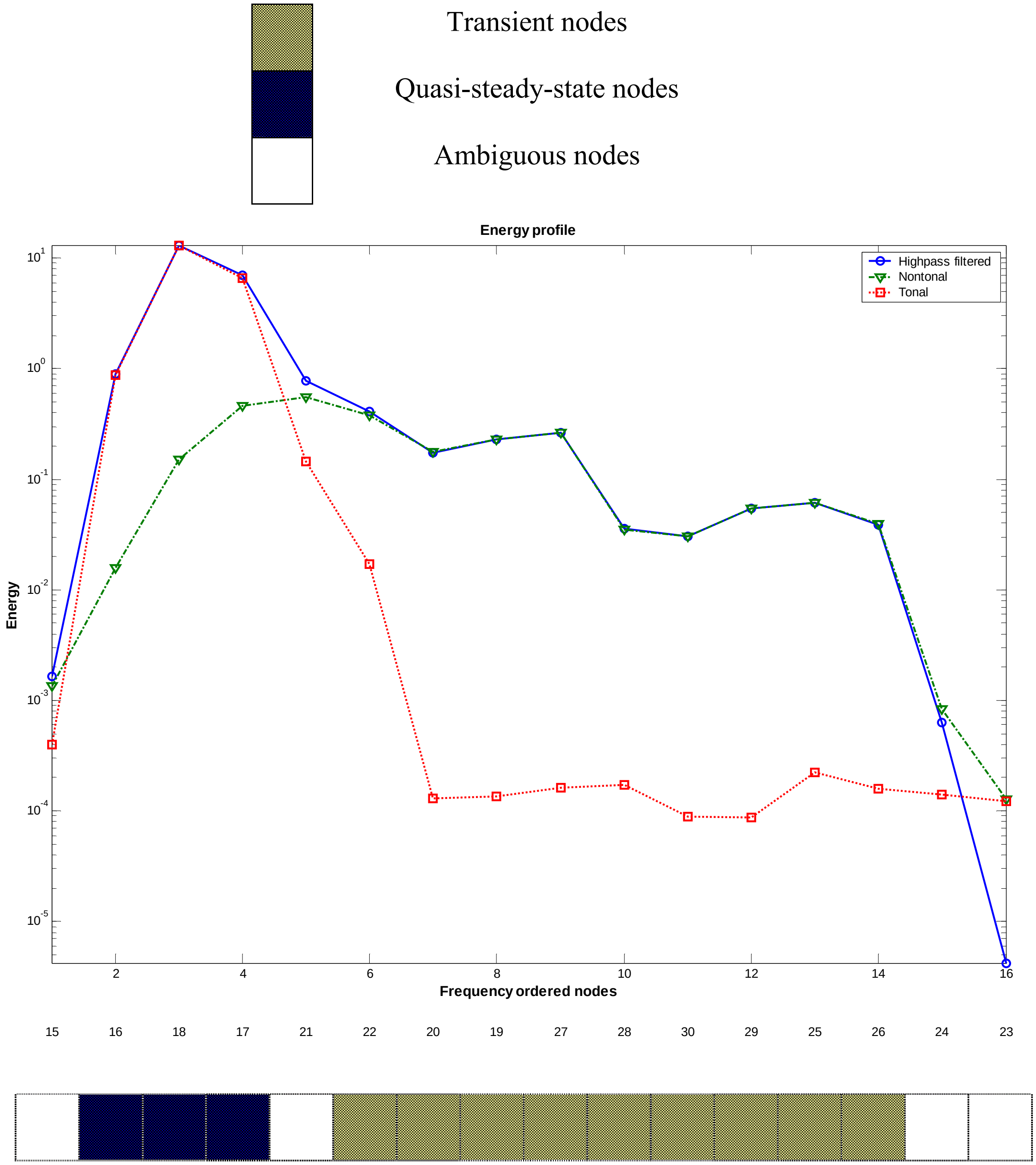

Figure 4.4: Example of node classification.

4.1.3 Incorporation of Variable Frame Rate Processing

We propose that ambiguous nodes include both transient and quasi-steady-state information that could not be isolated using frequency domain processing by wavelet packets alone. Variable frame rate processing, was investigated as a method to separate transient information from quasi-steady-state information in these nodes. Wavelet coefficients of the ambiguous nodes were included in the synthesis of the transient or quasi-steady-state speech component based on the VFR analysis.

Variable frame rate processing can identify time segments of speech where speech feature vectors are changing rapidly and time segments where speech feature vectors are relatively stationary. The approach to classification used here assumes that the time segments with rapidly changing feature vectors are associated with transient speech, while the time segments with slowly changing feature vectors are associated with quasi-steady-state speech.

The feature vector used for the variable frame rate algorithm is the Mel-frequency cepstral coefficients (MFCC). The flow chart of Figure 2.14, as discussed in Chapter 2, shows the process by which MFCC feature vectors are created from speech. In this section, the setup of this process and the values of the parameter used will be described. The variable frame rate (VFR) algorithm of Le Cerf and Van Compernolle [23] [24], which was used in this study, will be revisited with particular attention to parameter settings.

The speech signal is framed using a Hamming window of length 25 ms with frame step size of 2.5 ms. Twelve Mel-frequency cepstral coefficients per frame are calculated. The log energy and the first derivative cepstra are included, bringing the total number of coefficients per frame to twenty-six. Twenty-seven filters are used in the filter banks at the mel-scaling and smoothing stage as described in Chapter 2. These filters are adjusted to cover the spectrum from 0 Hz to half the sampling rate (5512.5 Hz).

The Euclidean norm of the first derivative cepstra is computed. This norm is large when the MFCC of two successive frames are different and small when the MFCC are similar. It provides information about the transitiveness of a speech signal and will be referred to as the *transitivity function*. The transitivity function is quantized so that it has a value of 1 when it is greater that the threshold, and 0 otherwise. Transient speech is synthesized by multiplying the speech signal by the quantized transitivity function. The number of samples in the original transitivity function is equal to the number of frames of the original signal, and as a result, the transitivity function must be interpolated before multiplying so that it has as many samples as the speech signal itself. The interpolated and quantized transitivity function will be called the *quantized transitivity function* (QTF).

The quantized transitivity function is used to select coefficients of an ambiguous node to be included in the synthesis of transient and quasi-steady-state speech components, as illustrated in Figures 4.5, 4.6 and 4.7. These figure are a fictitious example that uses a level 2 wavelet packet tree to illustrate the synthesis method. f_0 is the original speech signal, f_1, and f_2 are the level 1 wavelet packet nodes, and f_3, f_4, f_5, and f_6 are the level 2 wavelet packet nodes. Node f_3 is a quasi-steady-state node, nodes f_5 and f_6 are transient nodes, and node f_4 is an ambiguous node. The wavelet coefficients of node f_4 are multiplied by the quantized transitivity function (QTF) to define a transient component and by (1-QTF) to define a quasi-steady-state component of these coefficients.

4.1.4 Synthesis of Speech Components

The fourth component of the algorithm involves synthesis of transient and quasi-steady-state speech components using the node grouping obtained as described above. These components were used as estimates of the tonal and nontonal components. The inverse wavelet packet transform (IWPT), which was discussed in Chapter 2, was used to synthesize the speech components from the wavelet packet representation. To synthesize the transient speech component, wavelet coefficients of quasi-steady-state nodes were set to zero, and to synthesize quasi-steady-state component, wavelet coefficients of transient nodes were set to zero. Ambiguous nodes were handled as described below. In the synthesis of the transient component, shown in Figure 4.6, the wavelet coefficients of node f_4 are replaced by their VFR estimate of the transient coefficients, and the wavelet coefficients of node f_3, which is a quasi-steady-state node, are replaced by zeros. The estimate of the transient component is synthesized from the nodes f_5, f_6 and the transient part of f_4.

In the synthesis of the quasi-steady-state component, shown in Figure 4.7, the wavelet coefficients of nodes f_5 and f_6, which are transient nodes, are replaced by zeros, and the wavelet coefficients of node f_4 are replaced by their VFR quasi-steady-state coefficients. The estimate of the quasi-steady-state component is synthesized from node f_3 and the quasi-steady-state component of f_4.

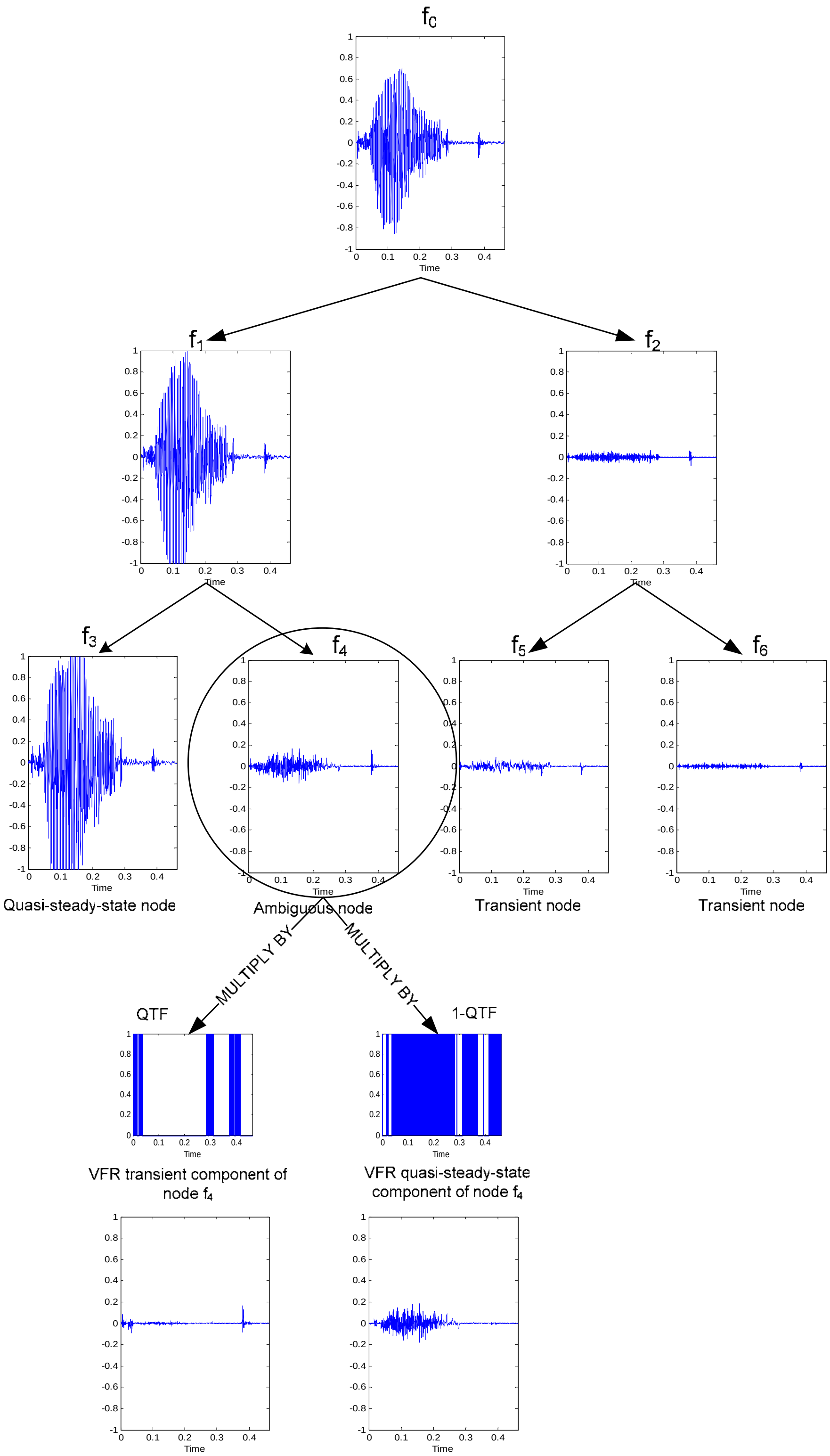

Figure 4.5: Wavelet packet decomposition and application of VFR.

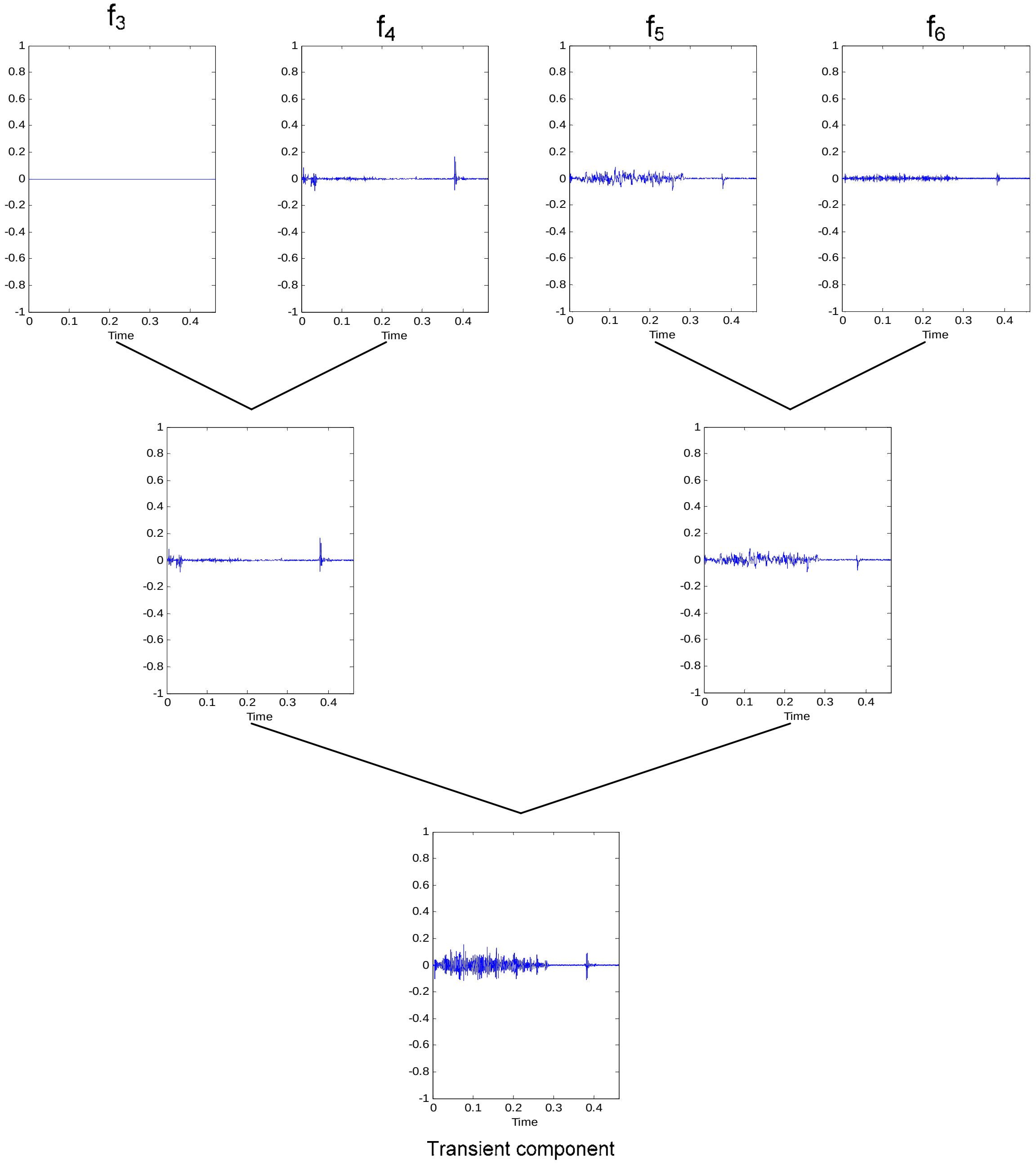

Figure 4.6: Synthesis of transient speech component

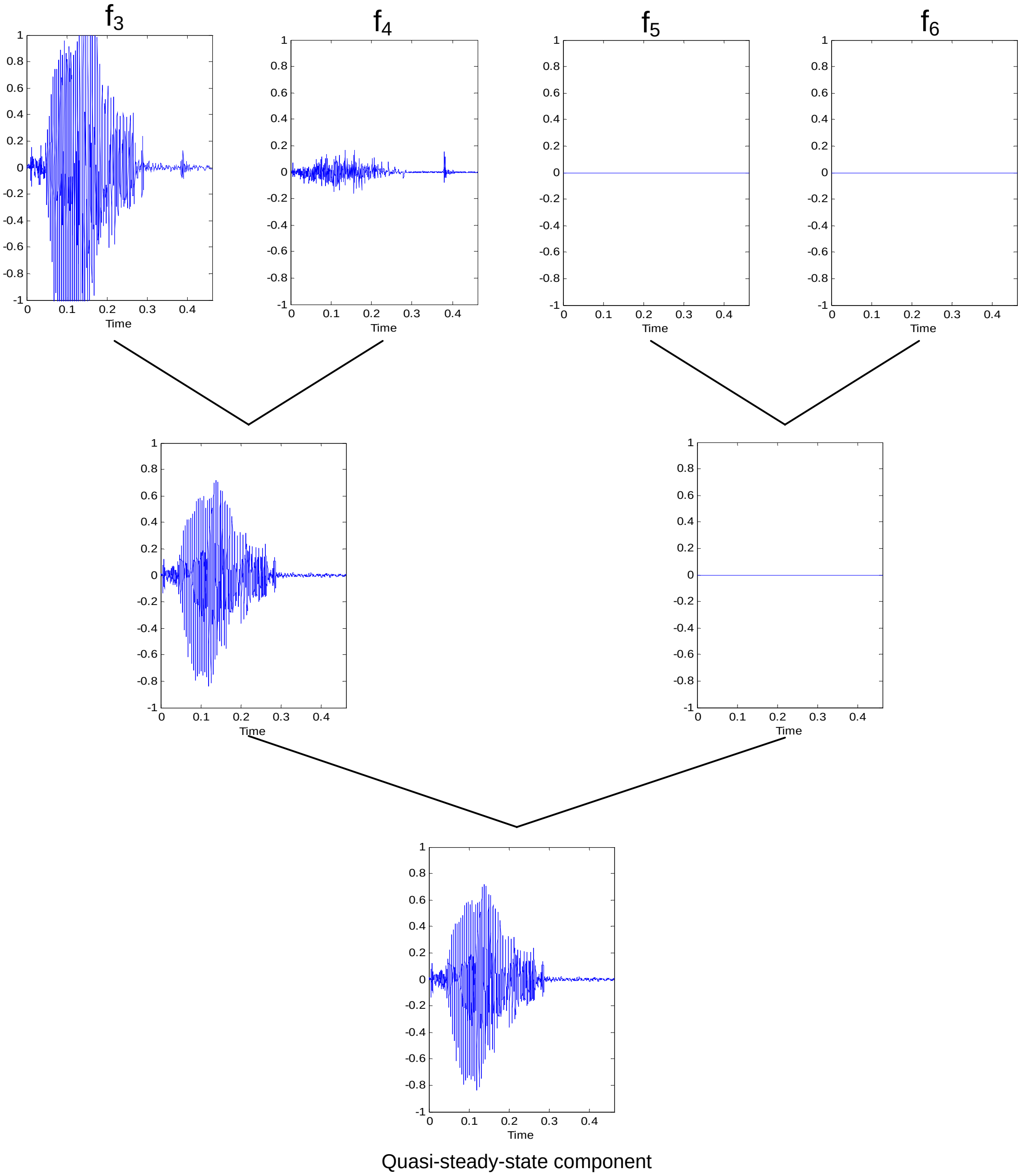

Figure 4.7: Synthesis of quasi-steady-state speech component

As a preliminary study to establish whether the variable frame rate algorithm detected transitions in speech, tests were carried out on a synthetic signal. The synthetic signal used will be referred to as the *tone-chirp-tone* signal, and is shown in Figure 4.8. This signal consists of a tone at a low frequency, a transition to a higher frequency and another tone at this higher frequency. The duration of both tones is 40 ms. The first tone has a 10 ms start period created by multiplying the tone by a window function, shown in Figure 4.9, having a 10 ms ramp. The ramp is created using a half period of a cosine function. The second tone has a 10 ms end duration formed in the corresponding way. Zero padding of 50 ms was inserted at the beginning and end of the tone-chirp-tone signal. The duration of the chirp (transition to the second tone) and the frequencies of the tones were varied to create four different test scenarios as given in Table 4.1.

In the tone-chirp-tone synthetic signal, the tones and the chirp are intended to model quasi-steady-state and transient speech, respectively. The transitivity function computed for the tone-chirp-tone signal had a minimum value of 0 and a maximum value of 16.7. To determine the threshold for each test situation of Table 4.1, the threshold was varied from 0 to 17, in increments of 0.5. This threshold will be referred to as the *transient-activity threshold*. A threshold value of 0 includes the entire tone-chirp-tone signal in the computation of the transient component, while a value of 17 includes the entire tone-chirp-tone signal in the computation of the quasi-steady-state component.

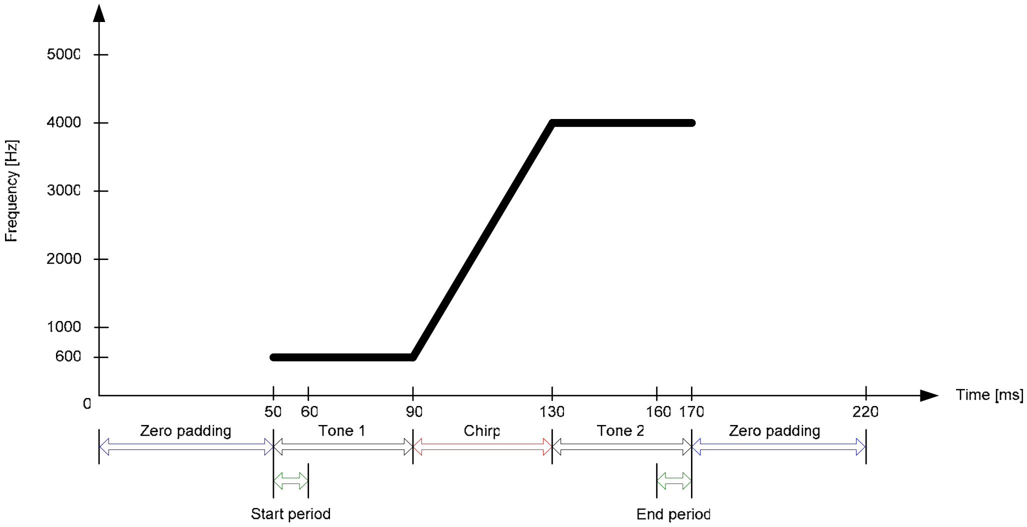

Figure 4.8: Spectrogram for the tone-chirp-tone signal with tones frequencies of 0.6 kHz and 4.0 kHz, and a tone duration of 40 ms.

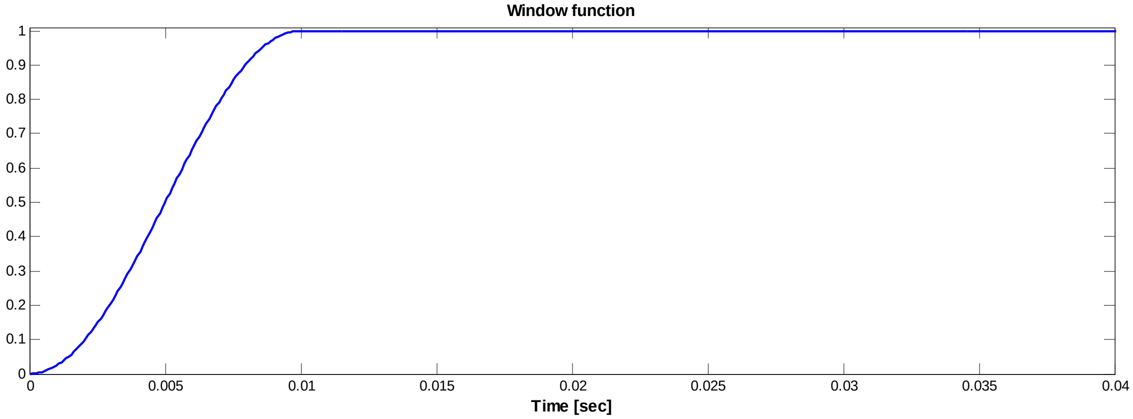

Figure 4.9: Window function used to create start and end periods of the tones.

Table 4.1: Test conditions evaluated for the tone-chirp-tone signal

Run	Tone 1 frequency	Tone 2 frequency	Chirp duration
1	0.6 kHz	1.9 kHz	40 ms
2	0.6 kHz	1.9 kHz	200 ms
3	0.6 kHz	4.0 kHz	40 ms
4	0.6 kHz	4.0 kHz	200 ms

Tests carried out on the tone-chirp-tone synthetic signal showed that the VFR algorithm was able to separate the chirp from the two tones. As an example, Figure 4.10 shows the tone-chirp-tone signal, its spectrogram, and spectrograms of the transient and quasi-steady-state components obtained as described above. The two tones had frequencies of 600 Hz and 4000 Hz, and the duration of the chirp was 40 ms. The spectrograms were computed using a Hanning window of lengths 10 ms and window overlap of 9 ms. Also show in the figure is the transitivity function, interpolated but not quantized. The transient-activity threshold was set to 7.

As seen in the figure, the transient component includes the chirp, and the quasi-steady-state component includes the two tones. The onset and offset of the tone-chirp-tone signal, as shown in part (f) of Figure 4.10, were also captured in the transient component. The transitivity function peaked during the chirp, and at the beginning and end of the tone-chirp-tone signal.

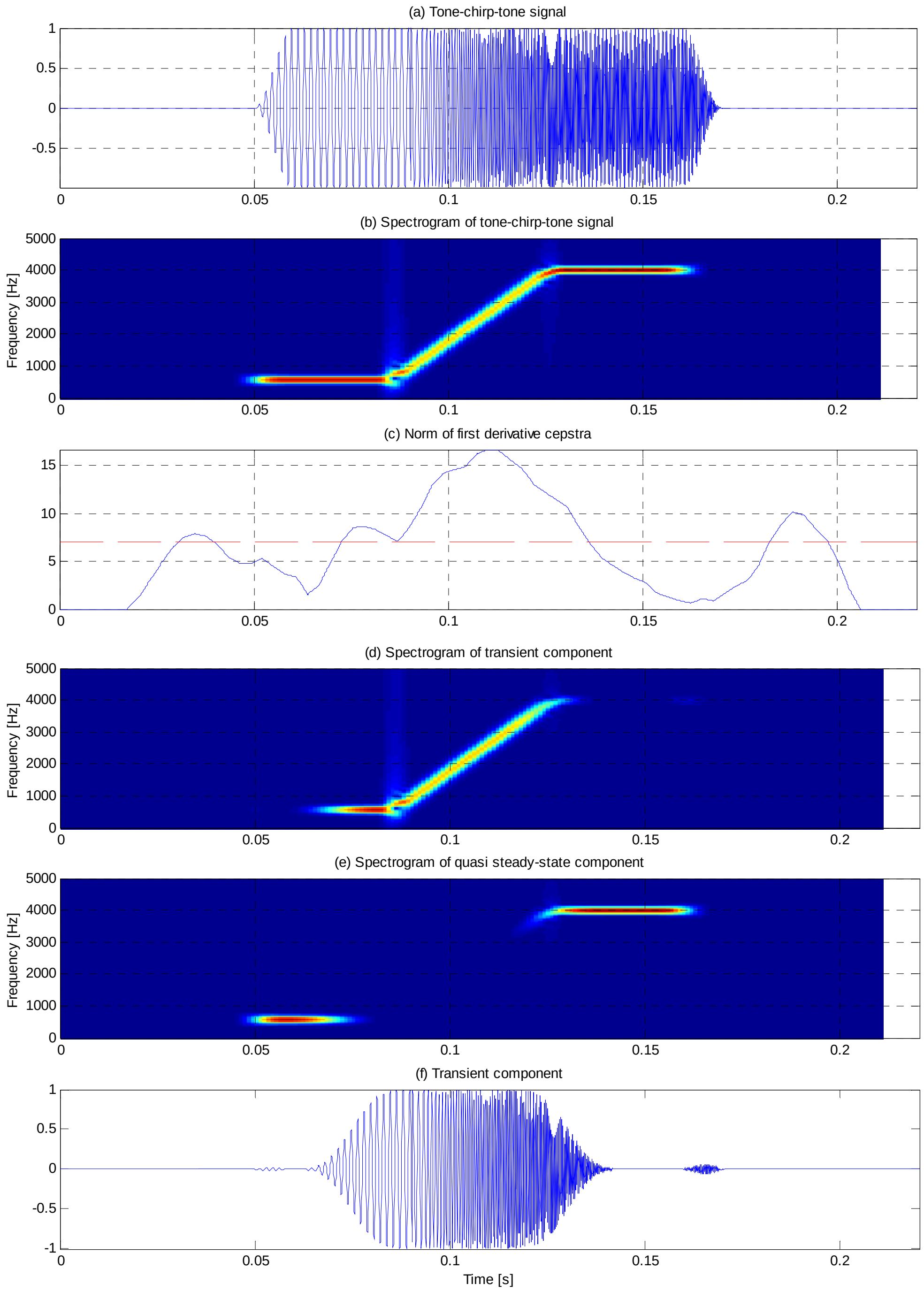

Figure 4.10: (a) Tone-chirp-tone signal, (b) spectrogram of tone-chirp-tone signal, (c) transitivity function and transient-activity threshold, (d) spectrogram of transient component (e) spectrogram of quasi steady-state component, and (f) transient component.

The variable frame rate technique was also tested with speech signals. As in the synthetic signal, the onsets and offsets of the speech signals were included in the transient component. Additionally, the transient component included the onset and offset of the strongest formant. The quasi-steady-state component, which included everything that was not captured in the transient component, had much more energy than the transient component. The proportion of the energy of the original signal captured by the quasi-steady-state component varied from word to word and also depended on the transient-activity threshold. A higher threshold included more of the original signal in the quasi-steady-state component, increasing the proportion of energy of this component.

Figure 4.11 shows the speech signal 'calm', as spoken by a female, processed using the variable frame rate algorithm. The speech signal was preprocessed by highpass filtering it with a cutoff frequency of 700 Hz. Also shown are the spectrogram of the speech signal, the interpolated transitivity function, and the spectrograms of the transient and the quasi-steady-state components. The spectrograms were computed using a Hanning window of lengths 15 ms and window overlap of 5 ms. The transient-activity threshold was 2.5. At this threshold, the quasi-steady-state component had 91 % of the total signal energy.

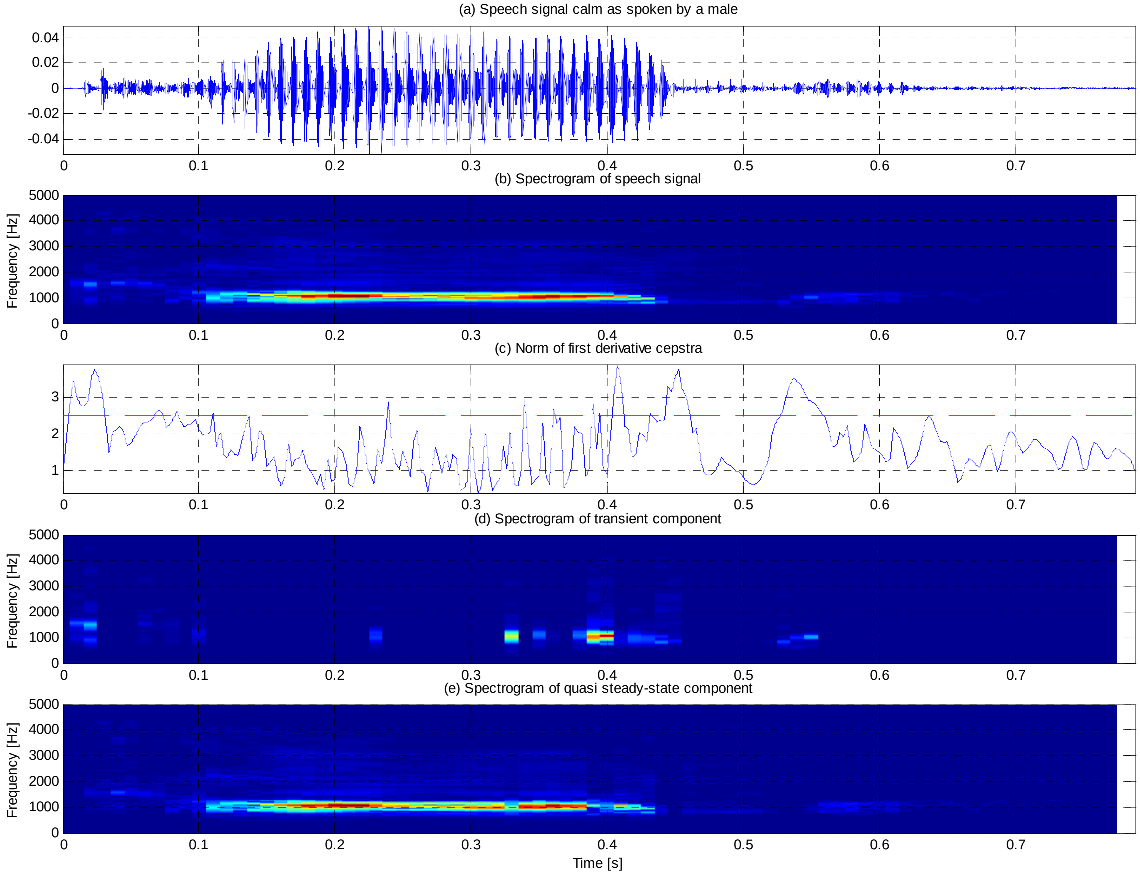

Figure 4.11: (a) Speech signal for the word 'calm' as spoken by a female speaker, (b) spectrogram of the speech signal, (c) transitivity function and transient-activity threshold, (d) spectrogram of transient component and (e) spectrogram of quasi-steady-state component.

To evaluate how closely the estimates of the transient and quasi-steady-state speech components synthesized using the algorithm that combines wavelet packet analysis with VFR approximated Yoo's nontonal and tonal components, the former were compared to the latter by comparing the magnitudes of the respective Fourier transforms and using an informal subjective listening test, conducted by the author.

As a means of comparing the transient component synthesized using wavelets to Yoo’s nontonal component across many words, the estimation errors were computed for 18 words using the mean-squared-error (MSE) between the spectra of the two components.

4.2 RESULTS

Results for choosing a wavelet function and a decomposition level will be presented first, followed by the classification of terminal nodes results. Finally the results for the synthesis of transient and quasi-steady-state speech will be given.

4.2.1 Wavelet Packet decomposition of Speech

The filter profiles observed for Daubechies wavelet function were classified, according to their type of slope, into four types; flat, upward, downward and irregular. No particular filter profile slope produced noticeably better results than the others. The db20 wavelet function was used for the decompositions because its filter profile was the closest to flat for the decomposition levels studied. It also had smaller side-lodes, narrower filter frequency responses and its time support of 3.54 ms seemed appropriate for identifying speech transients, which we expect to occur over several milliseconds.

Generally, the bandwidth of the filter functions and the size of the side lobes decreased as the wavelet order increased. In the evaluations, there was no observed pattern in the slope of the filter profiles as wavelet order increased, and the number of wavelet functions with each type of filter profile (upwards, downwards, relatively flat and irregular) was evenly distributed across wavelet orders.

The energy profiles of words were also investigated as a criterion to determine which wavelet function to use. For a given word and decomposition level, different wavelet functions were observed to have similar energy profiles. An example is shown in Figure 4.12 for the word ‘pike’ spoken by a female, decomposed at level 3 using db4, db20 and db38 wavelet functions.

The circles show the energy profile for highpass filtered speech, the squares show the energy profile for tonal speech, and the triangles show the energy profile for nontonal speech. The three wavelet functions produce very similar energy profiles despite the differences in their support size. As a result, the energy profiles of words did not provide a useful indication of which wavelet function to use.

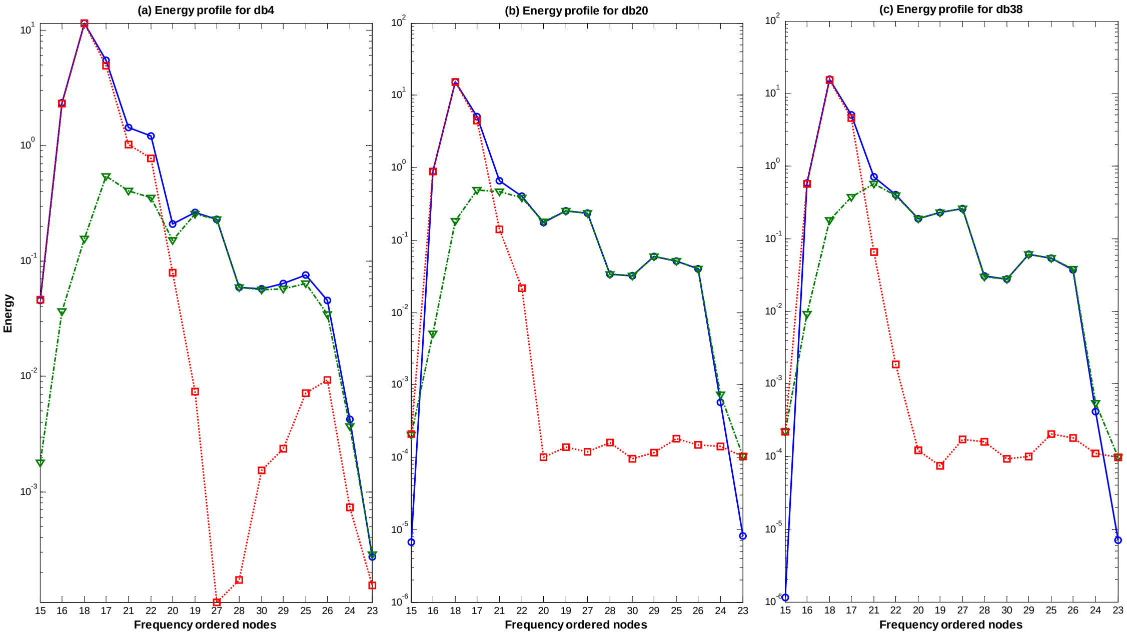

Figure 4.12: Energy profiles for (a) db4, (b) db20 and (c) db38 wavelet functions, for the word ‘pike’ spoken by a female.

The best ambiguity threshold value was the one at which the mean squared error (MSE) between the spectrum of the transient component and the spectrum of the nontonal component decreased the most when VFR was applied. To determine the best ambiguity threshold value for a decomposition level, the MSE between the spectra of the wavelet packet estimated transient component and the nontonal component was computed for 18 words and a range of threshold values. Then the MSE between the spectra of the wavelet packet estimated transient component with VFR processing and the nontonal component was also computed for the 18 words. ΔMSE, defined as MSE (without VFR) – MSE (with VFR), was interpreted as the gain in using VFR processing. A reduction in the MSE value (positive ΔMSE) indicated that the transient component estimate with VFR was better than the estimate without VFR processing. On the other hand, an increase in the MSE indicated that the transient component estimate without VFR was better.

Figure 4.13 shows a plot of average ΔMSE, averaged across 18 words, versus the ambiguity threshold, δ, for decomposition level of 6. From the plot, the ambiguity threshold values that maximize the gain of using VFR processing are 3.0 and 4.0 dB. The ambiguity threshold value of δ = 3.0 dB will be used for subsequent computation, since it was determined to be good, not only for level 6, but also for levels 4 and 5.

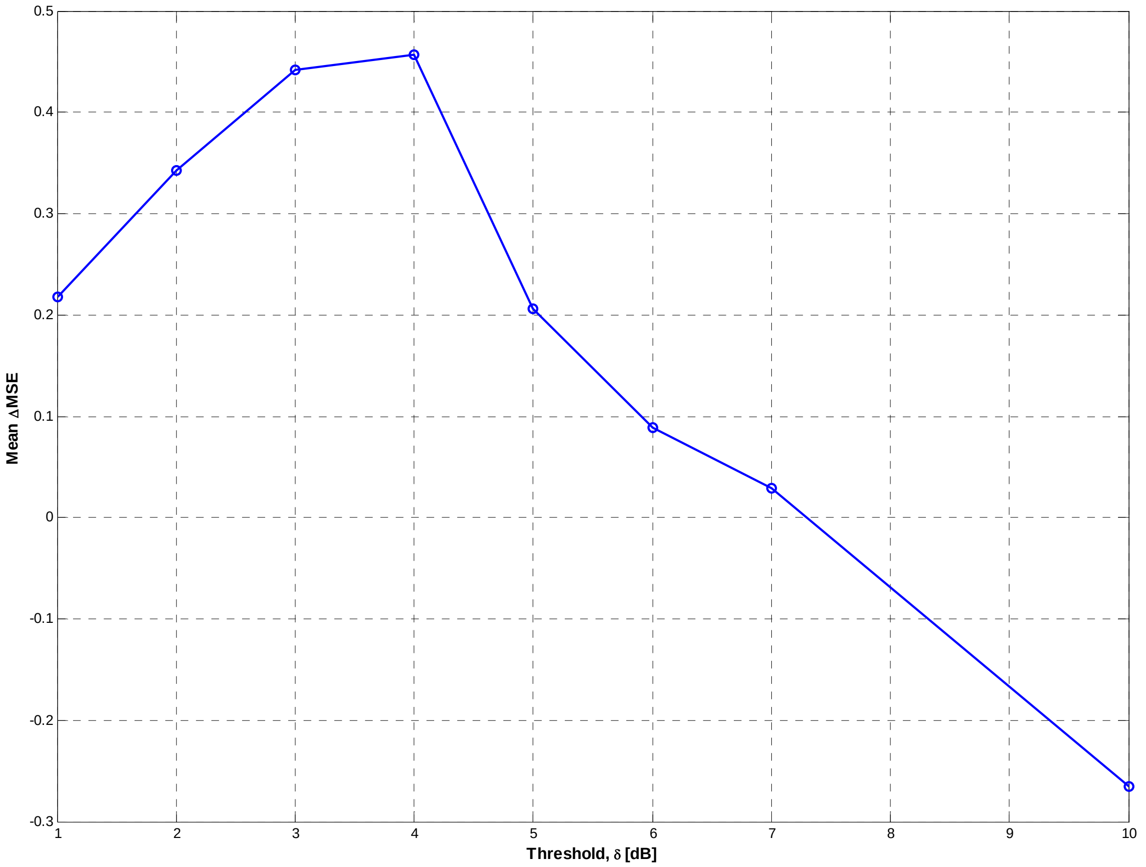

Figure 4.13: Determining the best ambiguity threshold, δ for decomposition level of 6.

When a decomposition level was chosen for a given word, the proportion of ambiguous nodes and the energy of these nodes were similar across decomposition levels. Table 4.2 shows the percentage of ambiguous nodes at decomposition levels of 3 to 6 for 18 highpass filtered speech signals. Table 4.3 shows the energies of these ambiguous nodes as a percentage of the energy of the highpass filtered speech. A db20 wavelet function was used in the decomposition and the ambiguity threshold was 3.0 dB.

The subscripts m and f denote whether the word was spoken by a female speaker or male speaker.

The average number of ambiguous nodes was lowest at level 3, but the average energy in ambiguous nodes was lowest at level 6. Level 6 was chosen, instead of level 3, as the best decomposition level because the differences in the average energy in ambiguous nodes across levels was more significant than the differences in the average number of ambiguous nodes across levels. Since the mean number of ambiguous nodes and mean energy of ambiguous nodes at other level were not very different from those observed at level 6, using other levels for the decomposition had little effect on the results.

Table 4.2: Percentage of ambiguous nodes for 18 words at decomposition levels 3 to 6 and ambiguity threshold of 3.0 dB.

	Level			
Word	**3**	**4**	**5**	**6**
pike$_m$	12.5000	25.0000	18.7500	23.4375
pike$_f$	12.5000	12.5000	12.5000	20.3125
calm$_m$	37.5000	25.0000	28.1250	26.5625
calm$_f$	25.0000	18.7500	25.0000	31.2500
nice$_m$	62.5000	50.0000	46.8750	39.0625
nice$_f$	0	18.7500	25.0000	21.8750
keg$_m$	12.5000	31.2500	34.3750	28.1250
keg$_f$	25.0000	25.0000	34.3750	34.3750
fail$_m$	12.5000	25.0000	28.1250	28.1250
fail$_f$	50.0000	31.2500	28.1250	28.1250
dead$_m$	37.5000	37.5000	40.6250	39.0625
chief$_f$	37.5000	37.5000	40.6250	32.8125
live$_m$	12.5000	25.0000	18.7500	25.0000
merge$_f$	25.0000	31.2500	28.1250	26.5625
juice$_f$	62.5000	37.5000	40.6250	29.6875
armchair$_f$	37.5000	31.2500	34.3750	31.2500
headlight$_m$	25.0000	31.2500	25.0000	26.5625
headlight$_f$	12.5000	31.2500	21.8750	28.1250
Mean	**27.7778**	**29.1667**	**29.5139**	**28.9063**

Table 4.3: Percentage of energy in ambiguous nodes for 18 words at decomposition levels 3 to 6 and ambiguity threshold of 3.0 dB.

	Level			
Word	**3**	**4**	**5**	**6**
pike_m	0.0009	2.5185	0.0009	1.3003
pike_f	0.0025	0.0001	0.0001	1.0956
calm_m	4.6353	4.0313	1.7472	2.4944
calm_f	3.2110	8.0377	5.1684	9.3160
nice_m	38.6399	32.7573	32.8854	18.9857
nice_f	0	9.8161	9.0530	10.0989
keg_m	4.8615	12.4233	16.7091	11.2798
keg_f	36.9813	21.1352	21.8985	31.6703
fail_m	12.2574	13.4497	15.0487	10.1699
fail_f	75.8187	24.6972	23.8000	17.7369
dead_m	23.3857	19.9028	16.2706	16.8531
chief_f	55.8857	35.8363	38.9554	25.5447
live_m	0.0012	5.1411	4.0116	6.9229
merge_f	17.2397	10.6731	13.7427	10.0224
juice_f	21.0630	10.0327	16.0793	5.0090
armchair_f	38.5106	22.6700	23.8268	18.8886
headlight_m	2.0049	12.9049	6.5212	8.8152
headlight_f	8.9237	20.6100	8.4095	11.2488
Mean	**19.0790**	**14.8132**	**14.1182**	**12.0807**

4.2.2 Classification of Terminal Nodes

To illustrate node classification, Figure 4.14 shows the energy profiles for highpass filtered, tonal and nontonal speech components for the word ‘pike’ as spoken by a female speaker. The decomposition level was 6 and a db20 wavelet function was used. The solid line with circles is the energy profile for the highpass filtered speech, the dotted line with squares is the energy profile for the tonal speech, and the dashed line with triangles is the energy profile for nontonal speech. For this example, at an ambiguity threshold value of 3.0 dB, quasi-steady state nodes include {70 68 67 75 76 78 77 73 74}, transient nodes are {71 87 88 90 93 94 92 91 83 84 86 85 81 82 80 79 111 112 114 113 117 118 116 115 123 124 126 125 121 122 120 119 103 104 106 105 109 110 108 107 100 102} and ambiguous nodes are {63 64 66 65 69 72 89 99 101 97 98 96 95}.

Table A4 in the appendix shows the node classification, at level 3 using an ambiguity threshold value of 3.0 dB, for 18 of the words studied, and Figure 4.15 shows, using filter frequency responses, the terminal node and their corresponding frequency ranges. The indices for level 4, 5 and 6 nodes that correspond to these levels are presented in Tables 3.1, 3.2 and 3.3.

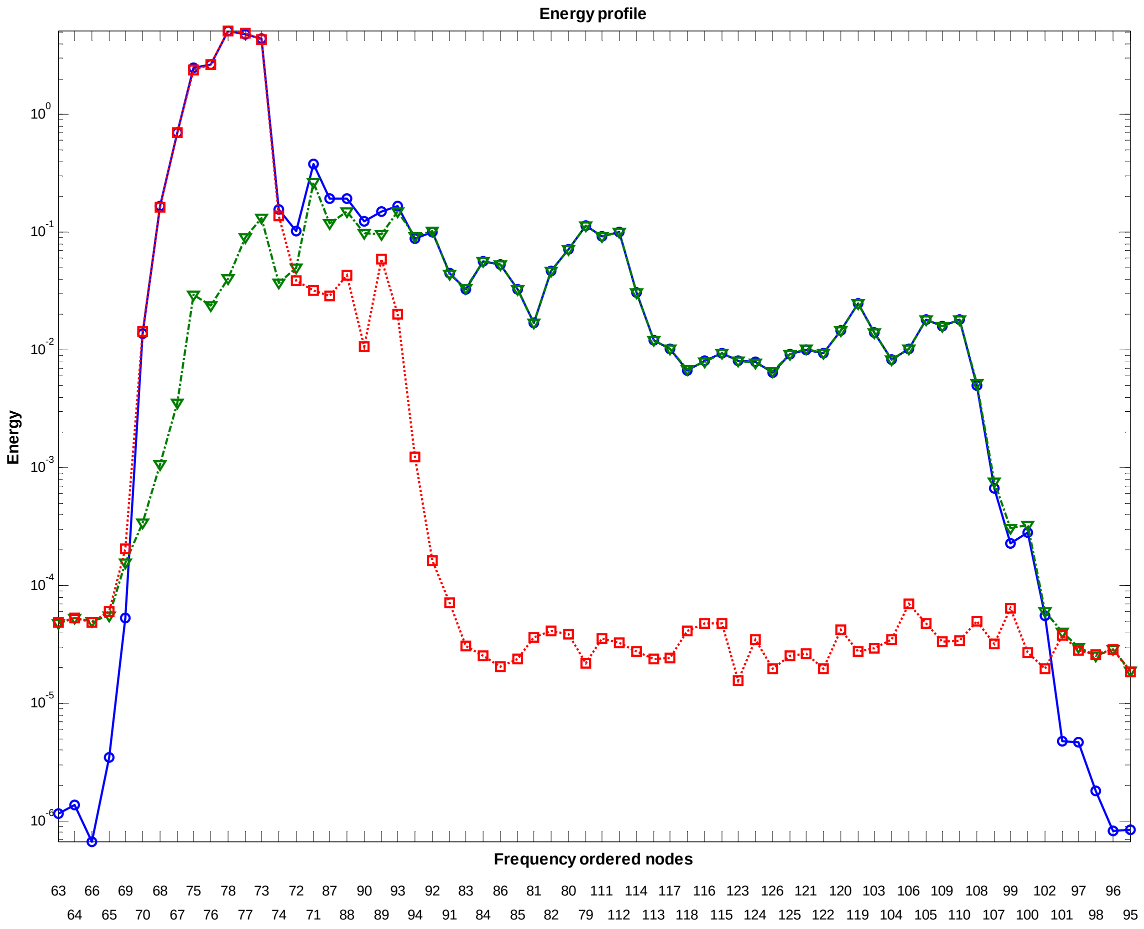

Figure 4.14: Node classification for the word ‘pike’ as spoken by a female.

Using energy profiles to categorize nodes and a db20 wavelet function, for most words, at least one of nodes 7, 8 and 10 or their children nodes were classified as quasi-steady-state nodes. These nodes correspond to the frequency range of 0 to 2100 Hz. In fact, if low frequency nodes (nodes 7 and 8) and high frequency nodes (nodes 12 and 11) which all have very low energy are ignored, most nodes in the frequency range 700 Hz to 2100 Hz were classified as quasi-steady-state nodes for most words. Additionally, nodes 9 and 13 and their children nodes, which include the frequency range of 2100 to 3400 Hz, were classified as transient nodes for most words. Ambiguous nodes were not restricted to any frequency interval.

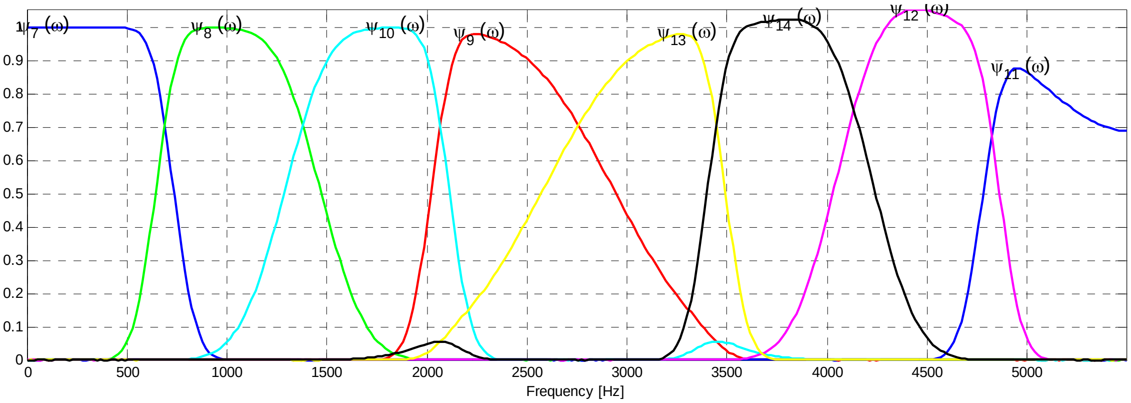

Figure 4.15: Terminal nodes and their corresponding frequency ranges.

4.2.3 Incorporation of Variable Frame Rate Processing and Synthesis of Speech Components

As a speech synthesis example, the top row of Figure 4.16 shows the spectra for the quasi-steady-state and transient components for the word ‘nice’ spoken by a male. A db20 wavelet function was used and the decomposition level was 6. Quasi-steady-state nodes were {63 64 66 70 75 76 78 77 73 74 72 71 87 88 90 89 93 118 116 124 125 121 122 119 106 105 109 102 101 97} and transient nodes were {65 69 68 67 94 92 91 83 84 86 85 81 82 80 79 111 112 114 113 117 115 123 126 120 103 104 110 108 107 99 100 98 96 95}.

The spectra of the tonal and nontonal components for the word are also shown in the figure as reference signals. Although the spectrum of the quasi-steady-state component has regions of low energy, it has spectral peaks (around 1380, 3290, 3540 and 4390 Hz) that are displayed by the tonal component. The biggest difference between the spectra of the transient and nontonal components was observed between 800 Hz and 1800 Hz. The transient component has very low energy in this frequency interval. Most of the nodes in this frequency interval were identified as and used to synthesize the quasi-steady-state component.

In a listening test, the transient component had the qualities of the nontonal component but was slightly more whispered. The quasi-steady-state component was more intelligible than the tonal component.

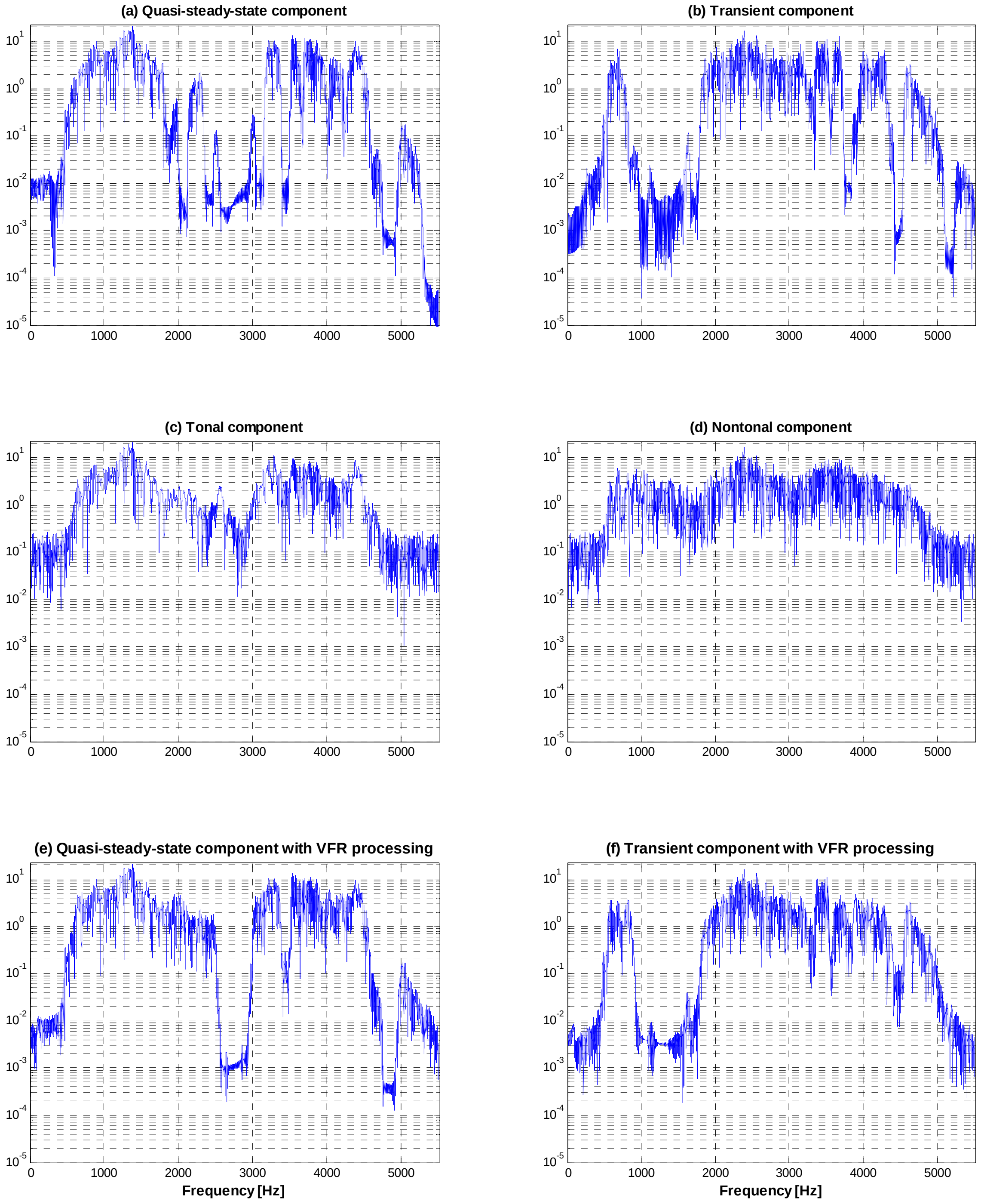

Figure 4.16: Spectra for (a) quasi-steady-state speech, (b) transient speech, (c) tonal speech (d) nontonal speech, (e) quasi-steady-state component with VFR processing, and (f) transient component with VFR processing for the word ‘nice’ spoken by a male.

When variable frame rate processing was applied to the ambiguous nodes, quasi-steady-state nodes were {70 78 77 73 74 72 71 87 88 90 89 93 118 116 124 121 106 105 109 102}, transient nodes were {68 83 84 86 85 81 82 80 79 111 112 114 115 123 110 108 107 99 100}, and ambiguous nodes were {63 64 66 65 69 67 75 76 94 92 91 113 117 126 125 122 120 119 103 104 101 97 98 96 95}. Parts (e) and (f) of Figure 4.16 show the spectra of the quasi-steady-state and transient components of the word ‘nice’, synthesized using an ambiguity threshold of 3.0 dB, when VFR processing was used. The bandwidth of regions of low energy observed when VFR was not applied to the estimations of the transient and quasi-steady-state components are reduced or eliminated by VFR. The spectra of the synthesized transient and quasi-steady-state speech approximated those of the nontonal and tonal speech more closely.

When the listening test was conducted, the transient and quasi-steady-state components synthesized using the combination of wavelet packets analysis and VFR sounded closer to the nontonal and tonal speech components.

Another synthesis example is shown in Figure 4.17. The top rows are the spectra of the quasi-steady-state and transient components for the word ‘chief’ spoken by a female. A db20 wavelet function was used and the decomposition level was 6. Quasi-steady-state nodes were {63 64 66 65 69 68 76 78 86 85 81 82 114 113 117 118 116 115 123 125 121 122 120 119 103 101 97 98} and transient nodes were {70 67 75 77 73 74 72 71 87 88 90 89 93 94 92 91 83 84 80 79 111 112 124 126 104 106105 109 110 108 107 99 100 102 96 95}.

Again the spectra of the tonal and nontonal components are also shown in the figure as reference signals. Except for the regions of low energy observed in the spectrum of the quasi-steady-state around 0.7, 2.7, 3.6 and 4.8 kHz, the spectrum of the quasi-steady-state component is similar to the spectrum of the tonal component. The spectrum of the transient component is similar to the spectrum of the nontonal component, except that regions of low energy, which were not present in the spectrum of the nontonal component, were observed around 2.4, 3.1, 3.4 and 4.1 kHz in the spectrum of the transient component.

When variable frame rate processing was applied to the ambiguous nodes, quasi-steady-state nodes were {68 76 81 114 113 117 118 116 115 121 122 120 119 101 97}, transient nodes were {70 67 77 73 74 72 71 87 88 90 89 93 94 92 91 83 111 124 126 106 105 109 110 108 107 99 100 102}, and ambiguous nodes were {63 64 66 65 69 75 78 84 86 85 82 80 79 112 123 125 103 104 98 96 95}. The spectra of the synthesized transient and quasi-steady-state speech approximated those of the nontonal and tonal speech more closely. Parts (e) and (f) of Figure 4.17 show the spectra of the quasi-steady-state and transient components of the word ‘chief’ as spoken by a female, when VFR processing was used. The bandwidth of regions of low energy observed when VFR was not applied to the estimation of the transient and quasi-steady-state components are reduced or eliminated by VFR.

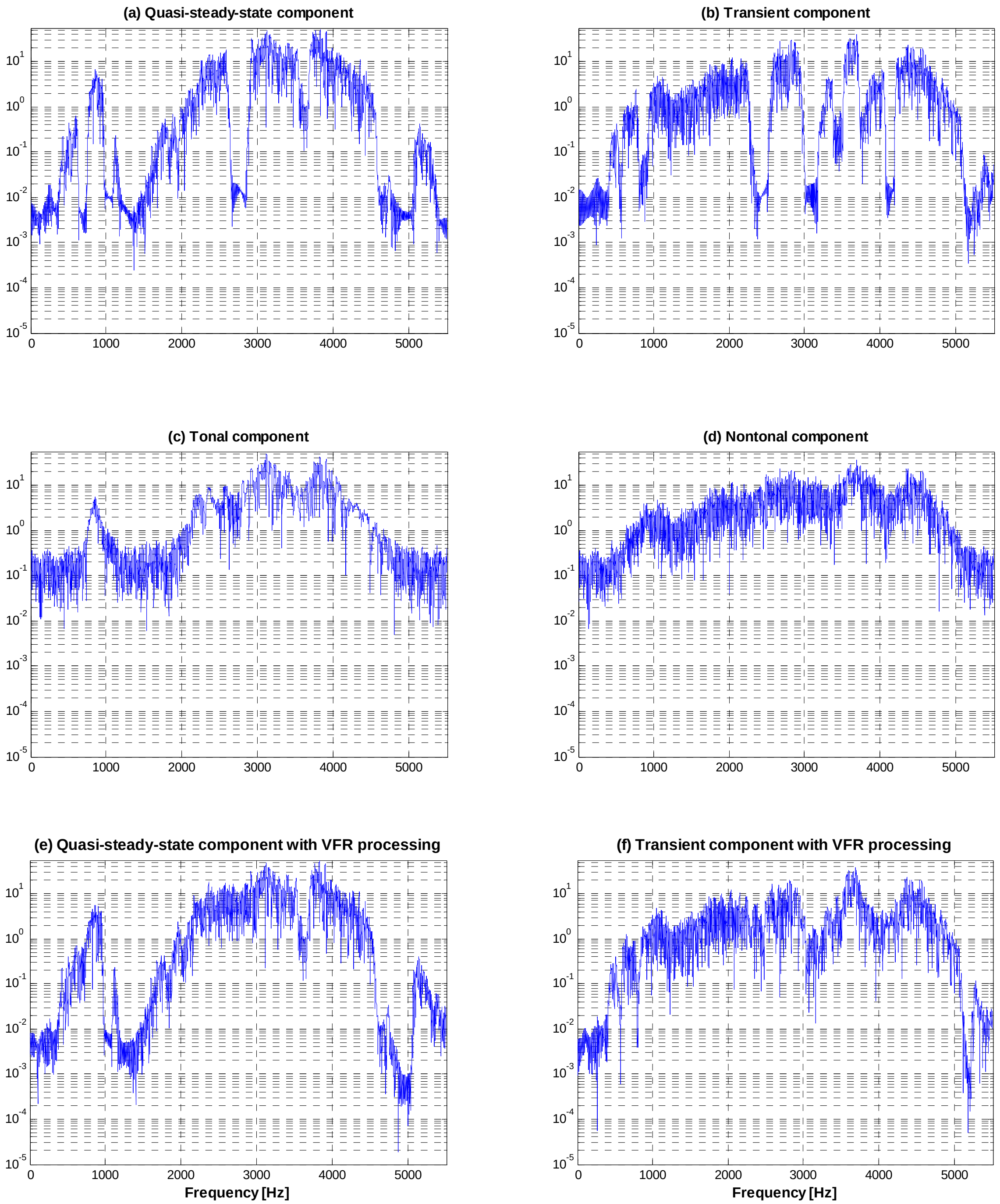

Figure 4.17: Spectra for (a) quasi-steady-state speech, (b) transient speech, (c) tonal speech (d) nontonal speech, (e) quasi-steady-state component with VFR processing, and (f) transient component with VFR processing for the word ‘chief’ spoken by a female.

When an informal listening test was conducted, the synthesized transient and quasi-steady-state speech for the word ‘chief’ spoken by a female with VFR processing applied sounded closer to the nontonal and tonal components than the synthesized transient and quasi-steady-state speech without application of VFR processing.

Generally, variable frame rate processing improved the synthesis of the transient and quasi-steady-state for most words studied. The gain in using VFR, measured using the mean-squared-error (MSE) between the spectra of the transient component estimate and the nontonal component, is given in Table 4.4 for 18 of the words studied. The decomposition level was 6 and a ambiguity threshold of 3.0 dB was used. A positive ΔMSE indicates that the VFR process improved the estimation of the nontonal component. An improvement, although small, in the MSE values was observed for most words.

Table 4.4: MSE improvements gained when VFR processing was used.

Word	**MSE Without VFR**	**MSE With VFR**	**ΔMSE**
pike_m	0.6187	0.6081	0.0106
pike_f	0.3990	0.4060	-0.0071
calm_m	0.0330	0.0356	-0.0025
calm_f	2.0300	1.9945	0.0355
nice_m	2.1415	1.8709	0.2706
nice_f	0.2394	0.2566	-0.0172
keg_m	6.1525	4.8595	1.2929
keg_f	6.5139	5.0876	1.4262
fail_m	10.6263	11.9313	-1.3050
fail_f	0.0883	0.0975	-0.0092
dead_m	1.2834	1.1156	0.1678
chief_f	12.9034	9.1226	3.7808
live_m	0.9183	0.9273	-0.0090
merge_f	8.6343	7.7955	0.8388
juice_f	2.2468	2.3478	-0.1010
armchair_f	14.2346	12.7887	1.4459
headlight_m	2.6789	2.5480	0.1308
headlight_f	0.0783	0.0691	0.0093
Mean	**3.9900**	**3.5479**	**0.4421**

5.0 DISCUSSION

An algorithm that identifies and selectively emphasizes speech transitions may enhance the intelligibility of speech in noisy conditions. Yoo et al described an algorithm that uses time-varying bandpass filters to decompose speech into tonal and nontonal speech components [53], [54] [55]. The tonal component predominately included steady-state formant activity and most of the signal energy, but was generally unintelligible. The nontonal component predominately included transitions within and between formants, was as intelligible as the original speech, and had much lower energy. This study evaluated the effectiveness of DWT, SWT and wavelet packets to identify transient and quasi-steady-state speech components that are close estimates of Yoo's components.

In the investigation of the DWT and SWT, the transient speech component was synthesized using wavelet levels that were identified as having predominately transient information, while the quasi-steady-state speech component was synthesized using wavelet levels that were identified as having predominately quasi-steady-state information. This meant that transient information present at a wavelet level identified as having quasi-steady-state information was not included in the synthesis of the transient component. Conversely, quasi-steady-state information present at a wavelet level identified as having transient information was excluded in the synthesis of the quasi-steady-state component. This may result in errors in estimating transient and quasi-steady-state components. The synthesized transient component had most of its energy concentrated in the frequency ranges that correspond to the wavelet levels that were identified as having predominately transient information, and the quasi-steady-state component had most of its energy concentrated in the frequency range that correspond to the wavelet levels that were identified as having quasi-steady-state information. For most words, the DWT and SWT identified the same wavelet levels as having transient or quasi-steady-state information; as a result the speech components synthesized using these two transforms were very similar, despite the additional redundancy offered by the SWT.

The wavelet packets and wavelet transforms identify transient and quasi-steady-state speech components in similar ways. In both algorithms, frequency ranges, which are interpreted as levels in the wavelet transform analysis, and as nodes in the wavelet packet analysis, are associated with either the transient or quasi-steady-state speech components. The advantage of using wavelet packets is that they offer a finer, more evenly spaced division of the frequency spectrum. When synthesizing speech components using wavelet packets, as with the wavelet transforms, the energy of the synthesized components is concentrated in the frequency ranges identified as having each kind of information (transient or quasi-steady-state). But since wavelet packets divide the spectrum into finer frequency ranges, the frequency ranges identified as having either kind of information generally have a narrower bandwidth.

An algorithm that incorporates wavelet packet analysis and variable frame rate processing for identifying transient and quasi-steady-state speech was presented. This algorithm included the processes of choosing a wavelet function, choosing a decomposition level, classifying terminal nodes of a decomposed speech signal, synthesis of transient and quasi-steady-state speech, and incorporation of the VFR process into the wavelet analysis.

In the process of choosing a wavelet function, a filters profile for a wavelet function was first defined and filter profiles for Daubechies wavelets of different orders were compared. The filter profile for the db20 wavelet function had a flatter profile, smaller side-lodes and narrower filter frequency responses than other functions. This wavelet function was used for the decompositions.

The average number and average energy of ambiguous nodes were used to select the best decomposition level. The average number of ambiguous nodes was lowest at level 3, but the average energy in ambiguous nodes was lowest at level 6. Level 6 was chosen, instead of level 3, as the best decomposition level because the differences in the average energy in ambiguous nodes across levels were larger than the differences in the average number of ambiguous nodes across levels. Using other levels for the decomposition did not change the final results significantly since the differences in the average number of ambiguous nodes and average energy of ambiguous nodes were small. A threshold for defining ambiguous nodes was referred to as the ambiguity threshold, and the best ambiguity threshold was determined to be 3.0 dB.

The energy profiles of the highpass filtered, tonal and nontonal speech were used to classify terminal nodes of the highpass filtered speech into three group; quasi-steady-state, transient, and ambiguous nodes. For the words studied, nodes classified as transient nodes or quasi-steady-state nodes were not restricted to any frequency ranges. But generally, at least one node in the frequency range of 0 to 2100 Hz was classified as a quasi-steady-state node, and at least one node in the frequency range of 2100 to 3400 Hz was classified as a transient node. This supports the idea that quasi-steady-state activity is predominately lowpass, while transient activity is predominately highpass.

A limitation of using wavelet packets without VFR processing was that region of low energy, which were not present in the spectra of the tonal and nontonal speech components, were observed in the spectra of the synthesized quasi-steady-state and transient components. These regions of low energy occurred in the frequency ranges where nodes were classified as ambiguous. Incorporation of variable frame rate processing improved the quality of the synthesized components. With the incorporation of variable frame rate processing, wavelet coefficients of an ambiguous node were either included in the synthesis of the transient component or quasi-steady-state component depending on the value of the quantized transitivity function. This reduced the bandwidth of the regions of low energy in the spectral estimates and produced transient and quasi-steady-state speech components that were the closest estimates of Yoo's nontonal and tonal speech components, respectively.

To synthesize transient and quasi-steady-state speech components, the proposed wavelet-based algorithm depends on the knowledge of Yoo's nontonal and tonal speech components. In the future, investigations will be carried out to formulate a wavelet-based algorithm that can decompose speech into transient and quasi-steady-state speech components without knowledge of Yoo's speech components.

The algorithm of Yoo is computationally intensive and unsuitable for real-time applications. The WP-VFR algorithm may provide a method to identify transient speech components with significantly less computation time. This approach may provide a method to implement a real-time speech enhancement algorithm using transient speech information.

APPENDIX

LEVEL AND NODE CLASSIFICATIONS

Table A 1: DWT level classification for 18 words

Word	Quasi-steady-state levels	Transient levels
pike$_m$	2 3 4 6	1 5
pike$_f$	3 4 6	1 2 5
calm$_m$	3 6	1 2 4 5
calm$_f$	2 3 4	1 5 6
nice$_m$	1 2 3 6	4 5
nice$_f$	2 3 4 5 6	1
keg$_m$	5	1 2 3 4 6
keg$_f$	1 3 4	2 5 6
fail$_m$	2 3	1 4 5 6
fail$_f$	2 3 4 6	1 5
dead$_m$	5 6	1 2 3 4
chief$_f$	1 5 6	2 3 4
live$_m$	2 3 5 6	1 4
merge$_f$	1 3 4 5	2 6
juice$_f$	1 2 5	3 4 6
armchair$_f$	1 3 6	2 4 5
headlight$_m$	2 3 4 5 6	1
headlight$_f$	3 4 6	1 2 5

Table A 2: SWT level classification for 18 words.

Word	Quasi-steady-state levels	Transient levels
pike_m	2 3 4 6	1 5
pike_f	3 4 5 6	1 2
calm_m	3	1 2 4 5 6
calm_f	2 3 4 6	1 5
nice_m	1 2 3 5 6	4
nice_f	2 3 4 5 6	1
keg_m	5	1 2 3 4 6
keg_f	1 3 4 5	2 6
fail_m	2 3	1 4 5 6
fail_f	2 3 4 5 6	1
dead_m	5 6	1 2 3 4
chief_f	1 5 6	2 3 4
live_m	2 3 5 6	1 4
merge_f	1 2 3 4 5	6
juice_f	1 2 5	3 4 6
armchair_f	1 3 6	2 4 5
$\text{headlight}_\text{m}$	2 3 4 6	1 5
$\text{headlight}_\text{f}$	3 4 6	1 2 5

Table A 3: WP Node classification for 18 words decomposed at depth 4.

Word	Quasi-steady-state nodes	Transient nodes
pike_{m}	16 18 17 21 22 24	15 20 19 27 28 30 29 25 26 23
pike_{f}	16 18 17 23	15 21 22 20 19 27 28 30 29 25 26 24
calm_{m}	18 17 21 24	15 16 22 20 19 27 28 30 29 25 26 23
calm_{f}	16 18 17 21 22 20 24 23	15 19 27 28 30 29 25 26
nice_{m}	15 18 17 21 28 30 29 25 23	16 22 20 19 27 26 24
nice_{f}	15 16 18 21 22 25 26 23	17 20 19 27 28 30 29 24
keg_{m}	15 18 20 19 27 24	16 17 21 22 28 30 29 25 26 23
keg_{f}	16 18 19 27 28 30 23	15 17 21 22 20 29 25 26 24
fail_{m}	15 18 17 21 22 20 19 27 23	16 28 30 29 25 26 24
fail_{f}	15 16 18 19 23	17 21 22 20 27 28 30 29 25 26 24
dead_{m}	15 18 20 19 27 24 23	16 17 21 22 28 30 29 25 26
chief_{f}	15 18 20 19 27 28 29 23	16 17 21 22 30 25 26 24
live_{m}	15 17 21 22 24 23	16 18 20 19 27 28 30 29 25 26
merge_{f}	15 16 18 19 27 25 23	17 21 22 20 28 30 29 26 24
juice_{f}	18 19 27 28 30 29 26 23	15 16 17 21 22 20 25 24
armchair_{f}	18 17 21 19 27 28	15 16 22 20 30 29 25 26 24 23
headlight_{m}	15 16 18 17 21 22 24 23	20 19 27 28 30 29 25 26
headlight_{f}	16 18 17 23	15 21 22 20 19 27 28 30 29 25 26 24

Table A 4 WP Node classification for 18 words decomposed at level 3.

Word	Quasi-steady-state nodes	Transient nodes	Ambiguous nodes
pike$_m$	7 8 10	9 13 14 12	11
pike$_f$	7 8	10 9 13 14 12	11
calm$_m$	8	9 13 14 12	7 10 11
calm$_f$	7 8 10	13 14 12	9 11
nice$_m$	8 10	9	7 13 14 12 11
nice$_f$	7 8 10 12	9 13 14 11	8
keg$_m$	9 11	7 10 13 14 12	9 14
keg$_f$	7 8 13	10 12 11	13
fail$_m$	8 10 9	7 14 12 11	8 9 13 11
fail$_f$	7	10 14 12	7 8 13
dead$_m$	8 10	9 13 14 12	7 11
chief$_f$	7 8	9 13 14 12 11	10
live$_m$	10	7 9 13 14 12	8 11
merge$_f$	7	14	8 10 9 13 12 11
juice$_f$	13 14	10	7 8 9 12 11
armchair$_f$	8	7 12 11	10 9 13 14
headlight$_m$	8	9 12	7 10 13 14 11
headlight$_f$	8	13 14 12	7 10 9 11

BIBLIOGRAPHY

[1] Bahoura, M. & Rouat, J. (2001), "Wavelet Speech Enhancement Based on the Teager Energy Operator", *IEEE Signal Processing Letters,* vol. 8, no. 1, pp. 10-12, January 2001.

[2] Barros, K. A., Rutkowski*,* T., Itakura*,* F. and Ohnishi, N. (2002), "Estimation of Speech Embedded in a Reverberant and Noisy Environment by Independent Component Analysis and Wavelets", *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 888 – 893, July 2002.

[3] Brookes, M. (1997), *Voicebox: Speech Processing Toolbox for MATLAB*, Imperial College London, Department of Electrical and Electronics Engineering Speech Processing Research Team.

[4] Brown, K. L., Algazi, V. R. (1989), "Characterization of Spectral Transitions with Applications to Acoustic Sub-word segmentation and Automatic Speech Recognition", *Proceedings of IEEE ICASSP '89*, vol. 1, pp. 104 – 107, May 1989.

[5] Burrus, C. S., Gopinath, R. A. & Guo, H. (1998), *Introduction to Wavelets and the Wavelet Transforms: A Primer,* Prentice Hall, Upper Saddle River, NJ.

[6] Chen, S. H. & Wang, J. F., (2002), "Noise-robust pitch detection method using wavelet transform with aliasing compensation", *IEE Proceedings of Vision, Image and Signal Processing*, vol. 149, no. 6, pp. 327 – 334, December 2002.

[7] Daubechies, I. (1992), *Ten Lectures on Wavelets,* Philadelphia: SIAM, CBMS-NSF Regional Conference in Applied Mathematics 61.

[8] Daubechies, I. (1996), "Where do Wavelets Come From?-A Personal Point of View", *Proceedings of IEEE*, vol. 84, no. 4, pp. 510-513, April 1996.

[9] Daudet, L. & Torresani, B. (2002), "Hybrid representation for audiophonic signal encoding", *Signal Processing,* vol. 82, pp. 1595 – 1617, 2002.

[10] Davis, S. & Mermelstein, P. (1980), "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. Acoustics, Signal Processing,* vol. 28, pp. 357-366, 1980.

[11] Deller, J. R., Proakis, J. G. & Hansen, J. H. L. (1993), *Discrete-Time Processing of Speech Signals,* Macmillan Inc, New York.

[12] Ephraim, Y. & Malah, D., “Speech enhancement using a minimum mean square error short time spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, June 1984.

[13] Farooq, O. & Datta, S. (2001), “Mel Filter-like Admissible Wavelet Packet Structure for Speech Recognition”, *IEEE Signal Processing Letters,* vol. 8, no. 7, pp. 196-198, July 2001.

[14] Farooq, O. & Datta, S. (2003), “Wavelet-based denoising for robust feature extraction for speech recognition”, *Electronics Letters*, vol. 36, no. 1, pp. 163-165, January 2003.

[15] Fant, G. (1973), *Speech Sounds and Features*, MIT Press, Cambridge, MA.

[16] Favero, R. F. (1994), “Compound Wavelets: Wavelets for Speech Recognition”, *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis,* pp. 600-603, October 1994.

[17] Jansen, M. (2001), Noise Reduction by Wavelet Thresholding, Springer-Verlag, *Lecture notes in Statistics*, vol. 161, 2001.

[18] Joos, M. (1948), “Acoustic Phonetics”, *Language Monograph*, Linguistic Society of America, Baltimore.

[19] Jing, L. I. & Changchun, B. A. O. (2002), “A Pitch Detector based on the dyadic wavelet transform and the autocorrelation function”, *6th International Conference on Signal Processing*, vol. 1 pp. 414-417, August 2002.

[20] Kadambe, S. & Srinivasan, P. (1992), “Application of the Wavelet Transform for Pitch Detection of Speech Signals”, *IEEE Transaction on Information Theory,* vol. 38, no. 2, pp. 917-924, March 1992.

[21] Kadambe, S. & Srinivasan, P. (1997), “Adaptive Wavelet Based Phoneme Recognition”, *IEEE Proceedings of the 40th Midwest Symposium on Circuits and Systems*, vol. 2, pp. 720-723, August 1997.

[22] Kaiser, J. F. (1993), “Some useful properties of Teager’s energy operators”, *Proceedings of IEEE ICASSP ‘93*, vol. 3, pp. 149–152, Apr. 1993.

[23] Le Cerf, P. & Van Compernolle, D. (1992), “Frame and Frame Dimension Reduction Techniques for Automatic Speech Recognition”, *Proceedings of 11th IARP/IEEE Conference on Image, Speech and Signal Analysis,* pp. 717-720 August 1992.

[24] Le Cerf, P. & Van Compernolle, D. (1994), “A New Variable Frame Rate Analysis Method for Speech Recognition”, *IEEE Signal Processing Letters*, vol. 1, no. 12, pp. 185-187, December 1994.

[25] Liberman, A. M., Delattre P. C., Cooper, F. S., Gerstman, L. J. (1954), “The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants”, *Psychology Monographs: General and Applied*, vol. 68, no. 8, pp. 1 - 13, 1954.

[26] Liberman, A. M., Cooper, F. S. (1972), “In Search of the Acoustic Cues”, in Valdman, A. (Editor), *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*, pp. 329-338, Mouton, Netherlands, 1972.

[27] Liberman, A. M., Harris, K. S., Hoffman, H. S., Delattre, P. C., Cooper, F. S. (1958), “Effect of Third-Formant Transitions on the Perception of the Voiced Stop Consonants”, *The Journal of the Acoustical Society of America*, vol. 30, no. 2 pp. 122-126, February 1958.

[28] Liberman, A. M., Cooper, F. S., Shankweiler, D. P., Studdert-Kennedy, M. (1967), “Perception of the Speech Code”, *Psychological Review,* American Psychological Association, vol. 74, no. 6, pp. 431-461, 1967.

[29] Logan, B. (2000), “Mel Frequency Cepstral Coefficients for Music Modeling”, *International Symposium on Music Information Retrieval,* 2000.

[30] Mallat, S. & Hwang, W. L. (1992), “Singularity detection and processing with wavelets,” *IEEE Trans. Inform. Theory*, vol. 38, pp. 617–643, Mar. 1992

[31] Mallat, S. G. (1989), “A Theory for Multiresolution Signal Decomposition: The Wavelet Representation”, *IEEE Transactions Pattern Analysis Machine Intelligence*, vol. 11 no. 7, pp. 674-693, July 1989.

[32] Mallat, S. G. (1998), *A Wavelet Tour of Signal Processing,* Academic Press, Chestnut Hill, MA.

[33] Malvar, H. S. (1990), “Lapped transforms for efficient transform/subband coding”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 6, pp. 969-978, June 1990.

[34] Mandridake, E. & Najim, M. (1993), “Joint Wavelet Transform and Vector Quantization for Speech Coding”, *IEEE International Symposium on Circuits and Systems,* pp. 699-702, May 1993.

[35] Mertins, A. (1999), *Signal Analysis: Wavelets, Filter Banks, Time-Frequency Transforms and Applications,* John Wiley and Sons, West Sussex England.

[36] Molau, S., Pitz, M., Schluter, R. & Ney, H. (2001), "Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum", *Proceedings of IEEE ICASSP "01,* vol. 1, pp. 73-76, May 2001.

[37] Morlet, J., Arens, G., Fourgeau, I. & Giard, D. (1982), "Wave Propagation and Sampling Theory", *Geophysics,* vol. 47, no. 2, pp. 203-236, February 1982.

[38] Najih, A. M. M. A., bin Ramli, A. R., Prakash, V. & Syed, A. R. (2003), "Speech Compression Using Discrete Wavelet Transform", *Proceedings of the Fourth National Conference on Telecommunication Technology*, pp. 1-4, January 2003.

[39] Nason, G.P., B.W. Silverman (1995), "The stationary wavelet transform and some statistical applications," *Lecture Notes in Statistics*, vol. 103, pp. 281-299.

[40] Ponting, K. M. & Peeling, S. M. (1991), "The Use of Variable Frame Rate Analysis in Speech Recognition", *Computer Speech and Language,* vol. 5, pp. 169-179, April 1991.

[41] Potter, R. K., Kopp, G. A., Green H. C. (1947), *Visible Speech*, Van Nostrand, New York.

[42] Quatieri, T. F., Dunn, R. B. (2002), "Speech Enhancement Based on Auditory Spectral Change", *Proceedings of IEEE ICASSP "02,* vol. 1, pp. 257-260, May 2001.

[43] Rabiner, L. & Juang, B. H. (1993), *Fundamentals of Speech Recognition,* Prentice Hall, New Jersey.

[44] Rao, A. & Kumaresan R. (2000), "On Decomposing Speech into Modulated Components", *IEEE Transactions on Speech and Audio Processing,* vol. 8, pp. 240-254, May 2000.

[45] Ris, C., Fontaine, V. & Leich, H. (1995), "Speech Analysis Based on Malvar Wavelet Transform", *Proceedings of IEEE ICASSP '95*, vol. 1, pp. 389-392, May 1995.

[46] Shelby, G. A., Cooper, C. M. & Adhami, R. R. (1994), "A Wavelet-based Pitch Detector for Tone Languages", *IEEE International Symposium on Time-Frequency and Time-Scale Analysis*, pp. 596-599, October 1994.

[47] Teolis, A. (1998), *Computational Signal Processing with Wavelets,* Birkhauser, Boston, MA.

[48] Vaidyanathan, P. P. & Djokovic, I. (2000), “Wavelet transform”, in Chen, W. K. (Editor), *Mathematics for Circuits and Filters,* pp. 131-216, CRC Press LLC, Boca Raton, FL.

[49] Walter, G. G. (1994), *Wavelets and Other Orthogonal Systems With Applications,* CRC Press Inc., Boca Raton, FL.

[50] Xiaodong, W., Yongming, Li, Hongyi, C. (1998), “Multi-domain speech compression based on wavelet packet transform”, *IEEE Electronics Letters*, vol. 34, no. 2, pp. 154-155, January 1998.

[51] Yao, J. & Zhang, Y. (2002), “The Application of Bionic Wavelet Transform to Speech Signal Processing in Cochlear Implants using Neural Network Simulations”, *IEEE Transactions on Biomedical Engineering*, vol. 49 , no. 11, pp. 1299 – 1309, November. 2002.

[52] Yoo, S., Boston, J. R., Durrant, J. D., El-Jaroudi, A., & Li, C. C. (2003), “Speech Decomposition and Intelligibility”, *Proceedings of the World Congress on Medical Physics and Biomedical Engineering*, 29 August, 2003, Sydney, Australia.

[53] Yoo, S., Boston, J.R., Durrant, J.D., Kovacyk, K., Karn, S., Shaiman, S. El-Jaroudi, A. & Li., C.C. (2004) “Relative energy and intelligibility of transient speech components”, *Proc. of the 12th European Signal Processing Conference*, Vienna, Austria, pp. 1031-1034, September 6-10, 2004.

[54] Yoo, S., Boston, J.R., Durrant, J.D., Kovacyk, K., Karn, S., Shaiman, S. E-Jaroudi, A. & Li., C.C. (2005) “Relative energy and intelligibility of transient speech components”, *Proceedings of IEEE ICASSP ’2005*, March 2005.

[55] Yoo, S. (2003). *Speech Decomposition and Enhancement,* University of Pittsburgh, Pittsburgh, PhD proposal exam, June 2003.

[56] Zhu, Q. & Alwan, A. (2000), “On the Use of Variable Frame Rate Analysis in Speech Recognition”, *Proceedings of IEEE ICASSP ‘00,* vol. 3, pp. 1783-1786, June 2000.