ENHANCEMENT OF SPEECH INTELLIGIBILITY USING SPEECH TRANSIENTS EXTRACTED BY A WAVELET PACKET-BASED REAL-TIME ALGORITHM

by

Daniel Motlotle Rasetshwane

B.S.E.E., University of Pittsburgh, 2002

M.S.E.E, University of Pittsburgh, 2005

Submitted to the Graduate Faculty of

Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Daniel Motlotle Rasetshwane

It was defended on

June 30, 2009

and approved by

Ching-Chung Li, Professor, Department of Electrical and Computer Engineering

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

Patrick Loughlin, Professor, Department of Bioengineering

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

Susan Shaiman, Associate Professor, Department of Communications Sciences and Disorders

Dissertation Director: J. Robert Boston, Professor, Department of Electrical and Computer Engineering

ENHANCEMENT OF SPEECH INTELLIGIBILITY USING SPEECH TRANSIENTS EXTRACTED BY A WAVELET PACKET-BASED REAL-TIME ALGORITHM

Daniel Motlotle Rasetshwane, PhD

University of Pittsburgh, 2009

Studies have shown that transient speech, which is associated with consonants, transitions between consonants and vowels, and transitions within some vowels, is an important cue for identifying and discriminating speech sounds. However, compared to the relatively steady-state vowel segments of speech, transient speech has much lower energy and thus is easily masked by background noise. Emphasis of transient speech can improve the intelligibility of speech in background noise, but methods to demonstrate this improvement have either identified transient speech manually or proposed algorithms that cannot be implemented to run in real-time.

We have developed an algorithm to automatically extract transient speech in real-time. The algorithm involves the use of a function, which we term the transitivity function, to characterize the rate of change of wavelet coefficients of a wavelet packet transform representation of a speech signal. The transitivity function is large and positive when a signal is changing rapidly and small when a signal is in steady state. Two different definitions of the transitivity function, one based on the short-time energy and the other on Mel-frequency cepstral coefficients, were evaluated experimentally, and the MFCC-based transitivity function produced better results. The extracted transient speech signal is used to create modified speech by combining it with original speech.

To facilitate comparison of our transient and modified speech to speech processed using methods proposed by other researcher to emphasize transients, we developed three indices. The

indices are used to characterize the extent to which a speech modification/processing method emphasizes (1) a particular region of speech, (2) consonants relative to vowels, and (3) onsets and offsets of formants compared to steady formant. These indices are very useful because they quantify differences in speech signals that are difficult to show using spectrograms, spectra and time-domain waveforms.

The transient extraction algorithm includes parameters which when varied influence the intelligibility of the extracted transient speech. The best values for these parameters were selected using psycho-acoustic testing. Measurements of speech intelligibility in background noise using psycho-acoustic testing showed that modified speech was more intelligible than original speech, especially at high noise levels (-20 and -15 dB). The incorporation of a method that automatically identifies and boosts unvoiced speech into the algorithm was evaluated, showing that this process does not result in additional speech intelligibility improvements.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This dissertation is dedicated to the memory of my sister, Morwadi Esther Rasetshwane. I miss you.

1.0 INTRODUCTION

The aim of speech enhancement, discussed in Chapter 2, is to improve the intelligibility and/or quality of speech in order to facilitate better communication in noisy environments. Conventional speech enhancement techniques try to remove noise from a noisy speech signal with minimal impact on the speech itself. Another approach to speech enhancement, which has not received as much attention, is to modify the speech signal itself, by emphasizing certain acoustical cues, in order to make it more intelligible in noisy environments.

Several studies, also discussed in Chapter 2, have shown that speech transients are important acoustical cues for identifying and discriminating speech sounds. Speech transients are associated with consonants, transitions between consonants and vowels, and transitions within some vowels. Compared to the relatively steady-state vowel segments of speech, these transients have much lower energy and thus are easily masked by background noise.

Yoo *et al.*. and Tantibundhit *et al.*., in our research laboratory, identified speech transients and showed that selective amplification of speech transients is effective in improving the intelligibility of speech in background noise at moderate (SNR -10 dB) and severe (SNR of -25 to -15 dB) noise levels [1] [2] [3] [4]. However, because of their complexity, the algorithms of Yoo *et al.*. and Tantibundhit *et al.*. are computationally intense and have not been implemented to run in real time.

An algorithm proposed by Skowronski *et al.*, which modifies speech by increasing the energy of transitional regions and unvoiced speech, was implemented to run in real-time by Raghavan *et al.* [5] [6]. However, this algorithm was only evaluated at a moderate noise level (SNR of -6 dB). Another real-time algorithm that improves speech intelligibility by increasing the duration and intensity of transient speech was described by Jayan *et al.* [7]. However, they presented psycho-acoustic testing results for only one subject.

The goal of this study is to develop and evaluate an algorithm to automatically extract and enhance transient speech in real-time. The algorithm involves the use of a function, which we termed the transitivity function, to characterize the rate of change of wavelet coefficients of a wavelet packet transform representation of a speech signal. The transitivity function is large and positive when a signal is changing rapidly and small when a signal is in steady state. The extracted transient speech signal is used to create modified speech, for speech enhancement, by combining the amplified transient speech with original speech and adjusting the modified speech so that its energy is equal to that of original speech. The intelligibility of modified speech, relative to that of original speech, is evaluated using psycho-acoustic testing.

This thesis is organized as follows. Chapter 2, which is divided into two main parts, begins with a description of the nature of speech and methods that have been proposed for speech enhancement. Conventional speech enhancement methods, which we will refer to as noise reduction methods, are reviewed. These methods include spectral subtraction and Wiener filtering. Our speech enhancement method is based on modification of the original noise-free speech to emphasize transient speech to improve intelligibility, and several studies that have shown that transient speech is important for speech perception are discussed. This is followed by a review of works that have modified or processed speech to improve intelligibility. The second

part of Chapter 2 describes established tools and techniques that we will be using in the development and evaluation of our algorithm. These include the modified rhyme test, the short-time energy, the short-time zero-crossing rate, the short-time autocorrelation function, Mel-frequency cepstral coefficients and wavelet packets.

In Chapter 3, two alternate definitions of the transitivity function are presented and validated using a synthetic signal that models specific transient events that occur in speech. The transient extraction algorithm is then described. In the algorithm, speech is decomposed using the wavelet packet transform, and a transitivity function is computed for each sequence of wavelet packet coefficients. The wavelet packet coefficients are modified based on the transitivity functions and recombined to create transient speech. A method to boost unvoiced speech, which will be evaluated for incorporation into the transient extraction algorithm, is described.

In order to compare our transient and modified speech signals to transient, modified and processed speech signals obtained using methods proposed by other researchers, we developed three indices for characterizing and quantifying the extent to which transient speech is emphasized by a given method. These indices are described in Chapter 4. Chapter 4 follows with illustrations of transient and modified speech signals and comparison, using the three indices, of our transient and modified speech signals to transient, modified and processed speech signals obtained by other researchers.

The transient extraction algorithm includes parameters that influence the intelligibility of transient and modified speech signals. Use of the best values for these parameters will result in the most intelligible transient and modified speech. Chapter 5 describes psycho-acoustic experiments that were used to select these parameters and to evaluate the intelligibility of

modified speech. Results of these experiments are presented in Chapter 5. The findings of this study are discussed in Chapter 6, concluding remarks are given in Chapter 7 and future research areas are discussed in Chapter 8.

2.0 BACKGROUND

This Chapter is divided into two parts. The first part, which includes Sections 2.1 and 2.2, presents literature reviews on the nature of speech and speech enhancement techniques. The review on the nature of speech is used to describe parts of speech that constitute transient speech – the speech component that we are developing an algorithm to extract. As will be described, the aim of speech enhancement techniques is to improve the quality and/or intelligibility of speech. Conventional speech enhancement techniques try to improve the quality of speech by reducing the amount of noise in a noisy speech signal. An alternative approach to speech enhancement, which has not received as much attention, is to modify the speech signal itself before it is corrupted by noise to make it more intelligible in the presence of background noise. Both approaches are reviewed. The algorithm we are developing takes the latter approach.

The second part of Chapter 2, from Section 2.3 to 2.6, describes established tools and techniques that we will be using in the development and evaluation of our algorithm. These include the modified rhyme test, the short-time energy, the short-time zero-crossing rate, the short-time autocorrelation function, Mel-frequency cepstral coefficients and wavelet packets.

2.1 THE NATURE OF SPEECH

In speech production, the lungs begin the process by pushing air upwards [8]. The vocal folds may vibrate, causing the air that flows between them to vibrate. The vibration of the vocal folds is known as voicing, and speech sounds that are produced with the vocal folds vibrating are called voiced sounds. Sounds that are produced with no vibration of the vocal folds and turbulent airstream are called unvoiced sounds. The vibrating or turbulent air stream is then modified according to the vocal tract – the throat, mouth and nasal cavity. Movement of the tongue and lips produces a large number of modifications of the vibrating air stream and thus a wide variety of speech sounds.

Speech sounds can be classified as consonants or vowels. Vowels (which are voiced) are made with no major obstruction in the vocal tract so that the air passes through fairly easily. For a given speaker, vowels can be completely characterized by their formants [9]. Formants are resonant frequencies of the vocal tract and depend upon its shape and length. Consonants involve some type of obstruction or constriction in the vocal tract. In addition to being classified as voiced or voiceless, consonants can further be classified using the place of articulation, which describes where the obstruction occurs in the vocal tract, and the manner of articulation, which describes the nature (partial or total) of the obstruction.

Since vowels are produced during a vocal tract configuration that allows air to flow easily, vowels predominately include approximately constant frequency activity, which we will refer to as quasi-steady-state activity. Consonants and onset and offsets of vowel formants, which are produced during transitions of the vocal tract shape, are characterized by abrupt changes in frequency content. Vowels that are characterized by a shift in formant frequency may be considered as including transient activity, and consonants that include sustained segments,

called hubs, may be considered as including quasi-steady-state activity. In this study, we will refer to the collection of transitions into and out of vowel formants, the onset and offset of consonants, and rapidly frequency shifting vowel formants as transient speech.

2.2 SPEECH ENHANCEMENT

In a communication system where either the speaker or listener is in a noisy environment, or the transmission channel is noisy, the intelligibility and quality of speech may be severely degraded, making communication difficult if not impossible. This may also fatigue the communicators as the speaker may have to raise his/her voice while the listener may have to concentrate more. The aim of speech enhancement systems is to improve the quality and/or intelligibility of speech and to reduce communicator fatigue in order to facilitate better communication in noisy environments. Examples of applications where speech enhancement has provided substantial benefits include wireless communication, aircraft-to-control tower communication, within-aircraft communication, speech recognition, and speech coding [10]. There are two basic approaches to speech enhancement: noise reduction and speech modification.

Noise reduction techniques, such as spectral subtraction, optimum filtering, comb filtering, noise cancellation and subspace approaches, try to remove noise from a noisy speech signal with minimal impact on the speech itself. Most noise reduction techniques use an assumed model of the interfering signal (noise or competing speaker) in an attempt to reduce its effect. Although some of these speech enhancement techniques may provide improved speech intelligibility, their main focus is to provide improved speech quality. Consequently, evaluations of noise reduction techniques typically include objective measures, such as improvements in

signal-to-noise ratio and subjective assessments of speech quality, such as the mean opinion score (MOS). In a MOS test, subjects listen to speech samples and then rate on a scale of 1 to 5, where 1 = unsatisfactory, 2 = poor, 3 = fair, 4 = good and 5 = excellent [11]. Rarely have subjective measures of speech intelligibility been included in the evaluation of noise reduction techniques. Speech quality and intelligibility are different and should not be confused. Speech quality relates to how comfortable it is for a listener to listen to a speech utterance. The utterance does not necessarily have to convey meaning. Intelligibility relates to the ability of a speech utterance to convey meaning, that is, whether the listener can correctly identify words being spoken.

Speech modification techniques try to enhance features of speech that have been shown to be important for speech perception before speech is corrupted by noise. The goal is to produce modified speech with higher intelligibility in noisy environments. As the main focus is improved speech intelligibility, speech modification techniques have typically been evaluated using subjective measures of intelligibility like the modified rhyme test (MRT) [12] [13] [14]. Objective measures of intelligibility, like the articulation index (AI), have also been used to predict speech intelligibility [15] [16].

Section 2.2.1 provides an overview of noise reduction-based speech enhancement techniques that have been proposed over the years. To understand speech modification-based speech enhancement, an understanding of features of speech that are important for speech perception is necessary. Section 2.2.2 reviews studies of the importance of certain speech features for speech perception. Section 2.2.3 discusses speech enhancement by speech modification, including techniques that apply fixed filtering and techniques that identify and enhance specific speech cues, such as transient components.

2.2.1 Speech Enhancement By Noise Reduction

This Section reviews several noise reduction techniques, including spectral subtraction, Wiener filtering and the minimum mean-square error short-time spectral amplitude estimator, based in part on the review by Lim and Oppenheim [10].

Spectral subtraction is a noise reduction technique that tries to estimate the short-time spectrum of an additive noise that is corrupting a speech signal. The estimated short-time spectrum of noise is subtracted from the short-time spectrum of noisy speech to obtain an estimate of the short-time spectrum of original speech, and the estimated spectrum is combined with the phase of noisy speech to estimate the original speech. These operations can be viewed as an attempt to enhance the speech-to-noise ratio by attenuating the short-time spectrum when the speech-to-noise ratio is relatively low and not attenuating the short-time spectrum when the speech-to-noise ratio is relatively high. Various spectral subtraction methods differ on how the estimate of the short-time spectrum of the additive noise is obtained.

In the original spectral subtraction method, Weiss utilized the difference between the autocorrelation functions of voiced speech and noise to reduce noise in noisy speech [17] [10]. He described a method that includes a pseudo-cepstrum transform, which when used to transform noisy speech, moves most of the noise towards the origin. The noise is then removed by setting samples of the pseudo-cepstrum that are close to the origin to zero, and then reversing the transform. Boll [18] used the average value of the noise taken during nonspeech activity to estimate the noise spectrum. When the estimate of the short-time spectrum of noise was greater than the short-time spectrum of noisy speech, Boll set the estimate of the short-time spectrum of the clean speech to zero to avoid a negative value. Lim showed that Weiss's method is equivalent to that of Boll and is a generalization of spectral subtraction [10].

The spectral subtraction methods of Boll and Weiss result in a distortion called 'musical' noise. Musical noise can be described as ringing, warbling, or introducing a tonal quality into speech [18]. To reduce the musical noise, Berouti *et al.* reformulated spectral subtraction by multiplying the estimate of the short-time spectrum of noise by a factor α (greater than 1) before subtracting it from the short-time spectrum of the noisy speech [19]. Additionally, instead of setting the estimate of the short-time spectrum of clean speech to zero when the estimate of the short-time spectrum of noise was greater than the short-time spectrum of noisy speech, Berouti *et al.* set it to a non-zero value. Using $\alpha > 1$ results in an over-estimate of the average noise spectrum, which in addition to reducing the musical noise, further reduces the background noise. α was adaptively varied from frame to frame as a function of the frame speech-to-noise ratio.

Recently, Hu and Yu proposed an adaptive method for estimating the short-time noise spectrum that further reduces the 'musical' noise in the enhanced speech [20]. In their method, a weighted sum of the average value of the noise taken during nonspeech activity and the ratio of the noisy speech to the average value of the noise was used to obtain an estimate of the short-time spectrum of the noise.

Another noise reduction technique that has been widely used in speech enhancement is Wiener filtering. The Wiener filtering problem is to design a filter to recover a signal $d(n)$ from noisy observations $x(n) = d(n) + v(n)$, where $v(n)$ is the noise signal. Assuming that both $d(n)$ and $v(n)$ are wide sense stationary and uncorrelated, Wiener considered the problem of obtaining coefficients for a filter that produces the minimum mean-squared error estimate of $d(n)$ [21].

A Wiener filter may be used for noise reduction and noise cancellation. In noise reduction, a signal $d(n)$ is estimated from a noise-corrupted observation $x(n) = d(n) + v(n)$. To

obtain the coefficients of the Wiener filter, the autocorrelation of the noise must be determined from known properties or actual measurements of the noise signal.

As in the noise reduction problem, the goal of noise cancellation is to estimate a signal $d(n)$ from the noise-corrupted observation $x(n)=d(n)+v(n)$. However, unlike noise reduction which requires the autocorrelation function of the noise, noise cancellation uses the autocorrelation of a secondary signal $v'(n)$ that is highly correlated with $v(n)$ but not correlated with $d(n)$. The secondary signal $v'(n)$ may be obtained by placing a sensor (microphone in the case of speech) in the noise field near the signal source. An estimate of the speech signal, $\hat{d}(n)$, is obtained by first obtaining an estimate of the noise signal $v(n)$ from the secondary signal $v'(n)$ using a Wiener filter and then subtracting $v(n)$ from $x(n)$ [21].

Related to Wiener filtering and spectral subtraction is an enhancement technique proposed by Ephraim and Malah [22]. This method uses a minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimator, which is a function of the clean speech-to-noise ratio, referred to as the *a priori* SNR in their study. The error of this estimator depends on the estimate of the *a priori* SNR given the noisy speech signal. When applied to speech enhancement, the MMSE STSA estimator results in colorless residual noise, and thus a perceived higher quality and intelligibility of the enhanced speech.

Ding *et al.* proposed a speech enhancement method that is related to spectral subtraction and Ephraim and Malah's MMSE STSA estimator [23]. In their method, the magnitude squared spectra of speech are modeled as the exponential distribution and the estimation is performed in the power spectral domain under the MMSE criterion. Ding *et al.* compared the ability of their speech enhancement method to estimate the spectra of clean speech to that of spectral subtraction and to the method of Ephraim and Malah using a spectral distortion measure and found that their

method outperformed the latter two when the additive noise was white noise or traffic noise at SNRs of 0 to 20 dB.

Spectral subtraction and Wiener filtering can provide substantial improvements in speech quality. However, their performance tends to diminish as noise levels approach and fall below 0 dB. Evans *et al.* showed that the performance of spectral subtraction, as a pre-processor in automatic speech recognition, fell from 97 % to 17 % word recognition as the noise level of speech increased from clean speech to an SNR of -5 dB [24].

Also, the primary focus of noise reduction-based speech enhancement techniques is improved speech quality and not improved speech intelligibility. In fact, these techniques rarely improve speech intelligibility as was shown by Hu and Loizou [25]. Hu and Loizou evaluated the intelligibility enhancement of eight speech enhancement techniques: the generalized Karhunen-Loeve Transform (KLT) approach [26], the perceptual KLT approach [27], the Log Minimum Mean Square Error (logMMSE) algorithm [28], the logMMSE algorithm with speech presence uncertainty [29], the spectral subtraction algorithm based on reduced delay convolution [30], the multiband spectral subtraction algorithm [31], the Wiener filtering algorithm based on wavelet-thresholded multitaper spectra [32], and the Wiener algorithm based on a-priori SNR estimation [33]. They corrupted clean speech with four different types of noise: babble, car, street and train, processed the noisy speech using the speech enhancement techniques and then measured the intelligibility of the processed speech using formal listening tests. They found that most of the speech enhancement techniques were only able to maintain speech intelligibility at the same level as that of noisy speech, but not improve it.

2.2.2 Speech Perception and Transient Speech

As mentioned earlier, an alternative approach to speech enhancement is to modify the speech signal itself before it is corrupted by noise to make it more intelligible in the presence of background noise. Efforts in the speech community to enhance speech in this manner are reviewed after the discussion on the acoustical cues that influence speech perception.

The first task to study perception of speech is to find the cues - the physical stimuli - that control perception [34]. Since the invention of the spectrogram at AT&T Bell Laboratories, hundreds of articles on acoustical cues that influence the perceived phoneme have been published. A few of these articles that influenced the current study are reviewed here.

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 places of articulation [35]. Joos also noted that formant transitions are characteristically different for various stop consonant-vowel syllables [36]. Liberman characterized the formant transitions between stop consonant-vowel syllables and concluded that (1) the second formant transition can be an important cue for distinguishing place of articulation 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 [37]. 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, ŋ/. Characteristics of the spectrum during the release of the consonant as well as formant transitions between consonants and vowels are important cues for identifying the place of articulation [38].

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 [39]. This has led to fewer studies on the effect of third formant transitions on perception. However, a study by Liberman found that, when frequencies of the first and second formants and the transitions 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 consonants /b, d, g/ [34].

These studies relate the place of articulation of stop consonants to the patterns in transitions of formants observed on spectrograms. It was noted, however, that spectrographic patterns for a particular phoneme typically look 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/ [40]. The most important cues are sometimes among the least prominent parts of the acoustic signal [34]. The studies cited above also accentuate the importance of formant transitions as acoustic cues for identifying and distinguishing some phonemes. Although these studies were conducted in noise-free environments, we expect the same acoustic cues to be important for identifying and differentiating phonemes in noisy environments.

2.2.3 Speech Enhancement by Speech Modification

Modified speech that emphasizes speech transitions can be created by applying time-, frequency- or time-frequency-domain processes to original speech. Thomas and Niederjohn processed speech by highpass filtering with a cutoff of 1100 Hz and an asymptotic attenuation slope of +12 dB/octave followed by infinite amplitude clipping [41] [42]. The clipper output waveform was strictly binary and its axis crossings represented those of the filtered signal in timing and polarity

[42]. Psycho-acoustic testing showed that their filtered/clipped speech was more intelligible in band-limited (frequency range of 250 to 6800 Hz) white noise than unmodified speech when both speech signals are presented at the same SNR. The filter cutoff frequency and asymptotic attenuation slope were determined by psycho-acoustic testing of a range of values. In a related study, Thomas and Ohley showed that highpass filtering alone improves the intelligibility of speech over unmodified speech when both are presented in band-limited (frequency range of 250 to 6800 Hz) white noise at the same SNR [43]. The filter cutoff frequency and asymptotic attenuation slope, determined after psycho-acoustic testing of a range of values, were 1500 Hz and +18 dB/octave. Later, Niederjohn and Grotelueschen processed speech by highpass filtering followed by amplitude compression to maintain the output signal at a constant amplitude independent of the input amplitude [44]. Again the parameters for the highpass filter (cutoff frequency of 2000 Hz and asymptotic attenuation slope of +6 dB/octave) were optimized using psychoacoustic testing. The filtered/amplitude-compressed speech was more intelligible than both filtered/clipped speech and unmodified speech when all were presented in band-limited white noise at the same SNR. Niederjohn and Grotelueschen suggested that filtered/amplitude-compressed speech was more intelligible than filtered/clipped speech because clipping adds distortion. Niederjohn and Grotelueschen also evaluated the intelligibility of filtered/amplitude-compressed speech in noise recorded at a power generating plant and observed intelligibility improvements over original speech [45].

Thomas and Ravindran showed that the speech modification method of Thomas and Niederjohn [42] can improve the intelligibility of speech in which noise is added prior to filtering and clipping [46].

Villchur proposed a system for multiband speech amplitude compression [47]. He argued that without multiband compression, only amplitude ratios between successive speech elements can be changed and not that between elements that occur simultaneously. Villchur's system split a speech signal into low and high frequency channels, amplitude-compressed and then equalized the channel signals before combining them. Equalization was performed to ensure that the compressed speech signal was above the threshold of hearing. Psychoacoustic evaluations in band-limited random noise showed that speech modified by amplitude compression and equalization was more intelligible than unmodified speech. Ramasubramanian *et al.* implemented Villchur's 2-channel amplitude compression scheme using the discrete wavelet transform and achieved marginal improvements in intelligibility over Villchur's method [48].

Harris and Skowronski proposed an algorithm that moves energy from spectrally stationary regions to spectrally transitional regions of speech. The algorithm, which they termed energy redistribution spectral transition (ERST), used a normalized energy difference between adjacent frames of windowed speech to compute a spectral transition measure. Psycho-acoustic testing showed that speech processed using ERST was more intelligible than original speech [49].

Skowronski and Harris also proposed an algorithm that increases the energy of consonants relative to the energy of adjacent vowels. The algorithm, which they termed energy redistribution voiced/unvoiced (ERVU), used voicing detection based on a spectral flatness measure to discriminate between consonants and vowels. Psycho-acoustic testing showed that speech modified using ERVU was more intelligible than original speech [5]. The ERVU algorithm was later implemented to run in real-time by Raghavan *et al.* [6].

Chanda and Park described and implemented in real-time a method that applies a tunable bandpass filter to enhance consonants relative to vowels [50]. Their method highpass filtered consonants at a higher cut-off frequency than vowels, and then enhanced their intensity. Simulation showed that speech modified with their method had higher speech intelligibility index (SII) scores, especially for male speakers. SII is a measure that is highly correlated with speech intelligibility that is a function of the speech-to-noise ratio in different frequency bands [51].

Jayan *et al.* described a method to automatically detect regions of speech characterized by spectral transitions (referred to as landmark regions in their paper), and enhanced these regions by intensity and time-scale modification, without increasing the overall speaking rate [7]. They located the landmark regions using the rate of variation in energy and centroid frequency in five non-overlapping frequency bands. Listening tests conducted using non-sense syllables showed improvements in speech recognition especially at high noise levels (SNR = -9 and -12 dB).

Gordon-Salant visually segmented consonant-vowel nonsense syllable stimuli into consonant and vowel portions and then investigated the effects of (1) increasing the consonant duration by 100 %, (2) increasing the consonant-vowel ratio by 10 dB and (3) a combination of (1) and (2) [52]. Through psycho-acoustic testing performed using a 12-talker speech babble as the background noise, she showed that modifying speech in this manner improves speech intelligibility and reduces consonant confusion. Increases in the consonant-vowel amplitude ratio produced better performance than the other speech modification methods. In a similar study, Hazan and Simpson identified consonantal regions and transitional regions at vowel onsets and offsets of vowel-consonant-vowel (VCV) nonsense syllable stimuli and sentence material and then evaluated various degrees of enhancements of these regions on speech intelligibility [53].

Filtering these amplified regions to make them more discriminable was also investigated. Different enhancement strategies were applied to consonant categories (plosives, fricatives and nasals) based on the knowledge of consonant confusions of these consonant categories. Psycho-acoustic testing in speech-weighted noise showed that speech processed by this method yielded statistically higher intelligibility scores. The improvement in speech intelligibility was higher when a combination of increase in the consonant intensity and increase in the intensity of transitional regions was applied than when just one of the two was applied. The improvements in intelligibility were higher for VCV material than sentence material.

Speech modified by amplifying transients may also help in language comprehension. Tallal, having shown that language difficulties of language-learning impaired (LLI) children may result from a deficit in processing rapidly changing sensory input, investigated the effect of training LLI children with speech modified by increasing the duration of the speech signal while preserving the spectral content and naturality and by enhancing transitional elements [54] [55]. Evaluation of the same group of LLI children showed significant improvements in speech discrimination and language comprehension after daily training using modified.

In these studies, highpass filtering removed most of the energy associated with the first formant and voicing and increased the relative energy of the second formant, suggesting that the second formant is more important to speech perception than the first formant, as suggested earlier by [56]. Amplitude clipping, amplitude compression, and increasing the consonant-vowel ratio emphasizes consonants and transitional regions, i.e. transient speech, re-enforcing the suggestion that transient speech provides important cues for the identification and discrimination of speech sounds.

Speech intelligibility may also be enhanced by direct identification of transient speech, followed by amplification of this transient speech. Yoo *et al.* high pass filtered speech at a cutoff of 700 Hz and then applied three time-varying band-pass filters, based on a formant tracking algorithm by Rao and Kumaresan, to track and remove a quasi-steady-state component of speech [1], [2], [57]. Highpass filtering removed most of the voicing energy and first formant energy. 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 envelope 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 envelope 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 quasi-steady-state (QSS) speech component, and a transient component was estimated by subtracting the quasi-steady-state component from the original speech signal. The quasi-steady-state component was considered to contain most of the steady-state information of the input speech signal and the transient 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 1. Yoo *et al.* modified speech by combining the amplified transient component with the original speech and adjusting the energy of the modified speech to be equal to that of the original speech. Psycho-acoustic testing results showed that the modified speech was more intelligible than original speech at low signal-to-noise ratios.

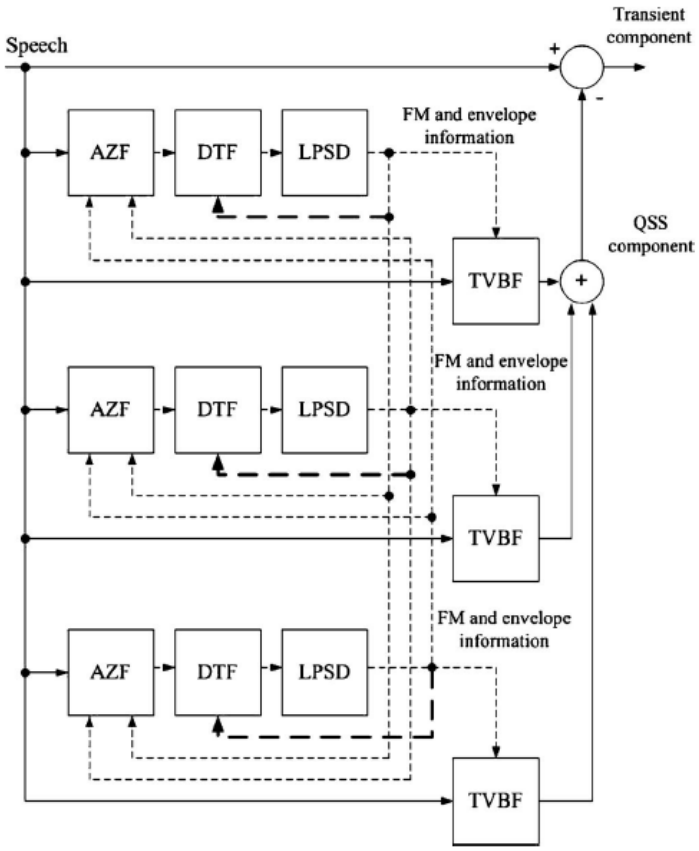

Figure 1: Block diagram of Yoo *et al.*'s speech signal decomposition method (From [2])

Another method that modified speech by identifying and emphasizing transient speech was proposed by Tantibundhit *et al.* [4]. Tantibundhit *et al.* expanded original speech using the modified cosine transform (MDCT) and then used a hidden Markov chain model to describe statistical dependencies of the MDCT coefficients and identify the most significant MDCT coefficients. Tonal speech was obtained as the inverse MDCT of the most significant coefficients. Tonal speech was subtracted from the original speech to obtain nontonal speech. Tantibundhit *et al.* expanded the nontonal speech using the discrete wavelet transform (DWT) and then used a hidden Markov tree to model statistical dependencies of the DWT coefficients and identify the most significant DWT coefficients. The inverse DWT of the most significant

DWT coefficients gave the transient speech signal. Tantibundhit *et al.* modified speech by combining the amplified transient component with the original speech and adjusting the energy of the modified speech to be equal to that of the original speech. Psycho-acoustic testing results showed that the modified speech was more intelligible than the original speech at severe signal-to-noise ratios.

The importance of speech transients as acoustical cues has also been used to improve the computational efficiency of automatic speech recognition systems. An example is a technique known as variable frame rate (VFR) analysis, where instead of using a fixed window step size when computing Mel-frequency cepstral coefficients (MFCC) or linear prediction coefficients (LPC) speech feature vectors for automatic speech recognition, the window step size is varied dynamically, using a small window step size when the speech signal is changing rapidly (retaining most of the frames) and a large window step size when the speech signal is changing slowly (discarding many frames). This reduces the computational load of speech recognizers without performance loss. The following are variants of variable frame rate analysis.

Ponting and Peeling proposed a VFR technique where the Euclidean distance between speech feature vectors of the current frame and the last retained frame was used in the frame picking decision [58]. 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 VFR technique of Ponting and Peeling by weighting the Euclidean distance between speech feature vectors with the log energy of the current frame [59]. 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. Le Cerf and Van Compernolle proposed a VFR method where the Euclidean norm of the first derivatives of MFCC feature vectors were used as the decision criteria for frame

picking [60] [61]. Their method discards a frame if the derivative measure of that frame is less than a threshold.

Brown and Algazi identified spectral transitions in speech using the Karhunen-Loeve transform and utilized them for sub-word segmentation and automatic speech recognition, improving recognition rates [62].

The two approaches to speech enhancement – noise reduction and speech modification – can be combined. Quatieri and Dunn described an adaptive Wiener filter whereby when the spectrum of noisy speech is changing rapidly, i.e. transient speech, little smoothing of the short-time spectrum is applied and when the spectrum is stationary, increased smoothing of the spectrum is applied [63]. Adapting the Wiener filter to transient speech helped to avoid the blurring of temporal fine structure in transient speech and resulted in enhanced speech that is of higher quality and intelligibility.

Speech may also be enhanced by emphasizing vowel activity. Cheng and O'Shaughnessy emphasized spectral peaks, which they suggested represented vowel activity, and de-emphasized spectral valleys of noisy speech improving the quality of speech [64].

2.3 EVALUATION OF SPEECH INTELLIGIBILITY

Speech intelligibility enhancement provided by speech enhancement methods can be evaluated using psycho-acoustic procedures such as the modified rhyme test [13] [14]. Generally, in a psycho-acoustic test, stimulus sounds are delivered to the ear of a subject, and behavioral responses elicited by these sounds are measured. The modified rhyme test (MRT), which was used for the psycho-acoustic experiments, is described below. The MRT is an attractive test

because it has been shown that repeated exposure to its material does not affect the levels of the responses, it can be automated and it is sensitive to consonant sounds.

The modified rhyme test (MRT), as its name suggests, is a modification of the rhyme test that was originally formulated by Fairbanks [65]. The MRT was proposed by House *et al.*, who used it to evaluate the ability of voice communication systems to transmit intelligible speech [12] [13]. The modified rhyme test draws stimuli from 50 sets of 6 rhyming monosyllable words, mostly of the form consonant-vowel-consonant (CVC), although a few CV and VC words are included. The 50 sets of rhyming words include 25 words in which the initial consonant is constant (e.g. pus-pup-pun-puff-puck-pub) and 25 words in which the final consonant is constant (e.g. lick-pick-tick-wick-sick-kick). The subject was provided with a response form showing 50 sets of 6 alternatives from which he/she was required to select his/her identification of the message. Fifty words, one from each set of rhyming words, were presented at different speech-to-noise ratios in the order shown in the subject's response form and the subject was instructed to draw a line through the item heard. At the end of the test, mean percentage correct scores at each speech-to-noise were calculated.

The modified rhyme test of House *et al.* is a closed-set test in that the response alternatives are available to the subject and the task of the subject is to identify the word heard from the set of possible answers. Mackersie *et al.* designed a word-monitoring modified rhyme test procedure using the word list of House *et al.* [14]. In a word-monitoring test, subjects are asked to listen to lists of words and to indicate when a target word has been heard. In Mackersie's test, which was completely computerized, a target word appeared on the computer monitor at the beginning of each trial and remained displayed as each of the six alternatives for the test item was presented auditorily. The subjects were required to push a button as soon as they heard the

target word. The subjects heard each of the six alternatives only once and did not have a second chance to hear the words. Also, if a subject did not indicate that any of the six words presented was the target word seen on the screen, the next target word appeared. As soon as the button was pushed, the trial was terminated and the response time, measured as the amount of time that elapsed between the end of the stimulus presentation and the subject response, was saved. The stimulus words were presented at six different SNRs with the SNR randomly selected for each trial. At the end of the test, percent correct score were calculated.

The modified rhyme test has the advantage that they do not require prior training of subjects and have minimal practice effects. House *et al.* showed that repeated testing of the same subjects results in similar percent correct scores [12] [13]. Mackersie's modified rhyme test is used in this study. A detailed description of this test is given in Section 5.1.

2.4 TIME-DOMAIN SPEECH PROCESSING METHODS

A major goal of much research on speech processing methods is to obtain a more convenient or more useful representation of the information carried by the speech signal. Time-domain speech processing methods allow for the extraction of features as a function of time. Examples of features that can be extracted in the time-domain are energy, zero-crossing rate and the autocorrelation function.

Over a short time segment, properties of the speech signal may be assumed to change relatively slowly with time. This assumption leads to a variety of short-time processing methods in which short segments of the speech signal are isolated and processed as if they were sustained sound with fixed properties. The processing of these short segments is often done at fixed

intervals with the segments, sometimes called analysis frames, overlapping. The results of the processing of each segment may be either a single number or a set of numbers.

This section, based on [66], discusses the short-time energy, the short-time average zero-crossing and the short-time autocorrelation function – time-domain speech processing methods used in this study. The application of these methods to voiced/unvoiced detection is also discussed. Voiced/unvoiced detection is used in phoneme identification and speech recognition to identify phones (speech sounds). In speech intelligibility enhancement, different segments of speech may be processed differently depending on whether they are voiced or not.

2.4.1 Short-time Energy

The amplitude of the speech signal is generally much lower for unvoiced segments than for voiced segments. The short-time energy of the speech signal provides a convenient representation that reflects these amplitude variations. In general, the short-time energy is defined as [66, pp. 120],

$$E_n = \sum_{m=-\infty}^{\infty} \left[x(m)w(n-m)\right]^2 \tag{1}$$

where $w(n)$ is the window sequence. This can be written as

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m) \tag{2}$$

where $h(n) = w^2(n)$. The signal power is filtered by a linear filter with impulse response $h(n) = w^2(n)$.

The choice of the filter or window function determines the nature of the short-time energy representation. If the window duration is too small, i.e. on the order of a pitch period or less (2 ms for a high pitch female or a child, 25 ms for a very low pitch male) [66], E_n will fluctuate very rapidly depending on the exact details of the waveform. If the window duration is too long, i.e. on the order of tens of pitch periods, E_n will change very slowly and thus will not reflect the changing properties of the speech signal. Clearly there is no single value for the window duration that is entirely satisfactory, especially considering the differences in pitch period between female/child and male. However a suitable practical choice is a window duration of 10-30 ms.

Another consideration in the selection of a window function is the type of window. For example, although both a rectangular window and a Hamming window are lowpass linear filters, a Hamming window produces more attenuation in the high frequencies and thus a smoother short-time energy representation than a rectangular window of the same duration. A rectangular window may also produce edge effects since it is not tapered at the edges.

2.4.2 Short-time Average Zero-Crossing Rate Function

A zero-crossing is said to have occurred if successive samples of a signal have different algebraic signs. For a sinusoidal signal the zero-crossing rate is proportional to the frequency of the sinusoid. The rate at which zero-crossings occur for a non-sinusoidal signal like speech is a simple measure of the frequency content. The zero-crossing rate is defined as [66, pp. 128],

$$Z_n = \sum_{m=-\infty}^{\infty} \left| \text{sgn}\left[x(m)\right] - \text{sgn}\left[x(m-1)\right] \right| w(n-m) \qquad (3)$$

where $\text{sgn}\left[x(n)\right] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases}$ and $w(n) = \begin{cases} \frac{1}{2N} & 0 \leq n \leq N-1 \\ 0 & otherwise \end{cases}$ with N the number of samples of the window function. All that is required to compute Z_n is to check samples in pairs to determine whether a zero-crossing has occurred and then compute the average over N consecutive samples.

Generally, voiced speech has most of its energy in the low frequencies (below 3 kHz) and unvoiced speech has most of the energy in the high frequencies. This implies that the short-time average zero-crossing rate, which is high for high frequency signals and low for low frequency signals, may be used to determine voicing. However, the determination of voicing using the short-time average zero-crossing rate is imprecise as there is no average zero-crossing rate to discriminate perfectly between voiced and unvoiced speech.

A more reliable determination of voicing uses a combination of the short-time average zero-crossing rate and the short-time energy. A combination of low short-time average zero-crossing rate and high short-time energy corresponds to voiced speech and a combination of high short-time average zero-crossing rate and low short-time energy corresponds to unvoiced speech. A combination of high short-time average zero-crossing rate and high short-time energy rarely occurs and is unclassified, while a combination of low short-time average zero-crossing rate and low short-time energy corresponds to silence. The correspondence of short-time average zero-crossing rate and short-time energy to voicing is summarized in Figure 2.

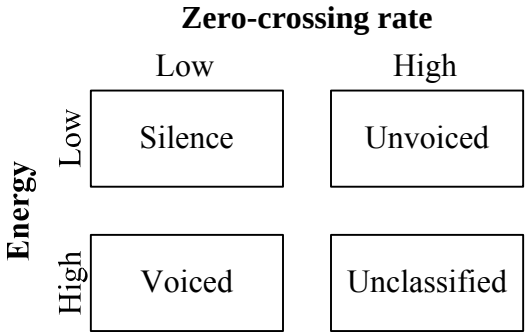

Figure 2: Use of the short-time average zero-crossing rate and the short-time energy to determine voicing.

Although there is no single threshold pair for the short-time average zero-crossing rate and the short-time energy for determining voicing for all speakers, use of two time-domain processing methods instead of a single method produces more reliable results.

2.4.3 Short-time Autocorrelation Function

The autocorrelation function of a deterministic signal $x(m)$ is defined as

$$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k) \tag{4}$$

The autocorrelation function representation of the signal is a convenient way to display certain properties of the signal. For example, if the signal is periodic with period P samples, then the autocorrelation function is also periodic with the same period, i.e. $\phi(k) = \phi(k+P)$. Also $\phi(0)$ is equal to the energy of the signal.

The short-time autocorrelation function is defined as [66, pp. 141],

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)x(m+k)w(n-k-m) \tag{5}$$

It can be shown that $R_n(k) = R_n(-k)$. Define $h_k(n) = w(n)w(n+k)$, then using $R_n(k) = R_n(-k)$, the short-time autocorrelation function can be written as

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)x(m-k)h_k(n-m) \tag{6}$$

That is, the value at time n of the k^{th} autocorrelation "lag" is obtained by filtering the sequence $x(n)x(n-k)$ with a filter with impulse response $h_k(n)$.

The short-time autocorrelation function is different from the true autocorrelation function. If $x(n)$ is periodic, the true autocorrelation is periodic. However $R_n(k)$ is not periodic but displays large peaks with decaying amplitude at the period of $x(n)$. The reduction in the amplitude of the peaks of $R_n(k)$ as k increases is due to having less data involved in the computation as k increases.

Although voiced speech is not truly periodic, its short-time autocorrelation exhibits large peaks located approximately at multiples of the "period" of the speech signal. Parts (a) and (b) of Figure 3 illustrates the near periodicity of the short-time autocorrelation functions of voiced speech. The autocorrelation functions, which are for two voiced segments from the same speech utterance, were computed using a rectangular window with a duration of 40 ms. The reduction in

the amplitude of the peaks of $R_n(k)$ as k increases is clearly visible. The short-time autocorrelation of unvoiced speech, also shown in Figure 3, has no periodicity peaks and looks like high frequency noise. Consequently voicing/unvoicing may be determined for a segment of speech by evaluating whether $R_n(k)$ for that segment exhibits periodic peaks or looks like high frequency noise. Additionally, the exact position of the peaks of $R_n(k)$ may be used to determine the pitch of the speech signal.

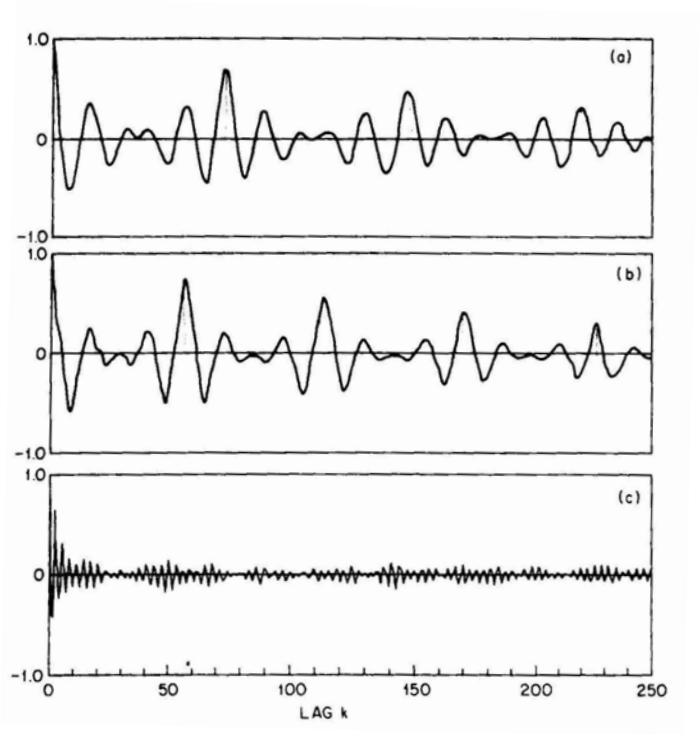

Figure 3: Short-time autocorrelation functions of voiced speech parts (a) and (b) and of unvoiced speech (c). The autocorrelation functions were computed using a rectangular window with a duration of 40 ms (From [66]).

To eliminate the attenuation and accentuate the peaks of the short-time autocorrelation function at the "period" of the signal, the modified short-time autocorrelation function may be used. The modified short-time autocorrelation function is defined as

$$\hat{R}_n(k) = \sum_{m=0}^{N} x(n+m)x(n+m+k) \quad 0 \le k \le K \tag{7}$$

$\hat{R}_n(k)$ is always computed over N samples, and samples from outside the interval n to $n+N-1$ are included in the computation.

An alternate procedure for emphasizing the peaks of the short-time autocorrelation function at multiples of the "period" of the speech signal is to apply a technique called center-clipping to the speech signal before computing the short-time autocorrelation function. In center-clipping, speech is passed through a non-linear transformation $y(n) = C[x(n)]$ where $C[\ ]$ is as shown in Figure 4. For speech samples above C_L, the clipping level, the output is the input minus the clipping level. For speech samples below C_L, the output is zero. Figure 5 compares the short-time autocorrelation function computed with no clipping and with clipping. Clearly the peaks at multiples of the "period" of the speech signal are easier to identify when center-clipping is used.

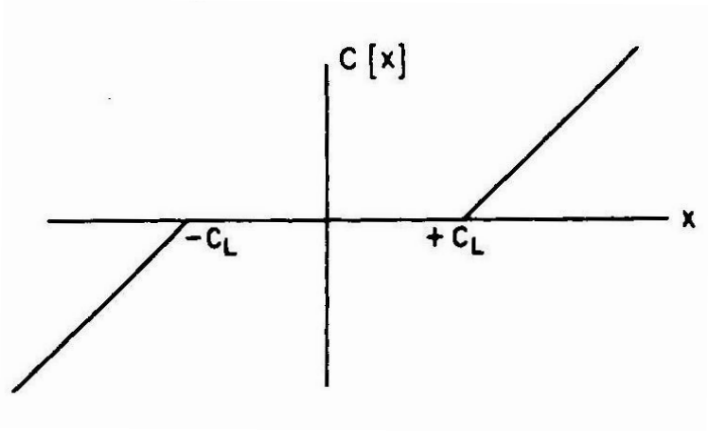

Figure 4: Center clipping (From [66]).

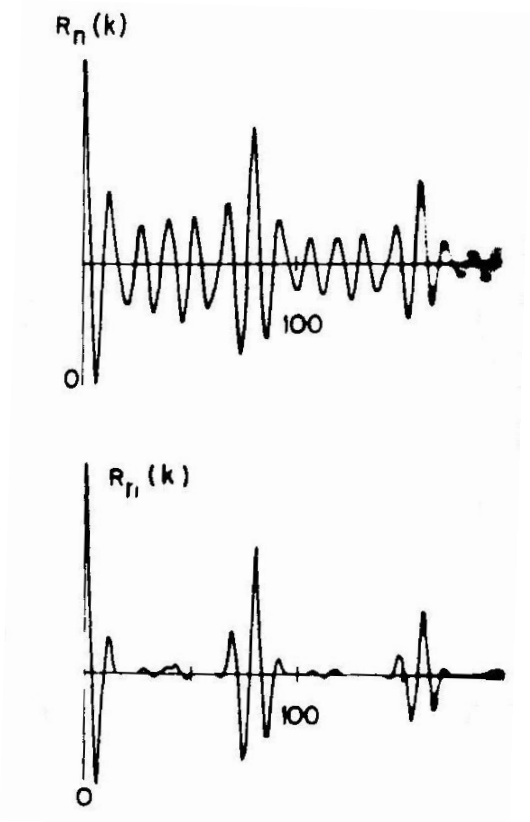

Figure 5: Short-time autocorrelation function computed with no clipping (top) and with clipping (bottom). The x-axis is the autocorrelation lag k (From [66]).

2.5 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 [67]. MFCC capture and preserve significant acoustic information better than linear prediction coefficients (LPC) [68]. MFCC have become the dominant features used for speech recognition and the following discussion of MFCC will follow the description of [69]. Our proposed transient speech extraction algorithm uses MFCC to compute a transitivity function that is used to characterize and emphasize transient activity in wavelet packet coefficients.

Figure 6 shows the process for creating MFCC features from a speech signal $s(n)$. The first step is to convert the speech into frames by applying a windowing function $w(n)$ of length M samples to obtain

$$s_i(m) = \sum_{m=0}^{M-1} s(m)\, w(i-m) \tag{8}$$

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. Subscript i indicates frame number.

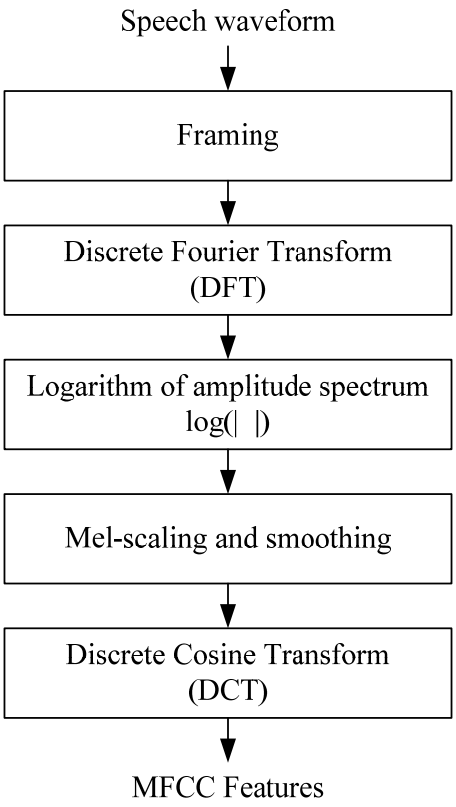

Figure 6: Process to create MFCC feature vectors from a speech waveform

The next step is to compute the discrete Fourier transform (DFT), $S_i\left(e^{j\omega}\right)$ for each frame i

$$S_i\left(e^{j\omega}\right) = \sum_{m=0}^{M-1} s(m) e^{-j\omega m} \tag{9}$$

Using the notation $S_i(\omega) = \left|S_i\left(e^{j\omega}\right)\right|^2$, the log spectrum is represented as

$$\log S_i(\omega) = \log\left|S_i\left(e^{j\omega}\right)\right|^2 \tag{10}$$

For a power spectrum that is periodic for a sampled data sequence and symmetric with respect to $\omega = 0$, the Fourier series representation of $\log S_i(\omega)$ can be expressed as [70]

$$\log S_i(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{-j\omega n} \tag{11}$$

where $c_n = c_{-n}$ are the real cepstral coefficients. The spectrum $S_i(\omega)$ discards the phase information but retains the amplitude information, which is regarded as the most important property for speech perception [69].

The Mel-scale is a scale that is based on a mapping between actual frequencies and pitch as perceived by the human auditory system. This scale is approximately linear up to 1000 Hz and logarithmic thereafter. In the next step of the computation, the Fourier spectrum $S_i(\omega)$ is Mel-scaled by warping the frequency using a filter bank where each filter's spacing and bandwidth is determined by a constant mel frequency interval. The spacing of this filter bank, with filters as shown in Figure 7, is approximately 150 mels and the width is 300 mels. Mel-scale filter banks, like critical bands, are arranged such that each frequency band contributes about equally to speech intelligibility, which emphasizes perceptually meaningful frequencies.

Denoting the log-energy output of the p^{th} filter as $\widetilde{S}_i^{\,p}$, $p = 1,2,...,P$, the Mel-frequency cepstral coefficients are computed as [70]

$$\tilde{c}_i(n) = \sum_{p=1}^{P} \left(\log \tilde{S}_i^p \right) \cos\left[n\left(k - \frac{1}{2} \right) \frac{\pi}{P} \right] \quad n = 1, 2, ..., L \tag{12}$$

where L is the desired length of the spectrum. Equation (12) includes the last step in Figure 6 – the computation of the discrete cosine transform (DCT). 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.

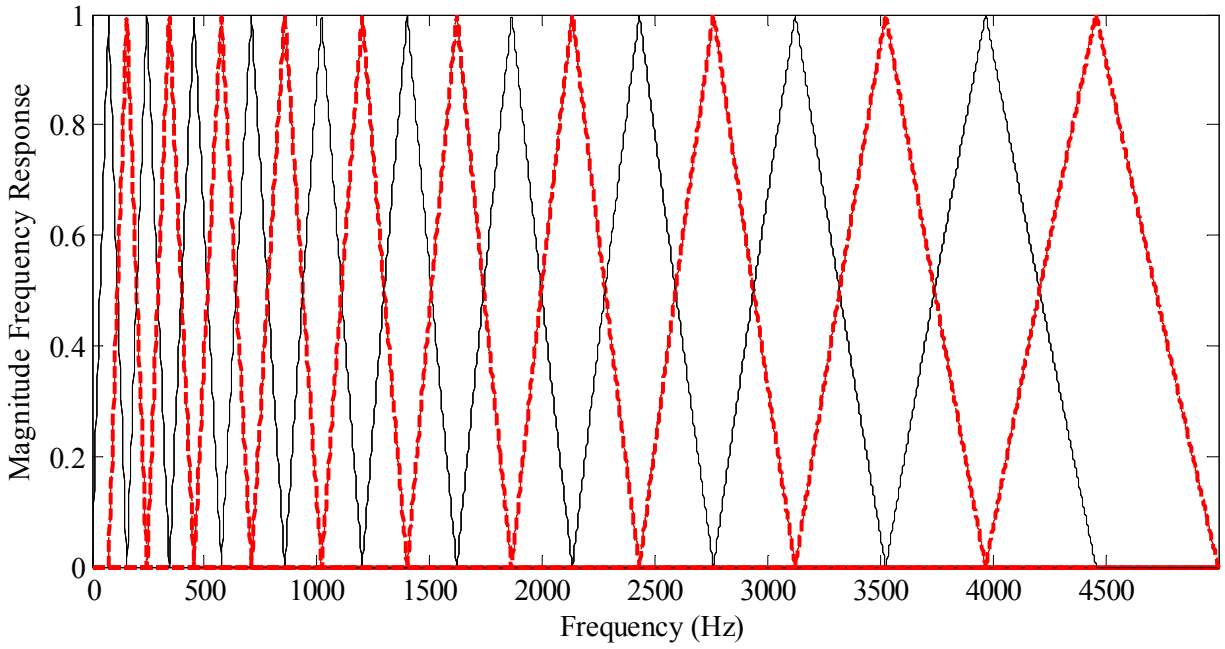

Figure 7: Mel-scaled filter bank. The spacing and bandwidth of each filter is determined by a constant mel frequency interval (spacing = 150 mels and bandwidth = 300 mels).

2.6 WAVELETS PACKETS

In subband signal processing, a signal is split into a number of subsignals. The subsignals may help emphasize specific aspects of the original signal or may be easier to work with than the original signal. The subsignals are sometimes called subband signals. The subband signals are often downsampled so that the data rates are the same in the subbands as in the original signal. The subband signals have to be sufficient to reconstruct the original signal. Wavelet transforms can be used for subband signal processing. This section describes wavelet packets – the subband signal processing method that was used in the transient speech extraction algorithm. For completeness, the continuous wavelet transform, multiresolution analysis, the discrete wavelet transform, signal decomposition and reconstruction using wavelets and factors considered in choosing a wavelet function are reviewed in Appendix A. The descriptions here and in Appendix A are based on [71] [72] [73] [74] [75] [76] [77] [78] [79] and [80].

The discrete wavelet transform (DWT) results in a logarithmic frequency resolution; high frequencies have wide bandwidths whereas low frequencies have narrow bandwidth [71]. The logarithmic frequency resolution of the DWT is not appropriate for some signals, and wavelet packets (WP) provide a method to segment the higher frequencies into narrower bands. This section discusses the 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 the binary tree) are also split by filtering and downsampling. The splitting of the low and high frequency spectra results in the full binary tree shown in Figure 8 and a completely evenly spaced

frequency resolution illustrated in Figure 9. In the DWT analysis, the high frequency band is not split into smaller bands.

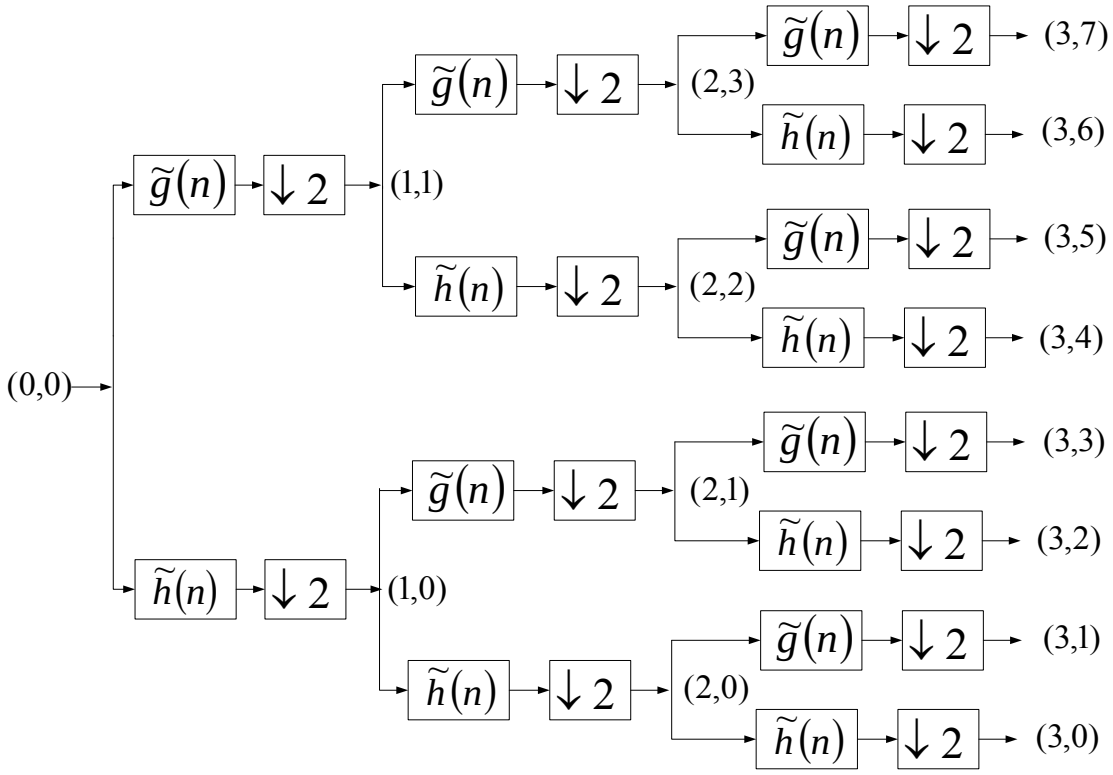

Figure 8: Three-stage full wavelet packet decomposition scheme

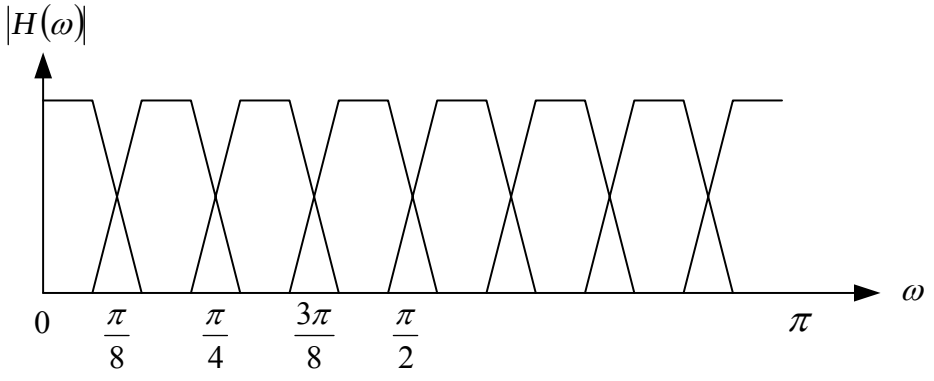

Figure 9: Ideal frequency response for the wavelet packet transform

In the structure of Figure 8, each subspace is indexed by its depth and the number of subspaces below it at the same depth. The original signal is designated depth zero.

The wavelet packet reconstruction scheme is achieved by upsampling, filtering with appropriate filters and adding coefficients. This scheme is shown in Figure 10. This WP reconstruction tree structure is labeled the same as the WP decomposition structure.

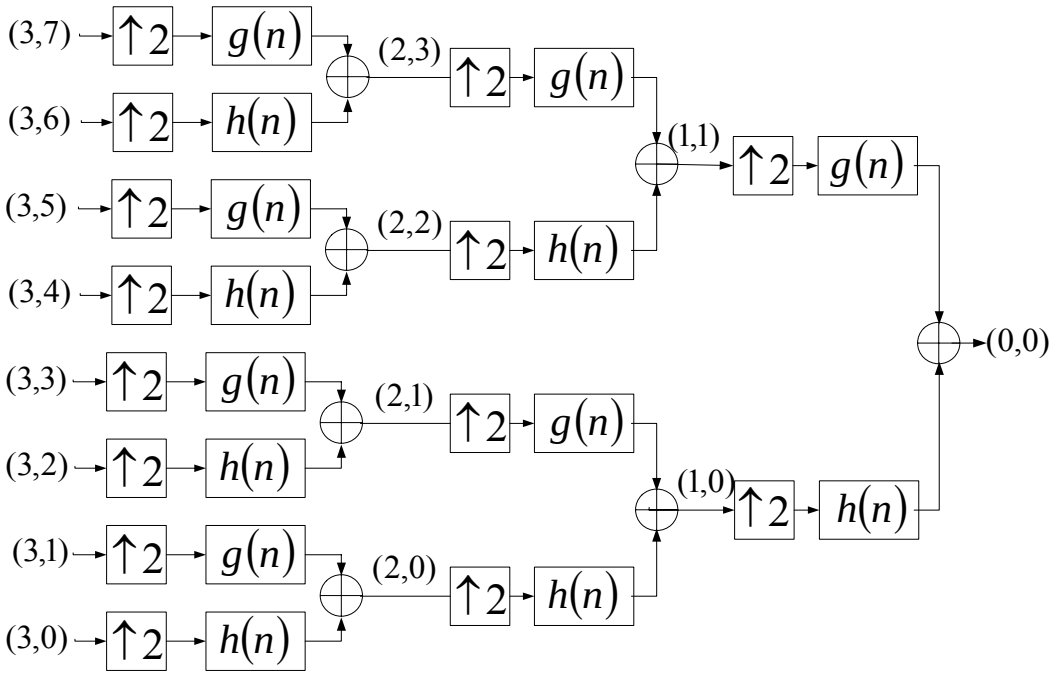

Figure 10: Three-stage full wavelet packet reconstruction scheme

The filters $\tilde{h}(n)$ and $h(n)$ are lowpass whereas filters $\tilde{g}(n)$ and $g(n)$ are highpass and satisfy the following properties [75] [79];

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. $|H(\omega = 0)| = 1|$ and $\tilde{h}(n) = O(n^{-2})$ at infinity, i.e. the asymptotic upper bound of $h(n)$ at infinity is n^{-2}.

$$|H(\omega)|^2 + |H(\omega + \pi)|^2 = 1$$

3.0 ALGORITHM FOR EXTRACTION OF TRANSIENT SPEECH

We have been investigating the use of wavelet packets for extraction and emphasis of transient speech for real-time speech intelligibility enhancement. The investigations involve the development of an algorithm that decomposes a speech signal into several sequences of wavelet coefficients using the forward wavelet packet transform, characterizes the rate of change and adjusts the wavelet coefficients based on how fast they are changing and synthesizes a transient speech signal using the inverse wavelet packet transform. Transient speech is used to create modified speech by amplifying and adding it to the original speech and then adjusting the energy of the modified speech signal so that it equals that of original speech. For the characterization of the rate of change of wavelet coefficients, a function that we called the transitivity function was developed. This function is large and positive when the wavelet coefficients have a rapidly changing frequency or amplitude and near zero when the wavelet coefficients are in steady-state.

Wavelet packets are attractive for our application because they provide subband decomposition, which allows the detection of transients occurring at different times in different frequency bands, and can be implemented in real-time.

Two alternate definitions for the transitivity function, one based on the short-time energy (STE) of wavelet packet coefficients and the other on Mel-frequency cepstral coefficients (MFCC) of wavelet packet coefficients, were formulated. The STE-based approach was initially given more attention and evaluated experimentally because this approach de-emphasized quasi-

steady-state activity more than the other approach and our initial goal was to maximize transient activity. The two methods for computing the transitivity function are both reasonable ways to detect transient speech. However, in informal listening tests, transient speech extracted using the MFCC-transitivity function had less 'garbling' artifact noise, a better speech quality and was more intelligible than transient speech extracted using the STE-transitivity function. Informal listening tests were conducted by listening to original, transient and modified speech of isolated words and sentences in noise at various SNRs and making judgments on their intelligibility.

In this chapter, STE and MFCC transitivity functions are defined. The transient extraction method, which can utilize either transitivity function to characterize the rate of change of the wavelet coefficients, is described. The creation of modified speech by combining amplified transient speech with original speech and adjusting the energy of modified speech so that its energy equals that of original speech is described. The incorporation of an unvoiced speech booster, which automatically detects and amplifies unvoiced speech segments, to the transient extraction algorithm is also described.

Demonstrations of the transitivity function and transient signals are presented using a synthetic signal.

3.1 THE TRANSITIVITY FUNCTION

As mentioned earlier two methods to compute a transitivity function were formulated; STE-transitivity function and MFCC-transitivity function. These functions are described here.

3.1.1 The Short-time Energy Transitivity Function

The short-time energy (STE) transitivity function of a sequence of wavelet packet coefficients $v_k[n]$, $0 \le k \le K-1$, for packet k of the decomposition of a speech signal can be computed as shown in Figure 11. $K = 2^L$ is the number of packets in the decomposition and L is the decomposition level. The short-time energy (STE) of $v_k[n]$ is computed using a window function $w[n]$ of length M as

$$E_{k,i} = \sum_{m=0}^{M-1} \left(v_k[m] w[i-m] \right)^2 \tag{13}$$

A Hamming window with M = 276 samples (25 ms at sampling frequency of 11025 Hz) and a window step size of 55 samples (5 ms) were used. A smoothed first derivative of the logarithm of the short-time energy is defined as the transitivity function and is computed as

$$f_{k,i} = \sum_{l=-4}^{4} a_l \log\left(E_{k,n} \right) \tag{14}$$

where the coefficients a_l are given by $a_l = \frac{-l}{60}$ [81]. The logarithm of the short-time energy emphasizes low energy regions compared to high energy regions. The subscripts denote that $f_{k,i}$ is the value of the transitivity function for the time interval included in the i^{th} window segment of the k^{th} wavelet packet. For a given frame, the transitivity function is large and positive when

the wavelet coefficients $v_k[n]$ of that frame have a rapidly changing amplitude. The transitivity function that is computed using the short-time energy will be referred to as STE-transitivity function.

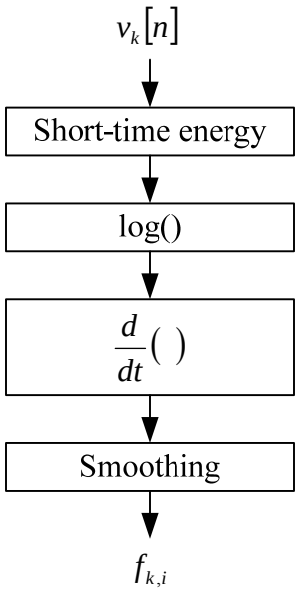

Figure 11: Computation of the transitivity function using short-time energy.

3.1.2 The Mel-Frequency Cepstral Coefficients-Based Transitivity Function

To reduce the computational load of automatic speech recognizers, Le Cerf and Van Compernolle proposed a variable frame rate method whereby the Euclidean norm of the first derivatives of Mel-frequency cepstral coefficients (MFCC) feature vectors was used as a decision criterion for frame picking [60] [61]. Frames whose value of this function was higher than a threshold were considered more relevant to speech perception as they included transient regions and were retained.

We expand their idea and apply it to formulate a method for computing the transitivity function of a sequence of wavelet coefficients $v_k[n],\ 0 \le k \le K-1$ as shown in Figure 12. $K = 2^L$ is the number of packets in the decomposition and L is the decomposition level.

First, 12 MFCC are computed using a 25 ms. Hamming window function and a window step size of 5 ms. The MFCC of each window segment of the wavelet coefficients are computed as

$$\tilde{c}_{k,i}(n) = \sum_{p=1}^{P} \left(\log \tilde{V}_{k,i}^{p}(\omega)\right) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{P}\right] \quad n = 1, 2, ..., L = 12 \qquad (15)$$

where $\tilde{V}_{k,i}^{p}(\omega),\ p = 1,2,...,P$ is the frequency warped spectrum of $v_{k,i}[n]$ obtained from the magnitude-squared spectrum $V_{k,i}(\omega) = \left|V_{k,i}\left(e^{j\omega}\right)\right|^2 = \left|\sum_{n=0}^{N-1} v_{k,i}(n) e^{-j\omega n}\right|^2$ of $v_{k,i}[n]$ by filtering $V_{k,i}(\omega)$ using mel-scaled filter banks with P filters as described in Section 2.5. $v_{k,i}[n] = \sum_{m=0}^{M-1} v_k[m] w[i-m]$ is the i^{th} frame of $v_k[n]$ obtained by windowing with window function $w[n]$. $L =$ 12 MFCC were used because $\tilde{c}_{k,i}(n) \approx 0$ when $L >$ 12. The first derivatives of the MFCC coefficients $\tilde{c}_{k,i}(n)$ are computed and smoothed to obtain

$$d_{k,i}(n) = \sum_{l=-4}^{4} a_l \tilde{c}_{k,i,n-l} \qquad (16)$$

where the coefficients a_l are given by $a_l = \frac{-l}{60}$ [81]. The Euclidean norm of the derivatives

$$f_{k,i} = \left\| d_{k,i}(n) \right\| = \sqrt{\sum_{n=1}^{12} d_{k,i}(n)} \tag{17}$$

is defined as the transitivity function $f_{k,i}$. The subscripts denote that $f_{k,i}$ is the value of the transitivity function for the time interval included in the i^{th} window segment of the k^{th} wavelet packet. For a given frame, the norm and hence the transitivity function is large and positive when the wavelet packet coefficients $v_k[n]$ have a rapidly changing amplitude or frequency.

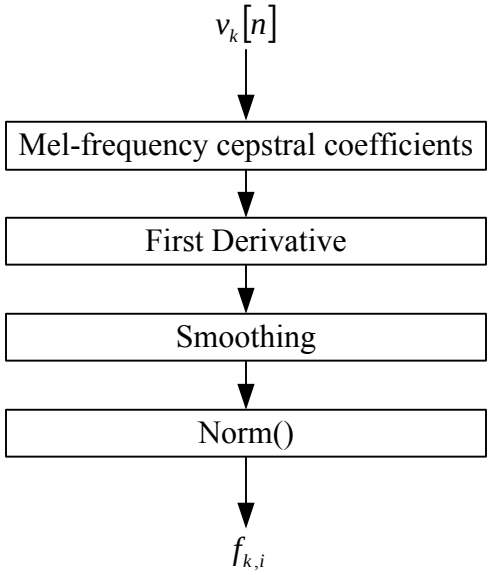

Figure 12: Computation of transitivity function using Mel-frequency cepstral coefficients.

3.2 ALGORITHM FOR EXTRACTION OF TRANSIENT SPEECH

A diagram of the method for extraction of transient speech is shown in Figure 13. The algorithm includes steps for pre-processing, wavelet decomposition, computation of transitivity functions and emphasis of speech transitions and wavelet reconstruction.

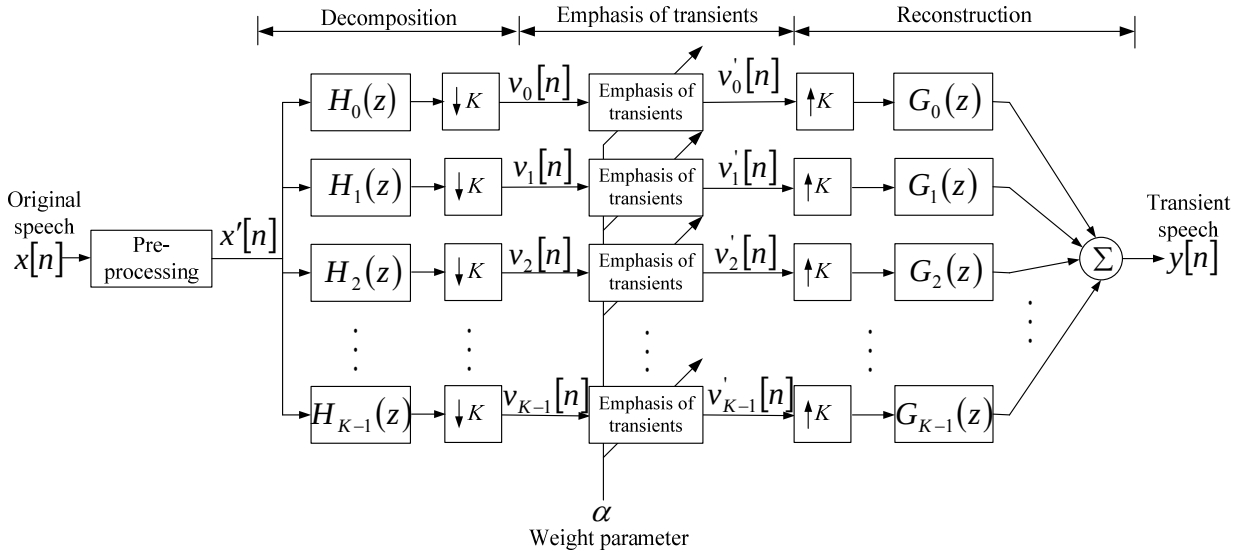

Figure 13: Transient speech extraction method.

3.2.1 Pre-processing

Pre-processing involves passing the speech signal $x[n]$ through a system that reduces the amount of energy of the first formant of speech. Without pre-processing, the transient speech signal obtained is dominated by low frequency transitions and does not contribute maximally to speech intelligibility enhancement [1]. Initially, pre-processing was performed using a 50^{th} order finite impulse response (FIR) highpass filter with a cutoff of 700 Hz. The magnitude frequency response of this filter, which we will refer to as HPF_0 hereafter, is shown in Figure 14. HPF_0

does not reduce the intelligibility of the speech signal, as was shown by [1], and highpass filtering alone can improve speech intelligibility [44] [5]. An experiment performed to select the best pre-processing filter is described in Chapter 5.

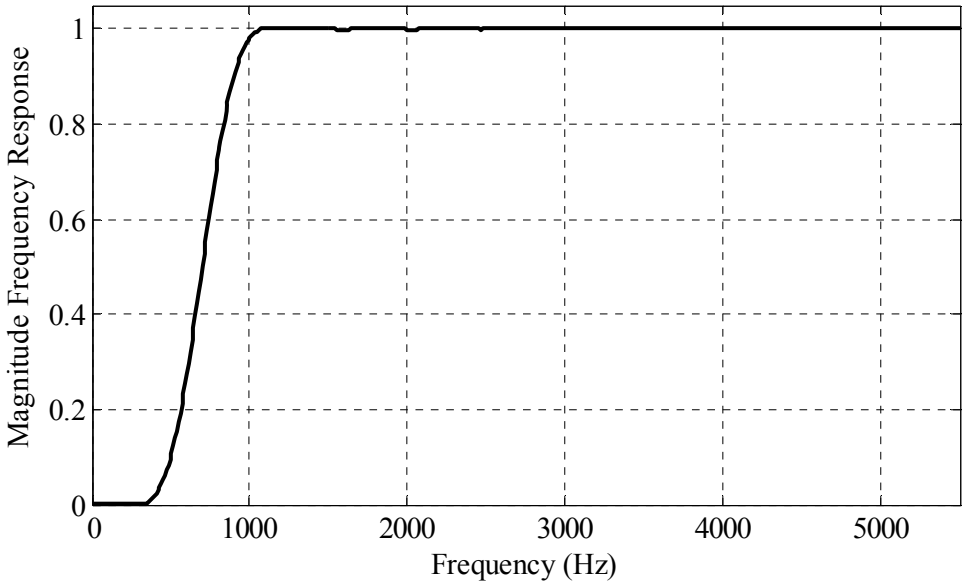

Figure 14: Magnitude frequency response of HPF_0 – a 50th order FIR filter with a cutoff frequency of 700 Hz. This filter was initially selected for pre-processing.

3.2.2 Wavelet Decomposition

In the wavelet decomposition stage, the pre-processed speech $x'[n]$ was decomposed at scale level 4 ($L = 4$) using the forward wavelet packet transform, resulting in $K = 2^L = 16$ packets $v_k[n],\ 0 \le k \le K-1$ each with a sequence of wavelet coefficients. Splitting the speech signal into $K = 16$ packets results in wavelet coefficients with good frequency resolution and with enough coefficients for a reliable computation of the transitivity function. The wavelet packet

transform utilized a Daubechies-18 wavelet function, which was found to have good frequency selectivity. Wavelet functions with shorter support size were less frequency selective while wavelet functions with longer support size increased the computation time of the algorithm. The wavelet and scaling functions and lowpass and highpass decomposition and reconstruction filters for the Daubechies-18 wavelet are described in Appendix B.

3.2.3 Emphasis of Speech Transients

The next stage is the emphasis of speech transients using the transitivity function. The steps involved in this stage are depicted in Figure 15. For each packet, a transitivity function $f_{k,i}$ that characterizes the rate of change of the wavelet coefficients $v_k[n]$ of that packet was computed. Subscript i indicates frame number. The transitivity function is used as a weighting function for the wavelet coefficients from which the transient speech signal is synthesized.

The transitivity function was computed using the two definitions, resulting in two distinct transient speech signals. The transitivity function produces one value per window segment of wavelet coefficients. To allow direct multiplication of the wavelet coefficients by the transitivity function, the transitivity function $f_{k,i}$ was linearly interpolated to have as many samples as there are wavelet coefficients in a packet, obtaining $f_k[n]$.

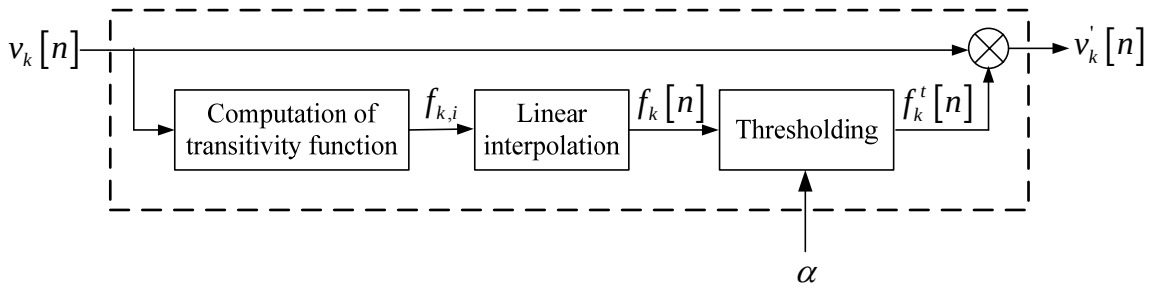

Figure 15: Emphasis of transients

Steady vowel segments of speech (quasi-steady-state components) have high energy compared to transition segments. Even if the relative rate of change during these regions is small, the value of the transitivity function may be significant compared to the values during speech transitions. To control the amount of quasi-steady-state energy that is included in the transient speech signal, a thresholding operation (labeled *Thresholding* in Figure 15) using a weight parameter α is applied to the transitivity function $f_k[n]$ producing $f_k^t[n]$. A sample of the computed transitivity function for packet k, $f_k[n]$ is set to zero when the value of the log short-time energy corresponding to that sample is greater than a packet threshold $T(k)$, chosen as:

$$T(k) = \log E(k)_{\min} + \alpha\left[\log E(k)_{\max} - \log E(k)_{\min}\right] \tag{18}$$

where $E(k)_{\min}$, $E(k)_{\max}$ are the minimum and maximum values of the short-time energy of the wavelet coefficients in packet k over the word or analysis interval, and $0<\alpha<1$ is the weight parameter. A sample of the transitivity function whose corresponding value of the log short-time energy is less than the packet threshold is left unchanged.

When $\alpha = 1.0$, $T(k) = \log E(k)_{\max}$ and no samples of the transitivity function are set to zero, i.e. $f_k^t[n] = f_k[n]$. In this case a large amount of quasi-steady-state activity is included in transient speech. When $\alpha = 0$, $T(k) = \log E(k)_{\min}$ and all samples of the transitivity function are set to zero, i.e. $f_k^t[n] = 0$. In this case, the entire speech signal is considered quasi-steady-state and excluded from the transient speech signal.

Equation (18) generates a packet specific threshold $T(k)$ that is a function of the short-time energy of the coefficients of packet k and the weight parameter α. A single threshold for all wavelet packets is not effective because the energy in different packets varies greatly. After thresholding, abrupt changes from non-zero-valued samples to zero-valued samples of the transitivity function are smoothed by replacing the seven zero-valued samples of the transitivity function following or preceding a non-zero-valued sample by a half period of the cosine function.

3.2.4 Wavelet Reconstruction

The wavelet coefficients were multiplied by the thresholded transitivity function $f_k^t[n]$ obtained for that packet, enhancing coefficients that correspond to transient speech. The resulting thresholded wavelet coefficients $v_k'[n]$ were used to synthesize a signal $y[n]$ that we call the transient speech signal.

3.2.5 Unvoiced Speech Booster

The transitivity function has larger peaks for transitions into and out of high energy formants than for transitions associated with low energy events such as unvoiced consonants. The incorporation of an unvoiced speech booster to the transient extraction method to increase the peaks of the transitivity function that correspond to unvoiced consonants was investigated. A diagram of the transient extraction method with unvoiced speech booster is shown in Figure 16. In addition to the transient extraction method, this version of the algorithm also includes a voiced/unvoiced detection method and an extension method. The output of the voiced/unvoiced detection process VUV_i is a voiced/unvoiced decision signal that has a value of one when a window segment of speech is voiced and a value of zero when it is unvoiced. The extension method extends VUV_i, which has as many samples as there are windowed segments used for computing it, to have as many samples as the packet wavelet coefficients giving $VUV[n]$.

Using the signal $VUV[n]$, the unvoiced speech booster processes each packet and sets samples of the transitivity function that occur when speech is unvoiced to the maximum value of the transitivity function for the packet. Samples of the transitivity function that occur when speech is voiced or silent unchanged are computed as described in the previous section.

The two voiced/unvoiced detection methods described in Chapter 2, were considered for use with the unvoiced speech booster method – a short-time energy/short-time average zero-crossing rate-based method and a short-time autocorrelation function-based method. The short-time autocorrelation function-based method produced better voiced/unvoiced detection results as described in Appendix C and its incorporation into the transient extraction algorithm was evaluated experimentally.

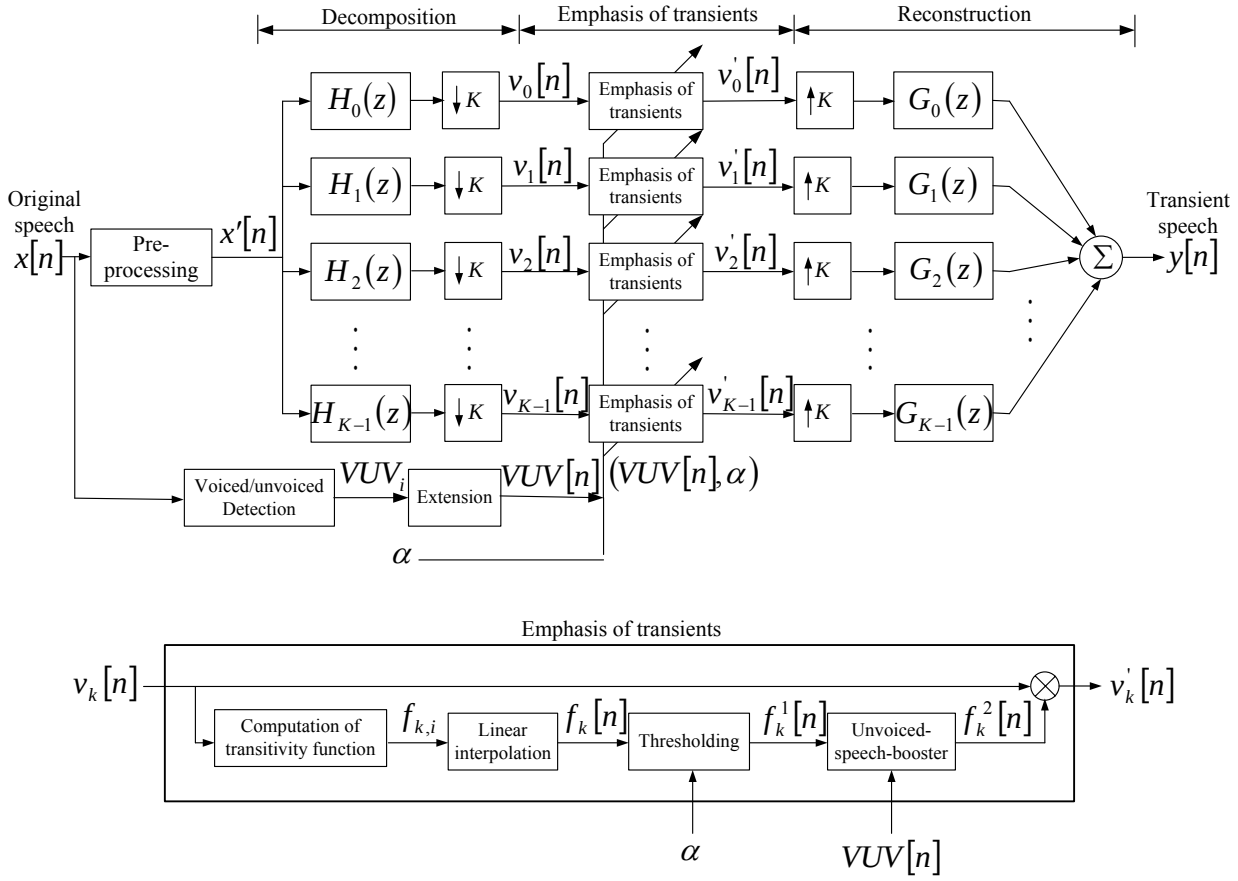

Figure 16: Transient speech extraction method with unvoiced speech booster. Compared to the transient extraction method of Figure 13, this version additionally includes a voiced/unvoiced detection method and the unvoiced speech booster.

3.2.6 Speech Modification

Modified speech $z[n]$ that emphasizes transient speech was formed by combining the transient speech signal $y[n]$ amplified by an enhancement factor β with the original speech $x[n]$, i.e.

$$z[n] = \rho(x[n] + \beta y[n]) \tag{19}$$

ρ is a scaling factor used to adjust modified speech so that its energy is equal to that of original speech. Including original speech gives the modified speech more voicing than transient speech and causes it to sound more natural. Modified speech (not transient speech) is the final speech intelligibility enhancing signal that would be presented to a listener in a communication system.

3.3 ILLUSTRATIONS OF TRANSITIVITY FUNCTIONS AND TRANSIENT SIGNALS

Transitions in a speech signal can manifest as change in amplitude, change in frequency, or change in both amplitude and frequency. To show that the transitivity function can identify these transitions, transitivity functions and the transient component of a synthetic signal are presented. The synthetic signal, with schematic shown in Figure 17, consists of two components. The first component, referred to as $C1$, is a steady tone of frequency 0.5 kHz with 50 ms. zero-padding at the beginning and end. The second component, beginning 50 ms. after and ending 50 ms. before the first, includes a tone of frequency F_1 and a transition via a linear chirp of duration t_{chirp} to another tone of frequency F_2. The first and the second tones of the second component will be referred to as $C2-1$ and $C2-2$, respectively. The duration of the two tones of the second component is 200 ms. The duration of component $C1$ is equal to the duration $C2-1$ + duration of $C2-2$ + duration of chirp + 100 ms. Both components were multiplied by a Tukey window to create gradual onsets and offsets. In the synthetic signal, the steady tones are intended to

model quasi-steady-state activity. The onset and offset of the tones and the chirp are intended to model transient activity.

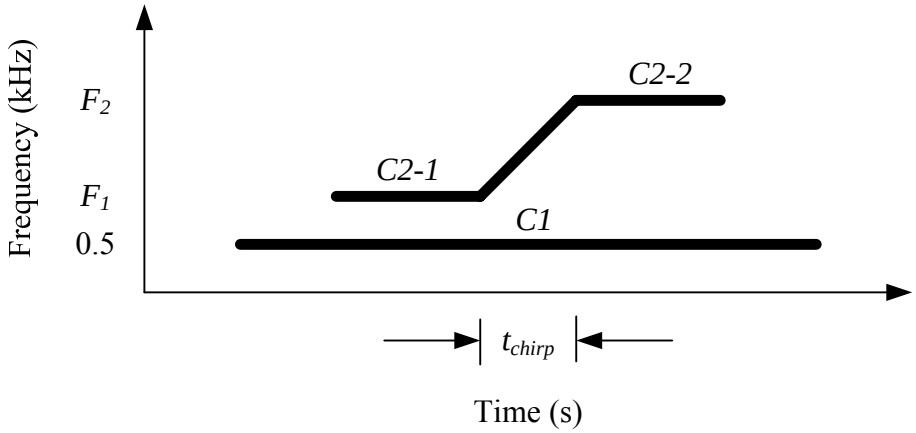

Figure 17: Schematic of synthetic signal used to evaluate and compare the transitivity functions.

To demonstrate the transitivity function and its use in identifying transients, Figure 18 shows the time-domain plot and spectrogram for the synthetic signal with F_1 = 1200 Hz, F_2 = 2600 Hz and t_{chirp} = 80 ms, and Figure 19 shows wavelet coefficients and transitivity functions for packets 1, 3, 5, 6, and 7 ($v_1[n]$, $v_3[n]$, $v_5[n]$ $v_6[n]$ and $v_7[n]$ in Figure 13) of the scale level 4 decomposition of this synthetic signal. The spectrogram was computed using a Hamming window of length 40 ms, a window step size of 0.1 ms. The spectrogram intensity values (z-axis) are logarithmic. The transitivity functions were computed using the MFCC-based method. Similar results but with less emphasis of the chirp were obtained using the STE-based method transitivity function. Wavelet coefficients $v_1[n]$ include 97 % of the energy of $C1$, $v_3[n]$ include 67 % of the energy of $C2-1$, $v_5[n]$ and $v_6[n]$ include 50 % of the energy of the chirp, and $v_7[n]$

include 45 % of the energy of $C2-2$. We expect the transitivity functions to have peaks at times that correspond to the onsets and offset of components $C1$, $C2-1$ and $C2-2$, and during the chirp. We also expect the thresholded wavelet coefficients, which are obtained by multiplying the wavelet coefficients by their corresponding transitivity functions, to be non-zero when the transitivity functions have peaks.

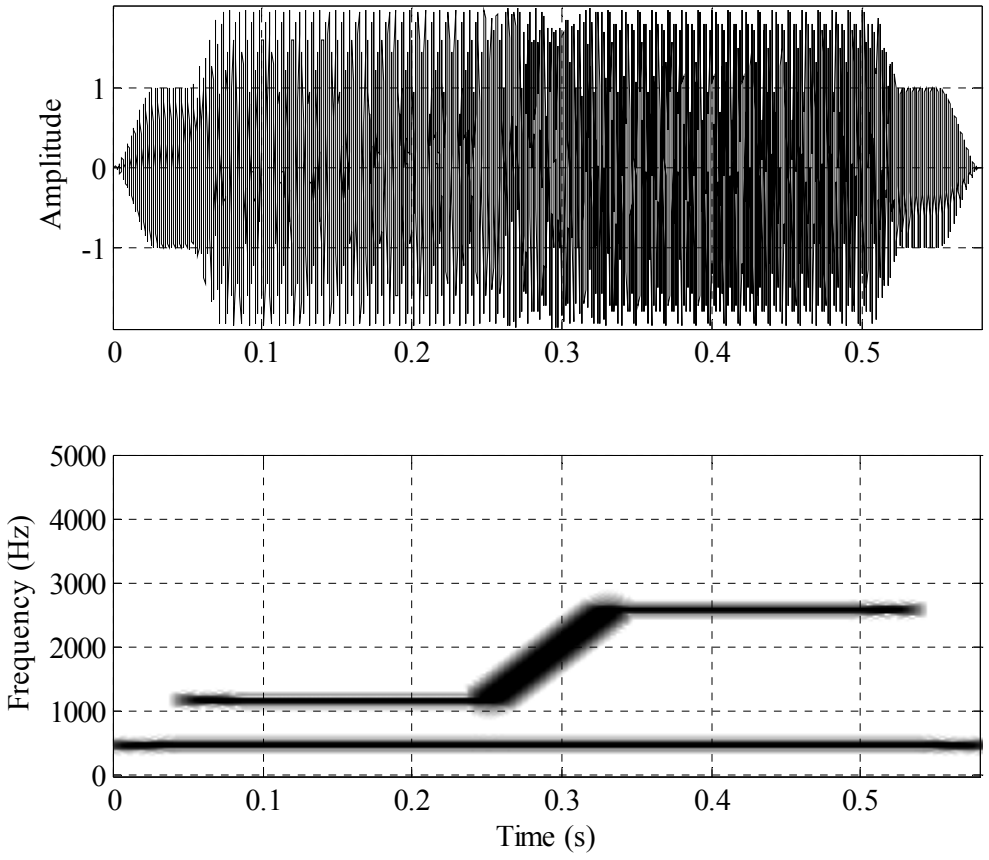

Figure 18: Time-domain plot and spectrogram for synthetic signal with F_1 = 1200 Hz, F_2 = 2600 Hz and t_{chirp} = 80 ms.

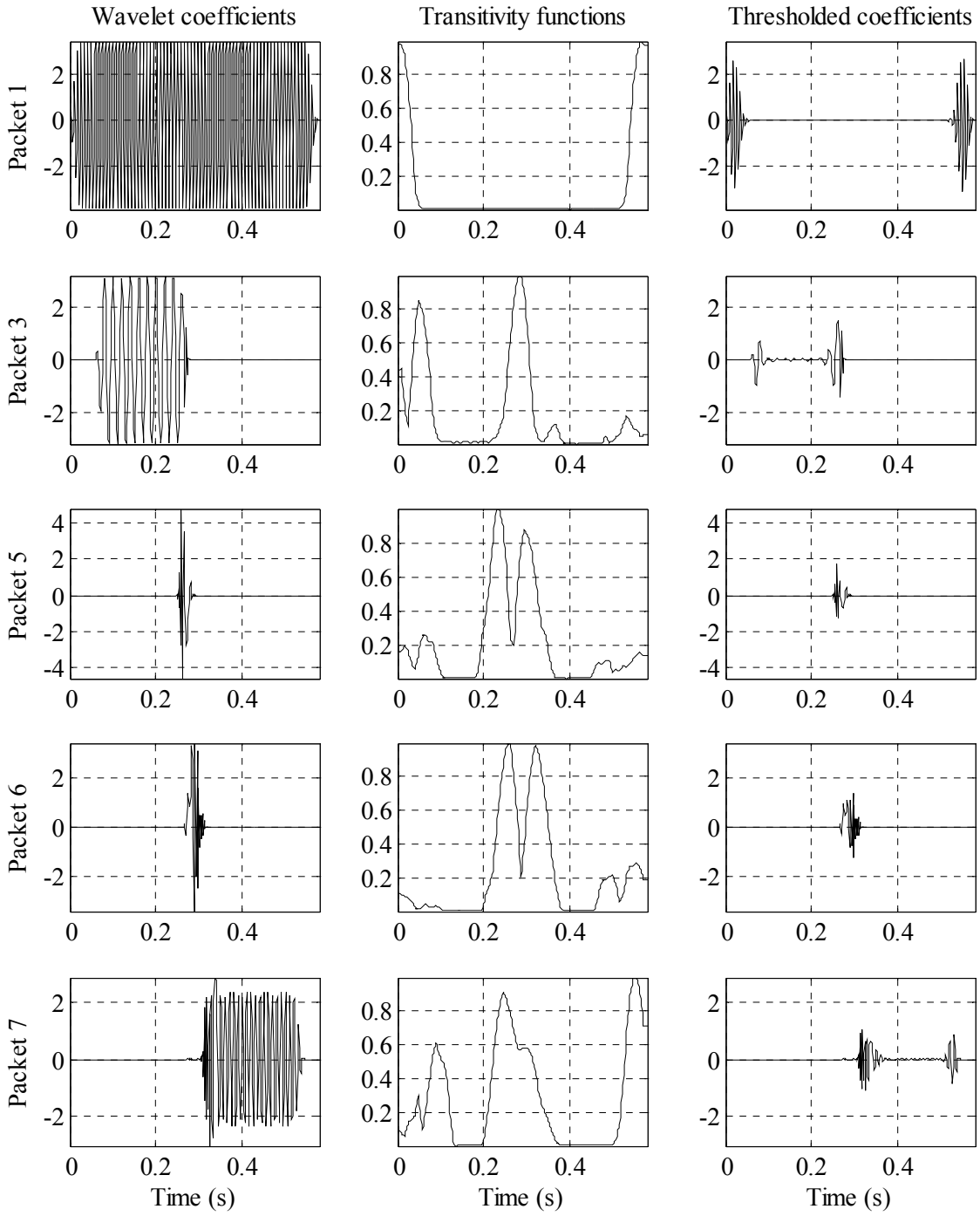

Figure 19: Demonstration of transitivity function and its use in identifying transients. Left column shows wavelet packet coefficients for packet 1, 3, 5, 6 and 7. The middle column shown transitivity functions computed from these coefficients. The right column shows thresholded wavelet packets coefficients.

The transitivity functions of $v_1[n]$, $v_3[n]$, $v_5[n]$ $v_6[n]$ and $v_7[n]$ have peaks at times that correspond to onset and offset of the components $C1$, $C2-1$ and $C2-2$ and the chirp, as expected. The thresholded coefficients for packets 1, 3, 5, 6 and 7 ($v_1'[n]$, $v_3'[n]$, $v_5'[n]$, $v_6'[n]$ and $v_7'[n]$ in Figure 13), are also shown in Figure 19. These coefficients are non-zero at times that correspond to onsets and offsets of the components and the chirp, as expected. The thresholded wavelet coefficients are used to synthesize the transient component of this signal. Although the transitivity function of packet 7 has a peak when the coefficients of the packet are in steady-state, (peak centered at t = 0.088 s), this does not affect the thresholded coefficients of this packet because the amplitude of the coefficients during this steady-state segment is very small.

To illustrate the transient signal, Figure 20 show a time-domain plot and spectrogram for the transient signal extracted from the synthetic signal of Figure 18. The transient signal emphasizes the onsets of components $C1$ and $C2-1$, the chirp and the offsets of component $C1$ and $C2-2$.

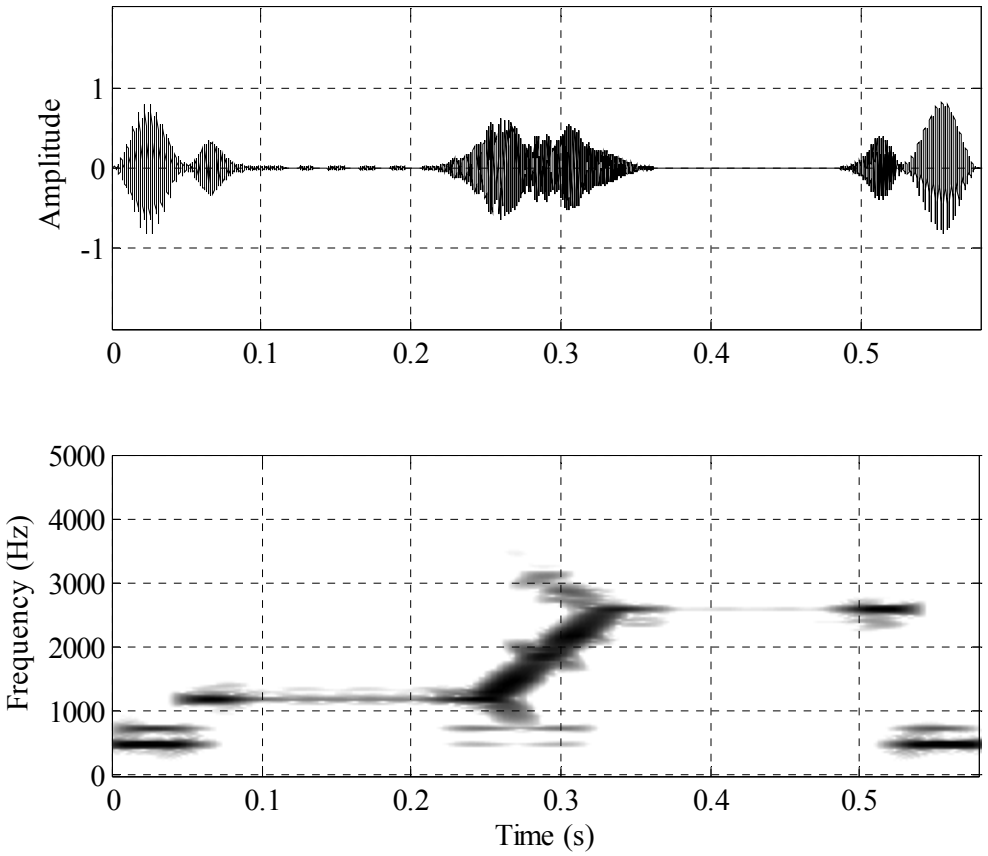

Figure 20: Time-domain plot and spectrogram for transient signal extracted from the synthetic signal of Figure 18. The onset, offset of the tones, as well as the chirp, are emphasized.

3.4 SUMMARY

A method for extracting transient speech from an original speech signal was described. The method involves decomposing a speech signal into 16 packets using the wavelet packet transform and computing a transitivity function to characterize the rate of change of the wavelet coefficients. The transitivity function, which is computed from Mel-frequency cepstral coefficients (MFCC) or from the short-time energy (STE), is used to emphasize the wavelet packet coefficients when they are changing rapidly. The inverse wavelet packet transform of the

transient-emphasized wavelet coefficients give a transient signal that emphasizes the onset and offset of vowel formants and unvoiced consonants. The transient extraction method includes a weight parameter, which when varied controls the amount of quasi-steady-state vowel activity that is included in the transient speech signal.

Although MFCCs are not traditionally applied to wavelet coefficients, their application in this case is reasonable as will be discussed in Chapter 6.

Synthetic signals, which model specific transient activities, were used to illustrate the transitivity function. The transitivity function has peaks during time segments when the frequency or amplitude of a signal is changing rapidly.

4.0 COMPARISONS OF TRANSIENT AND MODIFIED SPEECH

The transient extraction method has been applied to a wide range of speech material, including monosyllable consonant-vowel-consonant (CVC) words and sentences. This chapter presents results to illustrate transient and modified speech and the effect of different parameter values/options on transient speech. Comparisons of our transient and modified speech to transient, modified and processed speech obtained by other researchers are also presented. To facilitate these comparisons, indices that we developed and use to compare speech signals are first described.

4.1 INDICES FOR COMPARING SPEECH

In order to compare our transient speech signals extracted using different parameter values to each other and to transient speech signals extracted using methods of Yoo *et al.* and Tantibundhit *et al.*, some measure of a speech signal's "transient-ness" compared to original speech was needed. This same measure is required to compare our modified speech to modified or processed speech signals that have been proposed by other researchers for enhancement of speech intelligibility. This section describes three indices that were developed for this purpose. An index P was developed to compare the effect of a speech modification/processing method on a particular region of speech.

Important characterizations of transient speech are the extent to which onsets and offsets of formants are emphasized relative to steady-state regions of formants and the extent to which consonants are emphasized relative to vowels. Two indices were developed for these characterizations. Index R was developed to characterize the extent to which consonants are emphasized relative to vowels in a speech signal and index Q was developed to characterize the extent to which the onsets and offsets of formants are emphasized compared to steady segments in a speech signal [82]. These indices quantify differences in speech signals that are difficult to show using spectrograms, spectra or time-domain waveforms.

A common step in the computation of the three indices involves the placement of a time-frequency mask, rectangular blocks specified by a time interval (t_1, t_2) and a frequency interval (f_1, f_2), on a spectrogram of a test word and the calculation of the energy within this mask. An example of a time-frequency mask superimposed on a spectrogram is shown in Figure 21. The word is 'pack' (phonetically transcribed as /pæk/) spoken by a male. This phonetic transcription and subsequent transcriptions were obtained from http://dictionary.reference.com/. The word was sampled at 11025 Hz and the spectrogram was computed using a 5 ms Hamming window and a 1 ms window step size. The spectrogram intensity values (z-axis) are logarithmic. The figure also identifies parameters t_1, t_2, f_1 and f_2. In the example, the time-frequency mask is superimposed on the first formant of /æ/ and t_1 = 0.100 s., t_2 = 0.305 s., f_1 = 350 Hz and f_2 = 1200 Hz.

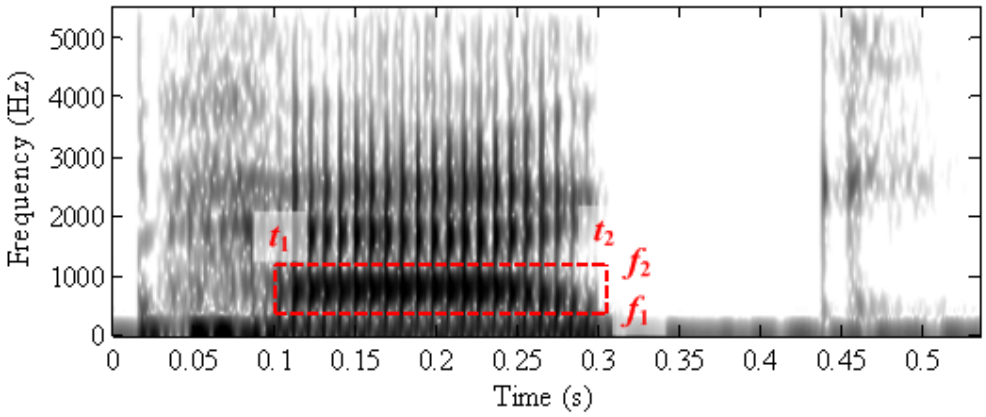

Figure 21: Demonstration of time-frequency mask and parameters t_1, t_2, f_1 and f_2, which are used in the computation of the energy within the time-frequency mask.

The energy within the time-frequency mask $E(t_1, t_2, f_1, f_2)$ can be computed as

$$E(t_1, t_2, f_1, f_2) = \sum_{t=t_1}^{t_2} \sum_{f=f_1}^{f_2} |S(t, f)|^2 \tag{20}$$

where $S(t, f)$ is the short-time Fourier transform used to compute the spectrogram.

4.1.1 Index *P*

Index P was used to compare speech signals. P is used to compare the effect of a speech modification/processing method on a particular region of speech. Specifically, P is used in Section 4.2 to show that transient speech extracted using the MFCC transitivity function retains

more energy of diphthongs than transient speech extracted using the STE transitivity function. Diphthongs are characterized by formants whose frequency changes with time.

To compute P, the energy of particular region of interest is computed for transient, modified or processed speech Equation (20). The energy for the same region is also computed for original speech. If we represent the energy in the region of interest in transient, modified or processed speech as $E_y(t_1, t_2, f_1, f_2)$ and the energy in the same region in original speech as $E_x(t_1, t_2, f_1, f_2)$, then index P is the ratio of the value of $E_y(t_1, t_2, f_1, f_2)$ to the value of $E_x(t_1, t_2, f_1, f_2)$, i.e.

$$P(x, y) = \frac{E_y(t_1, t_2, f_1, f_2)}{E_x(t_1, t_2, f_1, f_2)} \tag{21}$$

P will typically be expressed as a percentage. A large value of P indicates an increased emphasis of the particular region of interest of speech by a given speech modification/processing method.

4.1.2 Index *R*

Index R was developed to characterize the extent to which consonants are emphasized relative to vowels in a speech signal. Index R is used to quantify the difference in modified and processed speech signals and to compare these speech signals. In the computation of index R, time-frequency masks are manually placed on the initial consonant, vowel and final consonant of original and transient/modified/processed speech of a test word. Eighteen words, listed in Table

1, of the form consonant-vowel-consonant (CVC) obtained from a recording (male speaker) of the modified rhyme test (MRT) word list [13] were used as the test words.

Table 1: List of the 18 CVC words that were used for computation of index R. These words were obtained from the MRT word list

Mat
Man
Mad
Mass
Sass
Sat
Sap
Sack
Pad
Pass
Pat
Pack
Pane
Back
Bath
Tap
Tack
Tab

Figure 22 demonstrates the placement of time-frequency masks for the computation of R for the word pack /pæk/ spoken by a male. The spectrogram intensity values (z-axis) are logarithmic. The time-frequency mask on the left, with time and frequency intervals (t_1, t_2) = (0, 0.095) s. and (f_1, f_2), includes the initial consonant /p/, the mask in the middle, with time and frequency intervals (t_3, t_4) = (0.100, 0.305) s. and (f_1, f_2), includes the vowel /æ/ and the mask on the right, with time and frequency intervals (t_5, t_6) = (0.430, 0.535) s. and (f_1, f_2), includes the final consonant /k/. A single frequency interval of $(f_1, f_2) = (0, f_S/2 = 5512.5\text{Hz})$ was used for all three time-frequency masks.

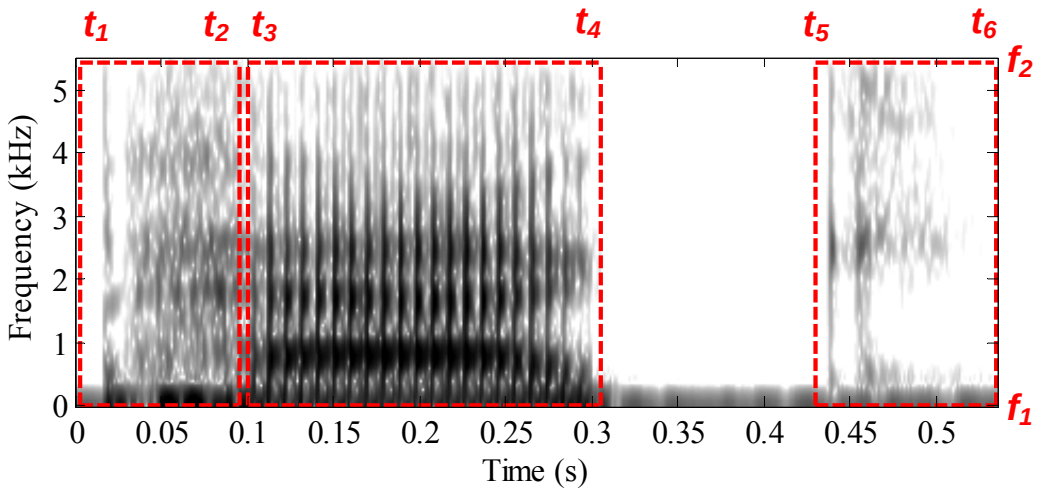

Figure 22: Demonstration of the placement of time-frequency masks on a spectrogram for computation of the index R. Time and frequency intervals for the masks involved in the computation of index R are shown. The word is 'pack' spoken by a male.

To compute R, first the energies in the time-frequency masks that include the initial consonant, vowel and final consonant are computed for both original and transient/modified/processed using Equation (20). That is, separate energy values are computed for the initial consonant $E(t_1,t_2,f_1,f_2)$, vowel $E(t_3,t_4,f_1,f_2)$ and final consonant $E(t_5,t_6,f_1,f_2)$. The ratio (C/V) of the sum of the energy in the consonants ($E(t_1,t_2,f_1,f_2)$ and $E(t_5,t_6,f_1,f_2)$) to the energy in vowel ($E(t_3,t_4,f_1,f_2)$) is then computed

$$C/V = \frac{E(t_1,t_2,f_1,f_2)+E(t_5,t_6,f_1,f_2)}{E(t_3,t_4,f_1,f_2)} \tag{22}$$

R is the logarithm of the value of C/V for modified/processed speech normalized by the value of C/V for original speech, i.e.

$$R = \log\left(\frac{C/V_{\mathrm{modified}}}{C/V_{original}}\right) \tag{23}$$

The range of the ratio $\frac{C/V_{\mathrm{modified}}}{C/V_{original}}$ is large and a logarithmic transformation was used to reduce the range. If a particular speech processing method emphasizes consonants relative to vowels, $R > 0$ and if it de-emphasizes consonants, $R < 0$. Two modified/processed speech signals are considered to provide similar emphasis of consonants if they have approximately equal values of R. A larger R value indicates an increased emphasis of consonants.

4.1.3 Index *Q*

Index Q was developed to characterize the extent to which the onsets and offsets of formants are emphasized compared to steady segments in a speech signal. This index was previously described in [82], where it was referred to as the 'transient index'. In the computation of index Q, time-frequency masks are manually placed on the onsets, steady-state segments and offsets of the 2nd, 3rd and 4th formants of test words. The eighteen test words listed in Table 1 were also used to compute Q. Words with relatively steady vowel segments and with 2nd, 3rd and 4th formants that were roughly coincident in time were used to provide a relatively unambiguous identification of onsets, offsets and steady-state segments of formants. The 1st formant was not included because it makes a limited contribution to intelligibility, as was shown in [1], [83]. Time intervals were selected such that a single interval captured the onsets, steady-state segments or offsets of formants 2 to 4 for each test word. The minimum width of the intervals for onsets and offsets was 0.05 s. The gap between onset or offset and steady-state segment was selected to be at least 0.01 s. The frequency intervals for each formant were selected such that the intervals of adjacent formants did not overlap.

Figure 23 demonstrates the placement of time-frequency masks on a spectrogram for the computation of index Q. The word is 'pack' (phonetically transcribed as /pæk/) spoken by a male. The spectrogram intensity values (z-axis) are logarithmic. The three time-frequency masks on the left, occurring during the time interval (t_1, t_2) = (0.07, 0.12) s., includes the onset of the formants of /æ/. The three time-frequency masks in the middle, occurring during the time interval (t_3, t_4) = (0.13, 0.25) s., include the steady-state segment of the formants of /æ/ and the three time-frequency masks, occurring during the time-interval (t_5, t_6) = (0.26, 0.31) s., include

the offset of the formants. The frequency intervals are: 2nd formant, (f_1, f_2) = (1200, 2150) Hz, 3rd formant, (f_3, f_4) = (2250, 3100) Hz and 4th formant, (f_5, f_6) = (3200, 4300) Hz.

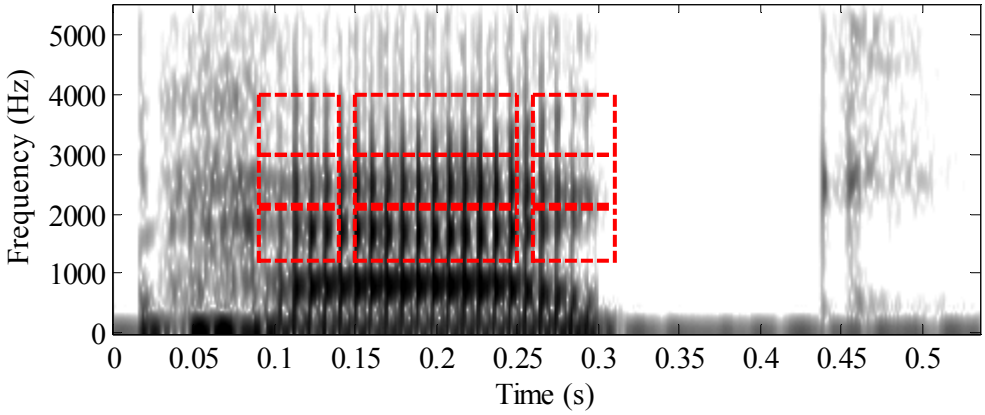

Figure 23: Demonstration of the placement of time-frequency masks on a spectrogram for computation of index Q. The word is 'pack' spoken by a male

To compute index Q, the energy in each of the nine time-frequency masks are computed using Equation (20). Using the mask energies, the ratio of energy in onset/offset to steady-state segments was defined for each formant. As an example, the ratio for the third formant is

$$Q_3 = \frac{E(t_1, t_2, f_3, f_4) + E(t_5, t_6, f_3, f_4)}{E(t_3, t_4, f_3, f_4)} \qquad (24)$$

where $E(t_1, t_2, f_3, f_4)$, $E(t_3, t_4, f_3, f_4)$ and $E(t_5, t_6, f_3, f_4)$ are the energies of time-frequency masks that include the onset, steady-state segment and offset, respectively. Similar ratios Q_2 and Q_4 were computed for the second and fourth formants.

Index Q for a transient, modified or processed speech signal is the logarithm of the sum of Q_i, $i = 2$ to 4, for the 3 formants of transient speech normalized by the sum of Q_i for the 3 formants of original speech,

$$Q = \log\left(\frac{\sum_{i=2}^{4} Q_{i,transient}}{\sum_{i=2}^{4} Q_{i,original}}\right) \qquad (25)$$

The range of the ratio $\sum_{i=2}^{4} Q_{i,transient} \Big/ \sum_{i=2}^{4} Q_{i,original}$ is large and a logarithmic transformation was used to reduce the range. If a particular transient speech signal emphasizes onsets and offsets relative to steady segment of formants, $Q > 0$. If it de-emphasizes onsets and offsets, $Q < 0$. Two transient speech signals are considered to provide similar emphasis of onsets and offsets of formants if they have approximately equal values of Q. A larger value of Q indicates increased emphasis of onsets and offsets of formants. This index can be applied to any transient speech signal without regard to how it was derived.

4.2 ILLUSTRATION OF TRANSIENT AND MODIFIED SPEECH

Results, using the word 'pack' from the modified rhyme test list, are presented here to illustrate transient speech and the effects of different weight parameter values and the unvoiced speech booster on the transient speech. The results are representative of results obtained for all the words in the list. The word 'pack', phonetically transcribed as /pæk/, includes a unvoiced bilabial

stop consonant /p/, a vowel /æ/, and an unvoiced velar stop consonant /k/. The word was spoken by a male and was sampled at 11025 Hz.

Quasi-steady state speech, obtained by subtracting transient speech from original speech, is also illustrated. Indices Q and $P(x, y)$ are used to make comparisons between different transient speech signals.

Transient speech extracted using the MFCC and the STE transitivity functions are also presented to illustrate the difference between the two. For this illustration, a sentence with several diphthongs is used, instead of the word 'pack'. Diphthongs are characterized by formants whose frequencies change with time. Using words with diphthongs for this illustration provides a clearer demonstration that the MFCC-transitivity function can capture changes in frequency better than the STE-transitivity function.

The time-domain waveform and spectrogram for the word 'pack' are shown in Figure 24. This spectrogram and spectrograms presented later were computed using a Hamming window of length 5 ms and a window shift of 1 sample (0.1 ms). The spectrogram intensity values (z-axis) are logarithmic.

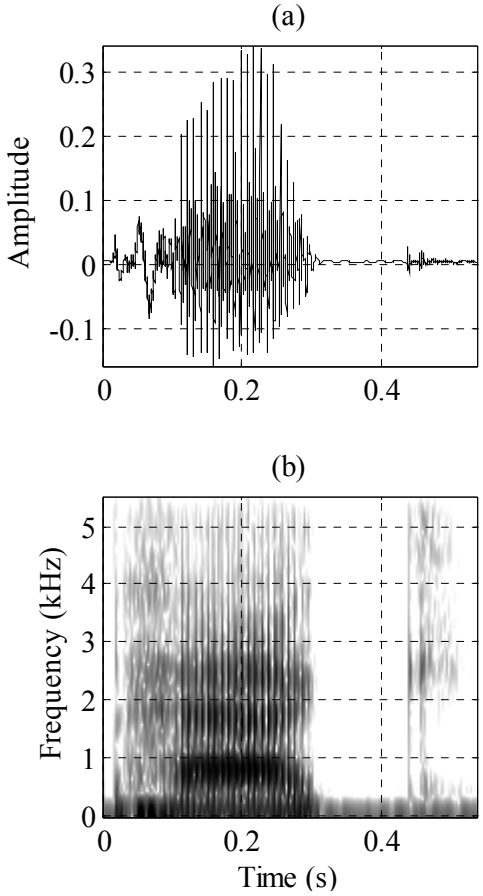

Figure 24: (a) Time-domain waveform and (b) spectrogram for the word ‘pack’.

To demonstrate the transient and quasi-steady state speech signals, Figures 25 and 26 show time-domain waveforms and spectrograms, respectively, for the transient and quasi-steady state components of the word 'pack' obtained without and with thresholding. In both figures, the left column shows transient speech and the right column shows quasi-steady-state speech. Figures 25 and 26 also demonstrate the effect of thresholding by comparing transient speech signals extracted using weight parameter values of $\alpha = 0.9$ and $\alpha = 1.0$. In both figures, the top row shows results for $\alpha = 1.0$, which is equivalent to extracting transient speech without thresholding and the bottom row show results for $\alpha = 0.9$. A weight parameter value of $\alpha = 0.9$

suppresses most of the quasi-steady state activity in the transient speech from 0.1 to 0.24 s. as can be seen in by comparing Figure 25 (a) to Figure 25 (b), and Figure 26 (a) to Figure 26 (b).

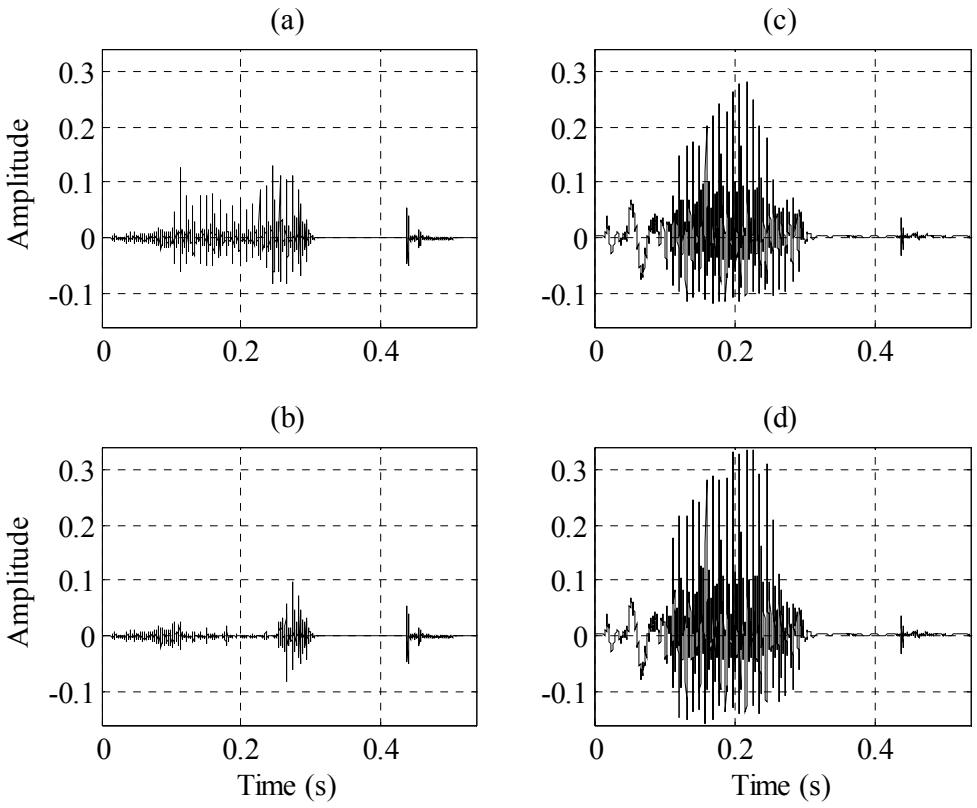

Figure 25: Time-domain waveforms for transient speech signals extracted (a) without thresholding ($\alpha = 1$) and (b) with thresholding using $\alpha = 0.9$. (c) and (d) show quasi-steady state speech signals obtained by subtracting (a) and (b) from original speech, respectively. The word is 'pack' spoken by a male.

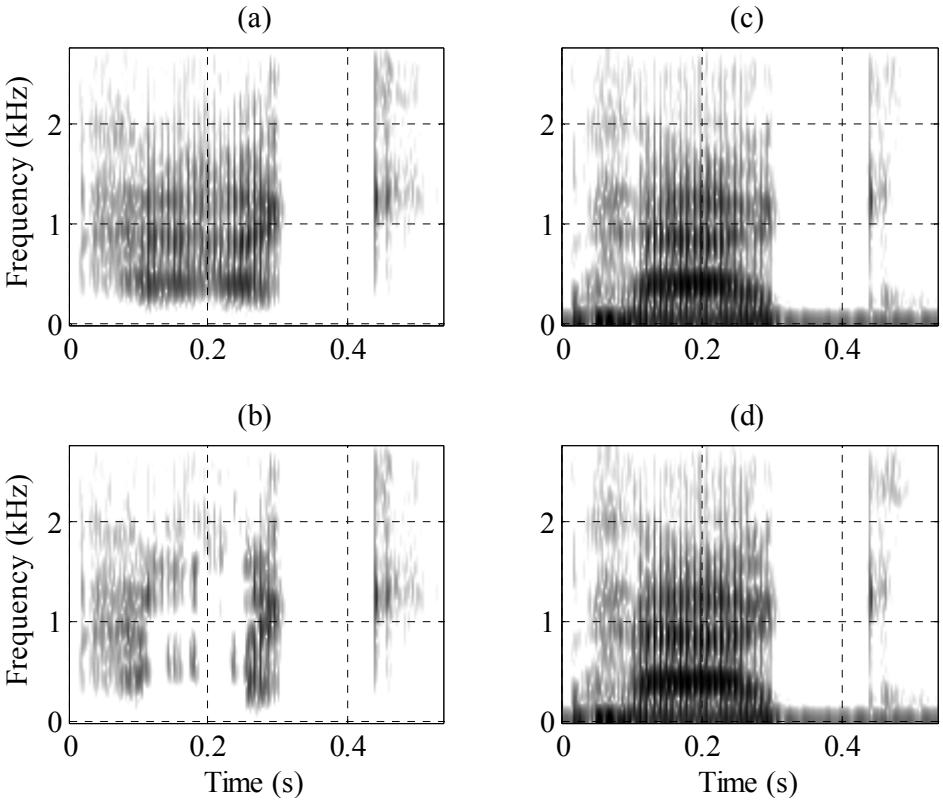

Figure 26: Spectrograms for transient speech signals extracted (a) without thresholding ($\alpha = 1$) and (b) with thresholding using $\alpha = 0.9$. (c) and (d) show quasi-steady state speech signals obtained by subtracting (a) and (b) from original speech, respectively. The word is 'pack' spoken by a male.

In both transient speech signals, the pre-processing highpass filter removed most of the energy of the first formant. The transient speech signals emphasize the onset, occurring at 0.09 s., and the offset, occurring at 0.26 s., of formants of the vowel /æ/. Also, the velar stop consonant /k/, from 0.44 s., is emphasized in the transient speech signals particularly at its beginning and ending.

The transient speech signals obtained without thresholding retain more vowel activity. The values of index Q for transient speech signals for the word ‘pack’ obtained without and with

thresholding are Q = 0.53 and Q = 3.86, respectively. The higher Q value for the transient speech signal obtained with thresholding shows that this signal emphasizes transient speech more than the transient speech signal obtained without thresholding. These results show that the amount of quasi-steady state energy in a transient speech signal can be controlled by varying the weight parameter, α .

Both quasi-steady-state speech signals de-emphasize onset and offset of formants. Although the quasi-steady state speech signals appear to be similar to the original, their values of index Q are less than zero (Q = -1.49 and Q = -1.29 without and with thresholding, respectively), which indicates de-emphasis of onset and offset of formants.

During informal listening tests, as weight parameter α for the thresholding operation decreased from one towards zero, there was an increase in a 'garble-like' artifact noise and an associated decrease in speech quality. However, transient speech was perceived to be more emphasized as α decreased.

To demonstrate the effect of the unvoiced speech booster (USB), Figure 27 shows spectrograms for transient speech signals for the word 'pack', obtained without and with the use of the unvoiced speech booster. Figure 27 (a) is a repeat of Figure 26 (a). The index P equals 116% without USB and 299% with USB, showing that the transient speech signal extracted with incorporation of the unvoiced speech booster includes more energy of the unvoiced consonant /k/. Index P can have values greater than 100% because the transitivity function is allowed to have values greater than 1 which result in boosting of energy by the extracted transient speech. The transient speech extracted without USB does not emphasize the energy of the consonant /k/ between frequencies of 1 and 2 kHz and above a frequency 3 kHz, as can be seen in Figure 27

(a). In informal listening tests, consonants were slightly more intelligible in noise and sounded more natural when the unvoiced speech booster was used.

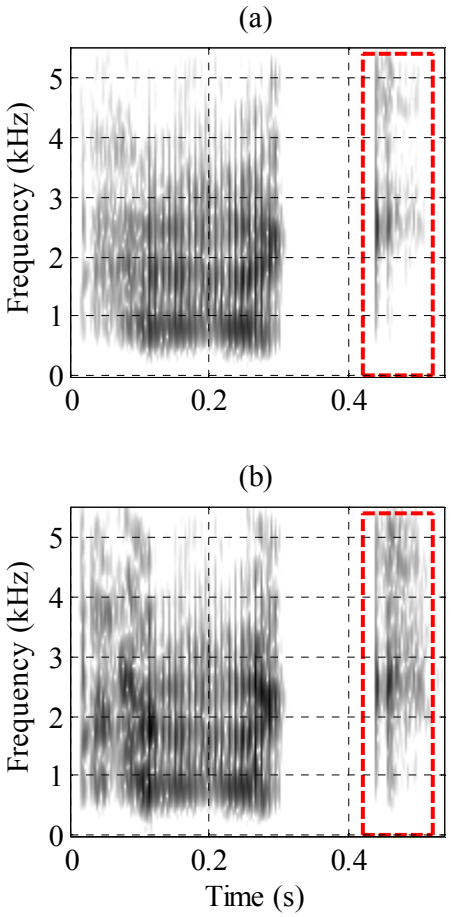

Figure 27: Demonstration of the effect of unvoiced speech booster. Spectrograms for MFCC-based transient speech signals obtained (a) without and (b) with utilization of unvoiced speech booster. The word is ‘pack’ spoken by a male.

A second example is presented to illustrate the difference between transient speech signals extracted using the MFCC-transitivity function and transient speech signals extracted with the STE-transitivity function. The sentence 'Here-is-a-nice-quiet-place-to-rest,' phonetically

transcribed as /hɪər-ɪz-eɪ-naɪz-‘kwaɪɪt-pleɪs-tu-rɛst/, spoken by a male is used for this example. This sentence was obtained from the CDROM that accompanies [84]. The sentence includes diphthongs, which demonstrate more clearly the differences between the two transient speech signals. The spectrograms for the original speech signal and the two transient speech signals are shown in Figure 28. The two transient speech signals were adjusted to have equal energy. The frequencies of the second formants of the vowels /ɪə/ in 'here' between 0.03 and 0.19 s., /aɪ/ in 'nice' between t = 0.40 and t = 0.56 s., /aɪɪ/ in 'quiet' between t = 0.78 and t = 0.9 s. and /eɪ/ in 'place' between t = 1.1 and t = 1.24 s are changing with time. Time-frequency masks were superimposed on these vowels and used to compute values of index $P(x, y)$ for comparing the transient speech signals. These values are presented in Table 2.

The two transient speech signals are similar in that they both emphasize onset and offset of vowel formants and consonants. However transient speech extracted using the STE-transitivity function does not include as much of the energy of second vowel formants that are changing in frequency as does the transient speech extracted using the MFCC-transitivity function. In informal listening tests, the MFCC-based transient speech had less 'garble-like' artifact noise and was more intelligible.

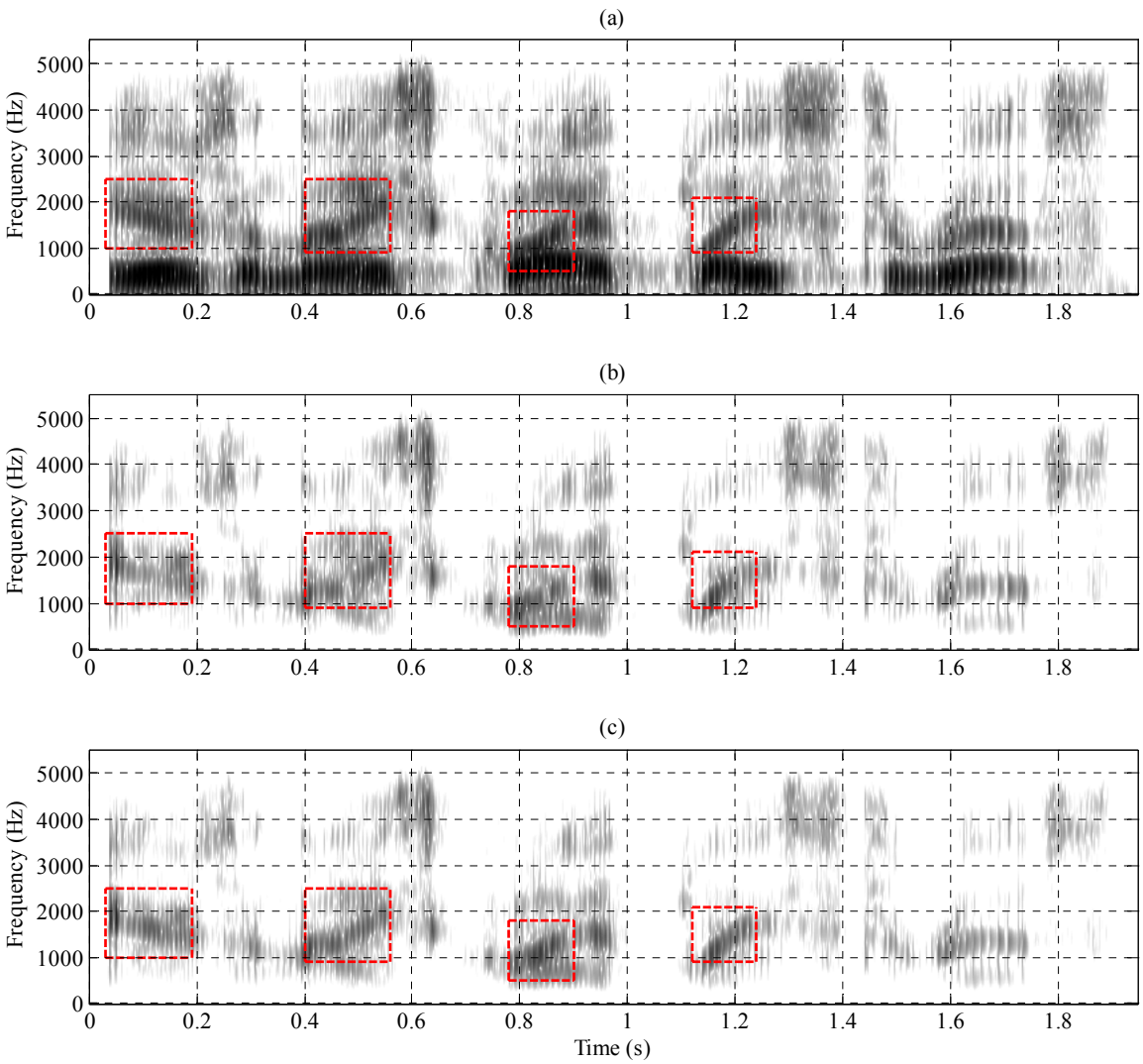

Figure 28: Spectrograms for (a) original speech and transient speech signals extracted using (b) STE-based transient speech signal (c) MFCC-based transient speech signal for the sentence 'Here-is-a-nice-quiet-place-to-rest,' spoken by a male. The rectangles superimposed on the spectrograms are time-frequency masks used to compute index P.

Table 2: Values for P for the time-frequency masks that include the second vowel formants of the vowels /ɪə/ in 'here', /aɪ/ in 'nice', /aɪɪ/ in 'quiet' and /eɪ/ in 'place'.

Vowel	P	
	STE-based transient speech	MFCC-based transient speech
/ɪə/ in 'here'	5.97 %	10.9 %
/aɪ/ in 'nice'	0.571 %	1.04 %
/aɪɪ/ in 'quiet'	14.6 %	31.4 %
/eɪ/ in 'place	7.15 %	9.11 %

4.3 COMPARISON TO OTHER SPEECH MODIFICATION METHODS

Indices Q and R and spectrograms were used to compare our transient and modified speech to transient, modified and processed speech obtained by other researchers. In particular, the following three sets of comparisons were made

Our transient speech was compared to transient speech obtained using the algorithms of Yoo *et al.* [2] and Tantibundhit *et al.* [4] to show that, like their transient speech, our transient speech emphasizes onsets and offsets of vowel formants relative to steady formant activity. Additionally, we demonstrate that our algorithm can extract transient speech signals that are

similar to those identified by Yoo *et al.* and Tantibundhit *et al.*, by adjusting the weight parameter.

In the second set of comparisons, our modified speech is compared to modified speech obtained using the algorithms of Yoo *et al.* [2] and Tantibundhit *et al.* [4] and to processed speech obtained using the algorithm of Villchur [47], Skowronski *et al.* [5] and Gordon-Salant [52] to evaluate the emphasis of onsets and offsets of vowel formants provided by our modified speech signal relative to these other modified/processed speech signals. Index Q and spectrograms are used for these comparisons. Villchur processed speech by splitting it into low and high frequency channels using filters, amplitude-compressing and equalizing each channel and then combining the signals from each channel [47]. Skowronski *et al.* processed speech by increasing the energy of unvoiced consonants relative to the energy of adjacent vowels [5]. Unvoiced segments were automatically identified using a spectral flatness measure to discriminate between consonants and vowels. Gordon-Salant processed speech by increasing the energy of manually identified consonants [52].

In the last set of comparisons, we evaluate the extent to which our modified speech emphasizes consonants relative to vowels. This emphasis of consonants is compared to that obtained by the algorithms of Yoo *et al.*, Tantibundhit *et al.*, Villchur, Skowronski *et al.* and Gordon-Salant [2] [4] [47] [5] [52]. Index R and spectrograms are used for these comparisons. Indices Q and R were computed following the methods described in Section 4.1.

Our transient speech was created using the MFCC-based algorithm without the unvoiced speech booster and our modified speech was created using this transient speech with, like Yoo *et al.* and Tantibundhit *et al.*, an enhancement factor of $\beta = 12$. Similar results were observed when using modified speech created with the STE-based transient speech.

The modified speech signals of Yoo *et al.* and Tantibundhit *et al.* were obtained directly from them. Skowronski and Harris' processed speech, which emphasizes unvoiced consonants, was created using their software implementation of their method (available at http://www.cnel.ufl.edu/~markskow/). Gordon-Salant's processed speech was created by identifying consonant and vowel segments in CVC words manually, computing the root-mean-square energy of these consonants and vowels, and then adjusting the amplitude of the consonants to achieve an increase in the consonant-to-vowel ratio of 10 dB. Villchur's modified speech was created using our software implementation of his method. The amplitudes of all modified speech signals were adjusted so that their energies equal that of original speech.

4.3.1 Comparison of Transient Speech

4.3.1.1 Identification of Weight Parameter

To identify weight parameter values for our algorithm that are required to extract transient speech signals that match transient speech of Yoo *et al.* and Tantibundhit *et al.*, transient speech for the 18 test words was extracted for a range of weight parameter (α) values from α = 0.7 to α = 1, and values of Q computed. The transitivity function, and hence the transient component, was essentially zero for α < 0.7. Figure 29 shows Q averaged over the 18 words as a function of α. As α decreases ($T(i)$ decreases, Equation (18)), the value of Q increases, indicating that steady formant segments in the transient speech signal are increasingly de-emphasized. The range of average Q values was (1.2, 18.01) and the minimum was obtained at α = 1.0.

Values of Q for the 18 test words (Table 1) were then averaged over the test words for transient speech of Yoo *et al.* and Tantibundhit *et al.*, and the average values were $Q_Y = 0.60$ and $Q_T = 4.53$, respectively. Values of α that result in our transient speech that are estimates of Yoo's and Tantibundhit's transient speech were then obtained from Figure 29, yielding $\alpha = 1.0$ ($Q_{\alpha=1.0} = 1.2$) for Yoo's transient speech and $\alpha = 0.875$ ($Q_{\alpha=0.875} = 4.60$) for Tantibundhit's transient speech. Since the minimum value of Q obtained using our method (1.2 at $\alpha = 1.0$) is greater than Q for Yoo's method, setting $\alpha = 1$ provided the best estimate of Yoo's transient speech that we could obtain. A closer match was obtained for Tantibundhit's method because the average value of Q for his transient speech is within the range (1.12, 18.01) of values of Q obtained with our method.

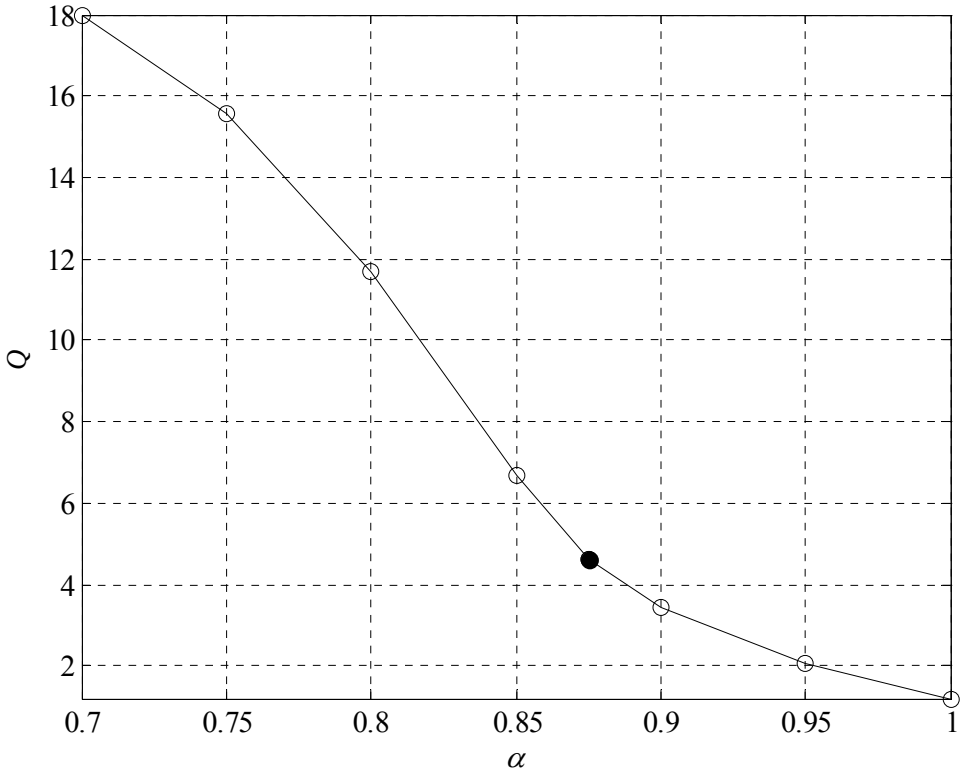

Figure 29: Q vs. α for our transient speech. Filled circle shows value of α that produces a value of Q that matches the value of Q obtained for Tantibundhit's transient speech.

4.3.1.2 Comparison of Transient Speech

Similarities and differences in transient speech components obtained with the three methods can be illustrated using the word 'pack', which is phonetically transcribed as /pæk/ and includes an unvoiced bilabial stop consonant /p/, a vowel /æ/ and an unvoiced velar stop consonant /k/. Figure 30 shows the spectrogram of 'pack' with the time-frequency masks used to compute the index Q superimposed. The word 'pack' is from the list of 18 test words of Table 1 which were used to compute the indices R and Q.

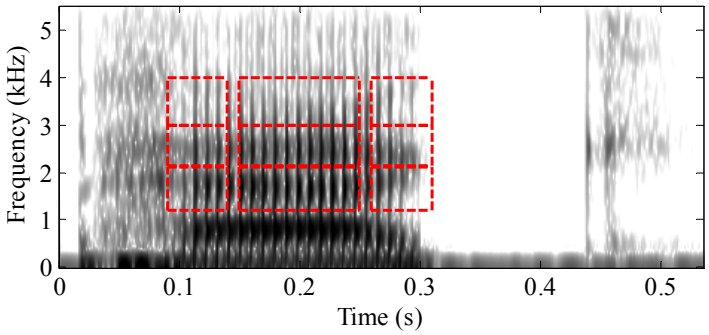

Figure 30: Spectrogram for the word 'pack', phonetically transcribed as /pæk/. The dashed rectangles are the time-frequency mask used to compute the index Q.

Figure 31 shows Yoo's and Tantibundhit's transient speech signals and our transient speech signals that are estimates of their signals. Our transient speech signals were extracted using α = 1.0 to match Yoo's transient speech and α = 0.875 to match Tantibundhit's transient speech. The top two panels of Figure 31 compare Yoo's transient speech to our transient speech extracted using α = 1.0, and the bottom two panels compare Tantibundhit's transient speech to our transient speech extracted using α = 0.875. All transient signals deemphasize the steady state segment of the vowel /æ/ and emphasize the onsets and offsets of formants of this vowel. The bilabial stop consonant /p/, occurring between 0 and 0.1 s and the velar stop consonant /k/, starting at 0.44 s, are emphasized by all four transient speech signals. Tantibundhit's transient speech retains more low frequency energy than the other transient speech because Tantibundhit did not apply highpass filtering.

The transient speech signals extracted using Yoo's method (Figure 31(a)) and our method with α = 1.0 (Figure 31(b)) are similar in that both retain some of the steady state vowel activity. The transient speech signals extracted using Tantibundhit's method (Figure 31(c)) and our method with α = 0.875 (Figure 31(d)) are similar in that both strongly de-emphasize steady-

state vowel activity. A difference between Tantibundhit's transient speech and our transient speech (α = 0.875) is that the former more completely removed energy associated with the steady-state segment of the 3rd formant, while our transient speech retained some of this energy.

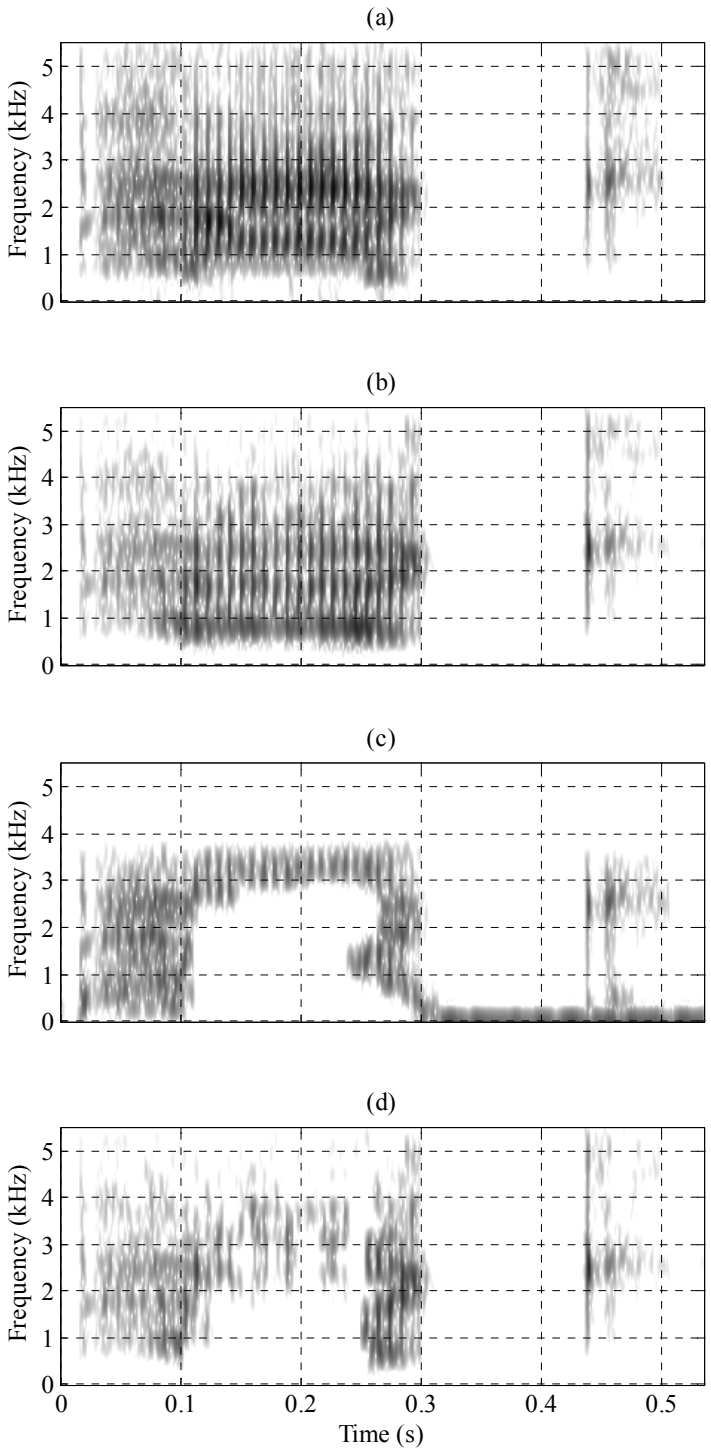

Figure 31: Transient speech extracted using (a) Yoo's method, (b) our method with α = 1.0, (c) Tantibundhit's method, and (d) our method with α = 0.875. The word is 'pack' /pæk/ spoken by a male.

For this word, values of the index Q are Q_Y = 1.26, $Q_{\alpha=1.0}$ = 1.12, Q_T = 4.52 and $Q_{\alpha=0.875}$ = 4.51. All four values are greater than zero, indicating that all four transient speech signals emphasize onsets and offsets of formants, relative to steady formant segments. $Q_{\alpha=1.0}$ is close to Q_Y indicating that our transient speech to match Yoo's transient speech emphasizes onset and offset of formants to a similar extent as Yoo's transient speech. $Q_{\alpha=0.875}$ is close to Q_T, indicating that our transient speech to match Tantibundhit's transient speech emphasizes onset and offset of formants to a similar extent as Tantibundhit's transient speech.

The average values of Q obtained for the 18 test words were Q_Y = 0.60, $Q_{\alpha=1.0}$ = 1.12, Q_T = 4.53 and $Q_{\alpha=0.875}$ = 4.60. All these values are greater than zero, indicating that our method, like the methods of Yoo *et al.* and Tantibundhit *et al.*, emphasizes formant onsets and offsets relative to steady-state segments of formants. Q_T is greater than Q_Y, which indicates that Tantibundhit's method de-emphasizes steady formant activity more than Yoo's method.

4.3.2 Comparison of Modified Speech to Illustrate Emphasis of Formant Onset and Offset

Index Q was applied to modified speech obtained using Yoo's method, Tantibundhit's method, our method with α = 1.0 (to match Yoo's modified speech), our method with α = 0.875 (to match Tantibundhit's modified speech) and processed speech obtained using the methods of Villchur, Skowronski and Gordon-Salant. Following Yoo *et al.* and Tantibundhit *et al.*, an enhancement factor value of β = 12 was used for both our modified speech signals. Creation of modified speech by adding amplified transient speech to original speech results in modified

speech that includes steady-state vowel segments and low frequency energy like original speech but also emphasizes onset and offset of formants and consonants.

The bar chart of Figure 32 shows the average values of index Q and standard error bars for the different versions of modified and processed speech for the 18 test words. Values of Q for all modified/processed speech are greater than zero, indicating that all methods emphasize the onset and offset of formants. Tantibundhit's modified speech and our modified speech to match his modified speech have the highest average values of Q, which was expected because both speech signals are created using transient speech that strongly emphasizes onset and offset of formants. The processed speech of Skowronski and Gordon-Salant have the lowest average values of Q, which was also expected because these processed speech signals were intended to specifically emphasize consonants and not formant onsets and offsets.

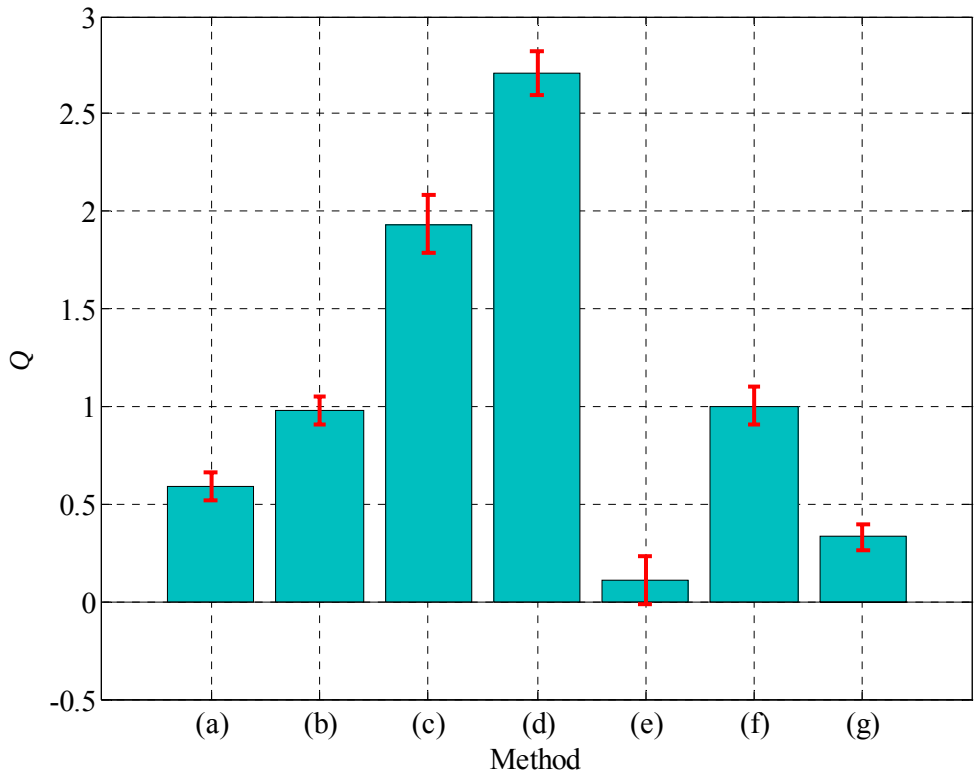

Figure 32: Average values and standard error of index Q for (a) Yoo's modified speech, (b) Our modified speech that is an estimate of Yoo's (α = 1.000), (c) Tantibundhit's modified speech, (d) Our modified speech that is an estimate of Tantibundhit's (α = 0.875), (e) Skowronski's processed speech, (f) Villchur's processed speech, and (g) Gordon-Salant's processed speech. Index Q is used to compare the relative emphasis of onset and offset of formants obtained with the different methods.

4.3.3 Comparison of Modified/Processed Speech to Illustrate Emphasis of Consonants

To evaluate the extent to which our algorithm emphasizes consonants relative to vowels and to compare this emphasis to that obtained using the methods of Yoo *et al.*, Tantibundhit *et al.*,

Skowronski *et al.*, Villchur and Gordon-Salant, index R was applied to these speech signals. The bar chart of Figure 33 show average values and the standard error bars of index R for the different versions of modified and processed speech.

The average value of R for our modified speech that is an estimate of Tantibundhit's modified speech, like the latter, is greater than zero, indicating that both modified speech signals emphasize consonants relative to vowels. Our modified speech to match Yoo's modified speech, together with Yoo's modified speech, showed a de-emphasis of consonants ($R < 0$). Speech processed using the methods of Skowronski *et al.*, Villchur and Gordon-Salant also showed emphasis of consonants ($R > 0$), with the greatest emphasis obtained by Gordon-Salant. The high average value of R and the narrow standard error interval for speech processed using the method of Gordon-Salant are due to the fact that the consonant-vowel boundaries used for her method and boundaries used for the computation of index R are matched very closely, since they are both manual processes performed by one person (the author).

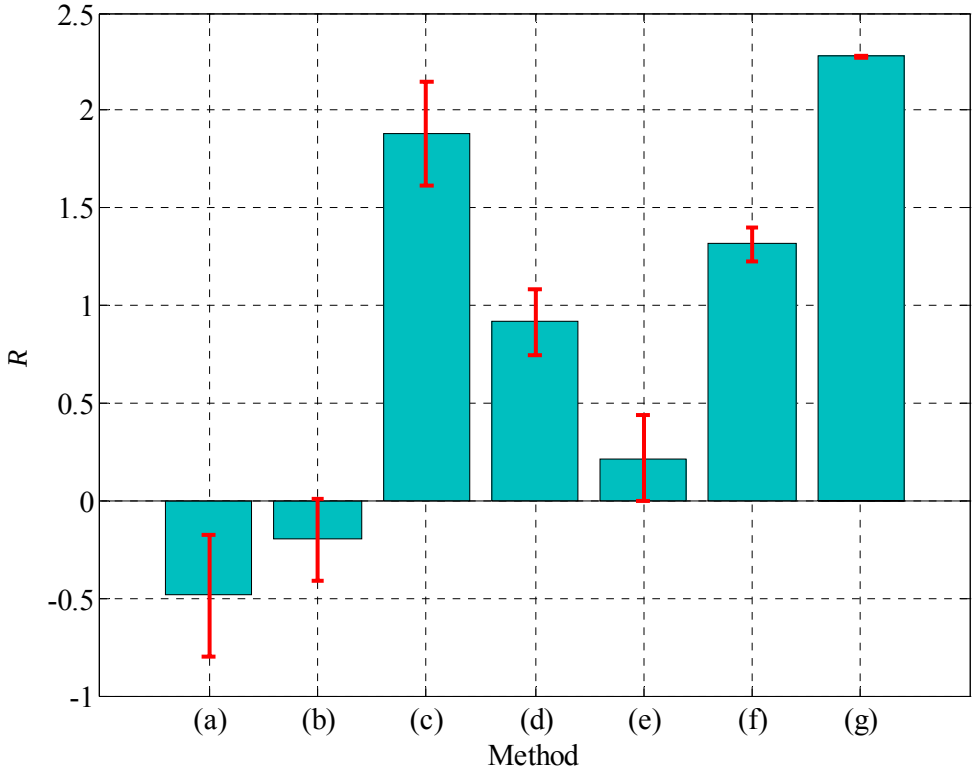

Figure 33: Average values and standard error bars of index R for (a) Yoo's modified speech, (b) our modified speech that is an estimate of Yoo's, (c) Tantibundhit's modified speech, (d) Our modified speech that is an estimate of Tantibundhit's, (e) Skowronski's processed speech, (f) Villchur's processed speech and (g) Gordon-Salant's processed speech. Index R is used to compare the relative emphasis of consonants obtained with the different methods.

4.4 SUMMARY

Three indices that can be used to compare speech signals processed using different methods were described. Index $P(x, y)$ is used to evaluate a single region of speech. A large value of index

$P(x, y)$ indicates an increased emphasis of the particular region of interest by a speech modification/processing method. Index R was used to characterize the extent to which a processed/modified speech signal emphasizes consonants relative to vowels. If a processed speech signal emphasizes consonants, $R > 0$ and if it de-emphasizes consonants, $R < 0$. Index Q was used to characterize the extent to which a processed/modified speech signal emphasizes onsets and offsets of formants relative to steady-state segments of formants. If a processed speech signal emphasizes onsets and offsets of formants, $Q > 0$ and if it de-emphasizes onsets and offsets of formants, $Q < 0$.

Examples of transient speech signals were presented to illustrate transient speech itself and to illustrate the effects of different weight parameter values and the unvoiced speech booster on the transient speech. Examples of transient speech were also presented to demonstrate the difference between transient speech signals obtained using the MFCC transitivity function and the STE transitivity function. For this demonstration a sentence with several diphthongs was used instead of a word because the differences between the two transient signals are more evident in speech material with diphthongs. The transient speech extracted using the MFCC-transitivity function is sensitive to changes in frequency that are not accompanied by changes in energy or amplitude. Transient speech extracted using the MFCC-transitivity function is also more intelligible than transient speech extracted using the STE-transitivity function.

We showed that our algorithm can extract transient speech signals that are similar to both Yoo's and Tantibundhit's transient speech components by adjusting the weight parameter. We also compared our modified speech to the modified speech signals of Yoo *et al.* and Tantibundhit *et al.*, and the processed speech signals of Skowronski *et al.*, Villchur and Gordon-Salant [2], [4],

[5], [47], [52] and showed that the relative emphasis of consonants provided by our modified speech can be increased by use of the unvoiced speech booster.

5.0 PSYCHOACOUSTIC EVALUATIONS AND SELECTION OF ALGORITHM PARAMETERS

This chapter describes procedures used to select parameters for the MFCC-based transient extraction and speech modification algorithm and presents results for the evaluation of this algorithm. The transient extraction algorithm includes a pre-processing stage to reduce the energy of the first formant and a weight parameter α to control the amount of quasi-steady-state energy in the transient speech signal. The creation of modified speech using the extracted transient speech involves use of an enhancement factor β. All of these parameters (pre-processing filter, α and β) influence the intelligibility of the modified speech signal. The algorithm also includes an unvoiced speech booster for increasing the energy of unvoiced speech. Psycho-acoustic experiments with the modified rhyme test (MRT) were used to select these parameters and to evaluate the intelligibility of modified speech with and without the unvoiced speech booster.

The best value for the parameters were selected one at a time using experiments in the order pre-processing filter, weight parameter, enhancement factor, with the later experiments using the best parameter values from the earlier experiments. The differences between parameter values that were observed in these experiments were small, and selection decisions were based on the highest percent correct values. Standard errors are presented with the data to show variability of the data.

5.1 METHODS FOR PSYCHOACOUSTIC EVALUATIONS

The modified rhyme test protocol software used by Yoo *et al.* [1] [2], developed from House *et al.* [12] [13] and Mackersie and Levitt [14], was used to measure the intelligibilities of original and modified speech. This test has been constructed to support twelve stimulus conditions, where a condition is a stimulus treatment (e.g. original speech, highpass filtered speech, modified speech) delivered at a given signal-to-noise ratio (SNR). Different stimulus conditions were used for the selection of each algorithm parameter and for the evaluation of the algorithm, as will be explained. A description of the modified rhyme test protocol follows.

Volunteer subjects with negative otological histories and hearing sensitivity of 15 dB HL or better by conventional audiometry (250-8000 Hz) were tested following an experimental protocol approved by the University of Pittsburgh Institutional Review Board. The modified rhyme test list includes 50 sets of monosyllabic rhyming words, 25 of which differ in the initial consonant and 25 differ in the final consonant. Each set consisted of six rhyming words with an interval between words of 0.25 s.

Subjects sat in a sound-attenuated booth and listened to sets of test words delivered monaurally in background noise through TDH-39 supra-aural headphones. The background noise used for all experiments conducted for this study was speech-weighted noise. The sound pressure spectrum level of speech-weighted noise is constant from 100 Hz to 1000 Hz and decreases at a rate of 12 dB/octave from 1000 to 5513 Hz [85]. Speech-weighted noise is effective at interfering with the recognition of speech sounds because it approximates the long-term sound pressure spectrum level of speech.

At the beginning of each trial, a target word appeared on the computer monitor and remained until all six rhyming words in the set were presented. Subjects were asked to identify

the target word from the set by pressing a mouse button as soon as they heard the target word. The subjects did not have a chance to hear the test words again or change their answer.

The test procedure included a training session and a testing session. The training session was used to familiarize the test subjects with the test and the test stimuli. The training session included 12 trials. The first 6 trials were presented without noise and the last 6 trials were presented at various noise levels. All the different speech stimuli for a given experiment were presented in randomized order, with and without noise.

The testing session included 300 trials – 25 trials at each of 12 stimulus conditions. The target words were randomly chosen from the modified rhyme test list. Each word appeared only once as a target word. The noise was presented for 1.83 sec. and windowed using a Tukey window to create a smooth onset and offset. The window rise and fall times were 0.25 sec. The orders of presentation of the speech stimuli and noise levels were randomized, but the noise level was kept the same for a given trial.

Test administration was computerized using MATLAB software (MathWorks, Inc.). Subject responses were recorded by the computer, test results were saved, and mean recognition scores for each subject and each noise condition were computed. The test procedures were monitored by a skilled examiner under supervision of a certified clinical audiologist.

5.2 SELECTION OF ALGORITHM PARAMETERS

Experiments performed to select the best parameters values for the transient extraction method are described here. Results of these experiments are also presented.

5.2.1 Selection of Pre-Processing Filter

In the development of the version of the algorithm that utilized the STE-transitivity function, we used a 50^{th} order FIR highpass filter with a cutoff frequency of 700 Hz, hereafter referred to as HPF_0, for pre-processing. This filter removed most of the energy associated with first formant. Without pre-processing, the extracted transient speech is dominated by low energy transitions and appeared not to contribute to speech intelligibility improvement during informal listening tests. Although HPF_0 provided additional speech intelligibility improvement to our modified speech, its parameters were not selected to maximize the intelligibility of modified speech. Yoo selected parameter for HPF_0 to eliminate the first formant and increase the processing efficiency of his algorithm. A different approach to reducing the low frequency energy of speech would be to use a filter, which will be referred to as HPF_S, that approximates the inverse of the long-term average spectrum of speech. HPF_S would act as an equalizer and distribute the energy of the speech approximately equally across the entire frequency range for most speech material. Byrne *et al.* measured the long-term average speech spectrum for 13 languages including English [86]. The combined speech spectrum for male and female speakers is shown in Figure 34. The magnitude response of filters HPF_S and HPF_0 are also shown in Figure 34. HPF_S was formed by inverting the long-term average spectrum of speech and then replacing the magnitude response below a frequency of 400 Hz with a linear polynomial to create a gradual magnitude increase from 0 to 400 Hz.

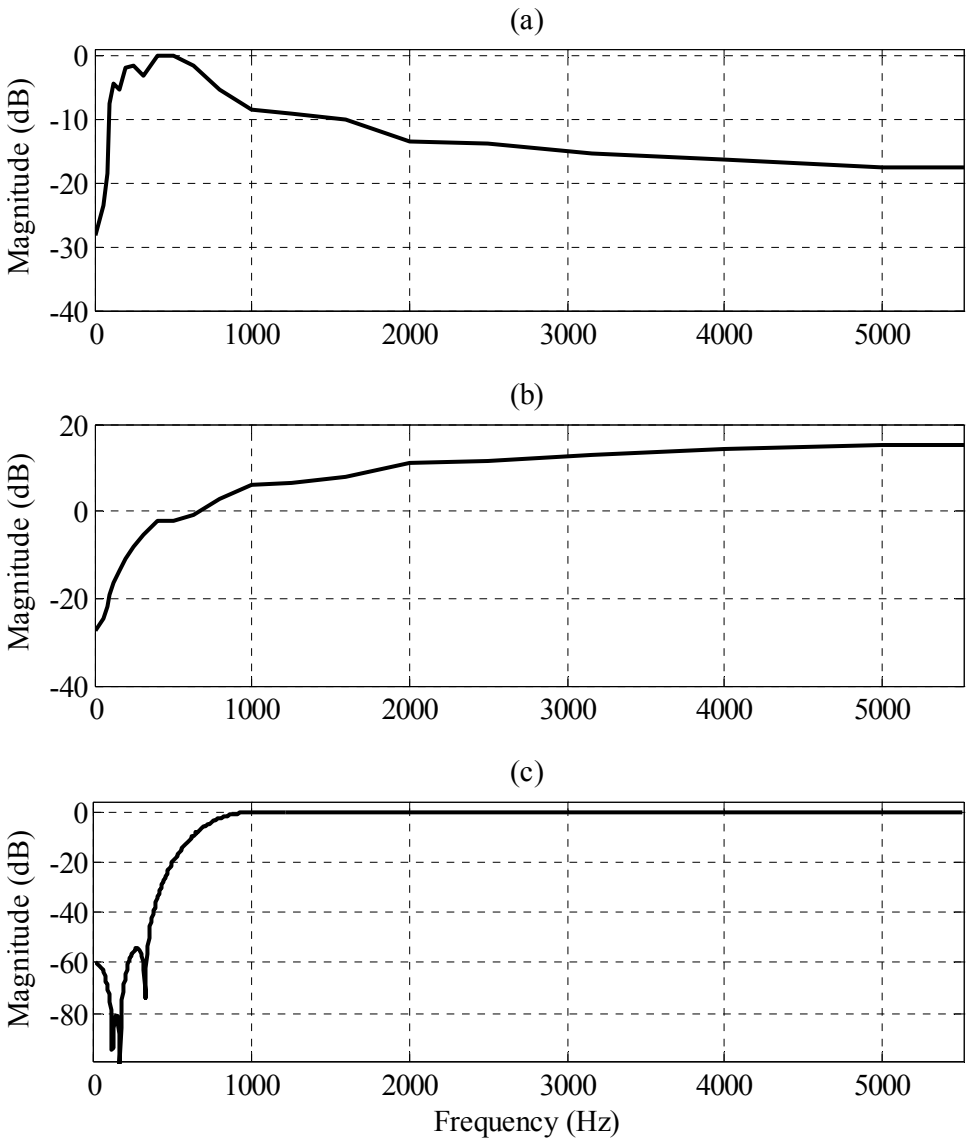

Figure 34: (a) Long-term average speech spectrum [86] (b) magnitude response of HPF_S and (c) magnitude response of HPF_0.

Use of HPF_S for pre-processing was compared to use of HPF_0 using measurements of the intelligibility of modified speech. In addition, these two pre-processing filters were compared to the condition where a pre-processing filter is not applied. The weight parameter and enhancement factor values used for this experiment were $\alpha = 0.9$ and $\beta = 12$. These parameters values were selected based on informal listening tests. The 4 stimulus treatments evaluated, using the MRT at SNRs of -25, -15 and -5 dB, are

- Original speech
- Modified speech without pre-processing
- Modified speech with pre-processing using HPF_0
- Modified speech with pre-processing using HPF_S

The average percent correct scores for nine subjects for the experiment are shown in Figure 35 and the standard errors are presented in Table 3. Of the two pre-processing filters, HPF_0 results in the most intelligible modified speech at all three SNRs, while the condition with no pre-processing filter results in the least intelligible modified speech. HPF_S provides about the same intelligibility improvement over original speech as HPF_0 at -25 dB SNR, but provides less improvement at -15 and -5 dB SNR. Based on these results, HPF_0 was selected as the best pre-processing filter for the transient extraction algorithm.

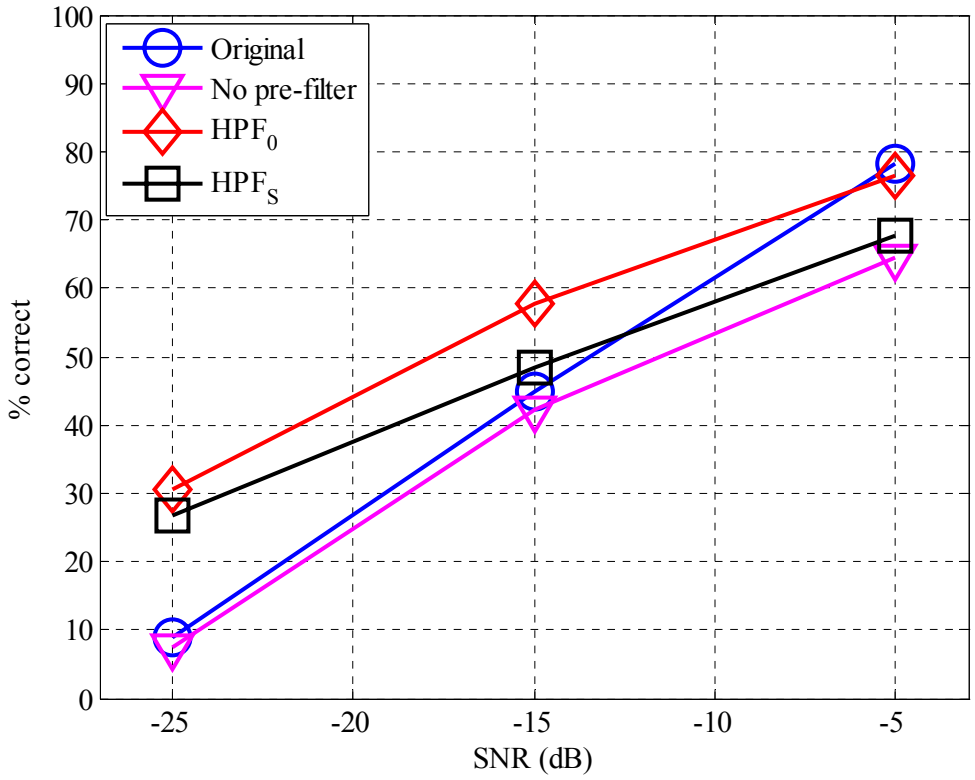

Figure 35: Modified rhyme test average percent correct scores for original speech and speech modified using no pre-filter, HPF_0 and HPF_S for pre-processing.

Table 3: Average percent correct scores and standard error original speech and speech modified using no pre-filter, HPF_0 and HPF_S for pre-processing

Speech type	SNR (dB)	Average percent correct	Standard error
Original	-25	8.89	5.07
	-15	44.89	2.38
	-5	78.22	3.20
Modified speech, no prefilter	-25	7.56	2.35
	-15	42.22	2.99
	-5	64.44	3.68
Modified speech, HPF_0 prefilter	-25	30.67	7.15
	-15	57.78	4.58
	-5	76.44	2.53
Modified speech, HPF_S prefilter	-25	26.67	4.99
	-15	48.44	2.53
	-5	67.56	2.44

5.2.2 Selection of Weight Parameter

The best weight parameter for the algorithm that utilizes HPF_0 for pre-processing was selected using the MRT. Weight parameter values of α = 0.85, 0.95 and 1.0 were evaluated with low (β = 18) and high (β = 36) enhancement factor values at SNRs of -20 and -5 dB. An enhancement

factor of 18 adds the same amount of energy of transient speech to original speech that Yoo *et al.* added when they formed their modified speech [2]. Values of $\alpha < 0.85$ were not evaluated because they result in a transient speech signal with greatly diminished speech quality. The 6 weight parameter/enhancement factor pairs (stimulus treatments) that were evaluated, at SNRs of -20 and -5, for modified speech in this experiment are:

- $\alpha = 0.85,\ \beta = 18$
- $\alpha = 0.85,\ \beta = 36$
- $\alpha = 0.95,\ \beta = 18$
- $\alpha = 0.95,\ \beta = 36$
- $\alpha = 1.0,\ \beta = 18$
- $\alpha = 1.0,\ \beta = 35$

The bar chart of Figure 36 shows the average percent correct word scores and standard error bars at -5 dB and -20 dB SNR and two values of β as a function of the weight parameter, α. The percent correct scores increase with increasing α at both -5 and -20 dB SNR. Based on these results, a weight parameter of α = 1.0 was selected for use in the transient extraction algorithm. As mentioned earlier, a weight parameter of α = 1.0 is equivalent to not applying thresholding to the transitivity function in the transient extraction algorithm. An enhancement factor value of β = 18 consistently resulted in higher percent correct scores than an enhancement factor of β = 36.

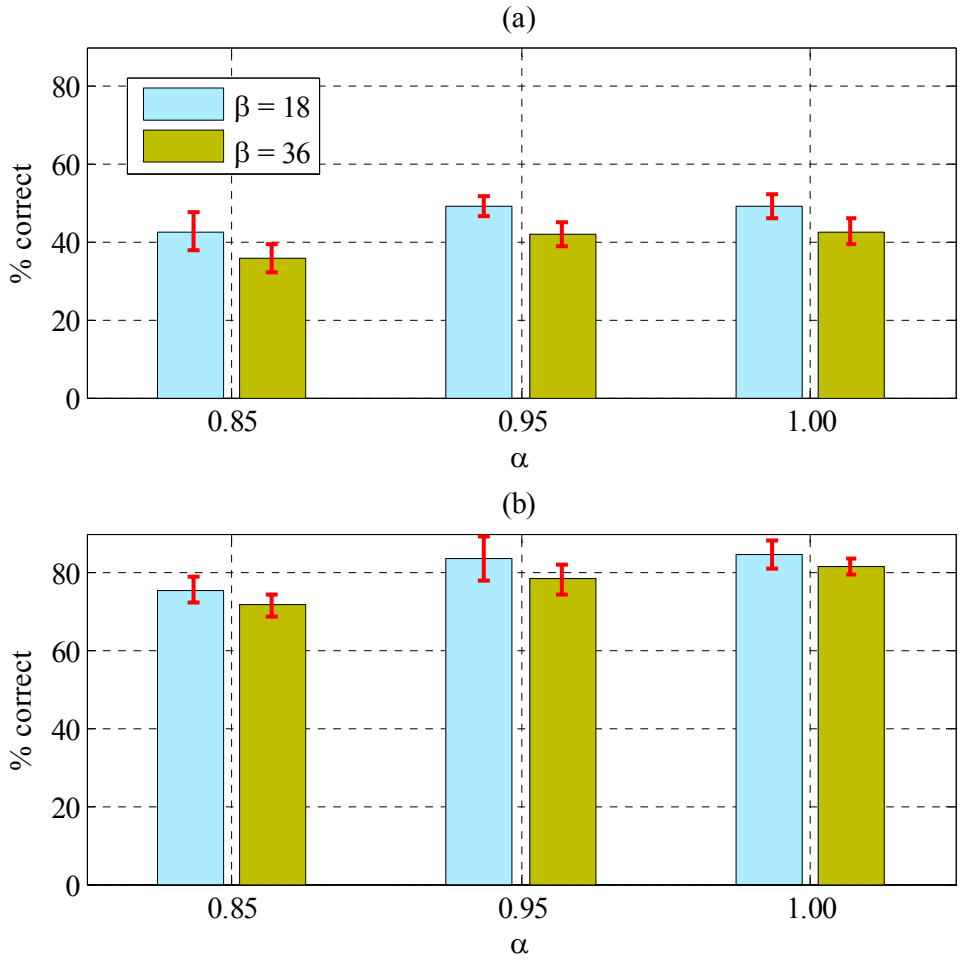

Figure 36: Average percent correct word scores and standard error bars for the modified rhyme test experiment to select the best weight parameter, α at (a) -20 dB and (b) -5 dB SNR.

5.2.3 Selection of Enhancement Factor

Having selected the best pre-processing filter and weight parameter, the best enhancement factor was selected in this experiment. Five values of β $\{\beta = 6, 12, 18, 24 \text{ and } 30\}$ were evaluated at SNRs of -20 and -5 dB using the MRT. Intelligibility measurements for original speech were also obtained in this experiment.

The bar chart of Figure 37 shows the mean percent word correct scores and standard error bar at -5 and -20 dB SNR as a function of β. A weight parameter of α = 1.0 was used to extract transient speech. At both SNRs, the highest percent correct responses were obtained with enhancement factors of β = 12 and β = 24. These two values produced nearly identical percent correct responses. A value of β = 24 was selected for use with the algorithm because it provides greater emphasis of transient activity and our objective was to improve intelligibility by emphasizing transient speech.

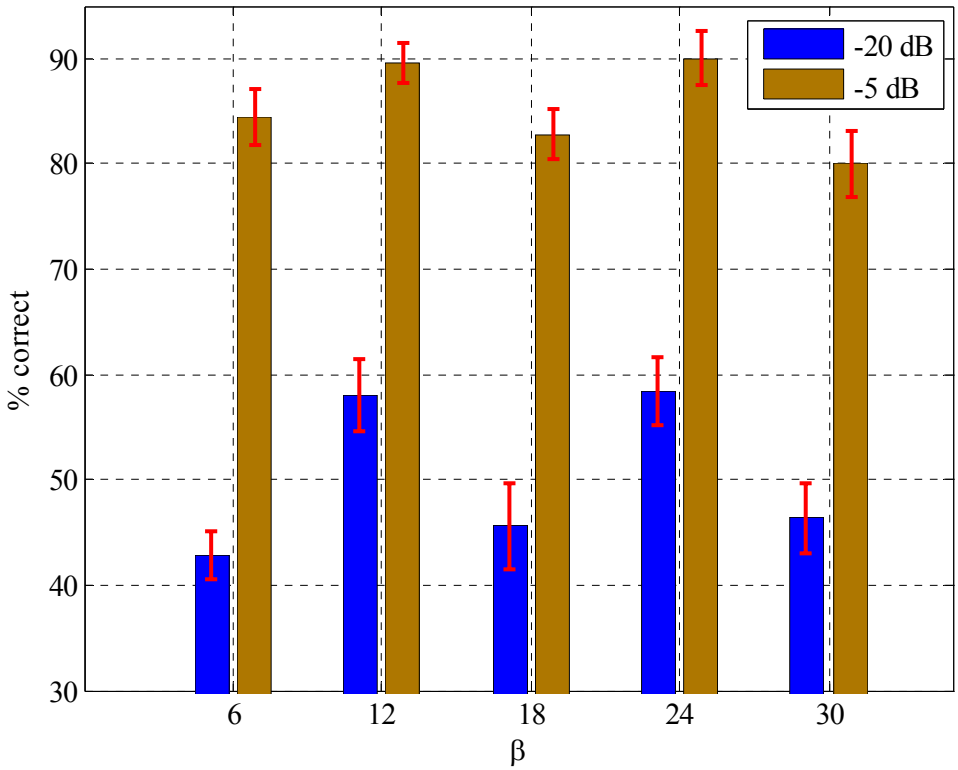

Figure 37: Mean percent word correct scores and standard error for experiment to select the best enhancement factor (β).

5.3 EVALUATION OF ALGORITHM

Having selected the pre-processing filter (HPF_0), weight parameter (α = 1.0) and enhancement factor (β = 24) for the MFCC-based algorithm, the intelligibility of modified speech created using these parameters with and without the unvoiced speech booster (USB) was measured. The intelligibility of original speech was also measured to provide a baseline for comparison. The three stimulus treatments evaluated, at SNRs of -20, -15, -10 and -5 dB in this experiment are:

- Original speech
- Speech modified by transient emphasis
- Speech modified by transient emphasis and unvoiced speech booster

Figure 38 shows mean percent word correct scores and the standard error bars for original speech and modified speech created with and without the unvoiced speech booster. The intelligibility of both modified speech (with and without the unvoiced speech booster) is higher than the intelligibility of original speech at all noise levels evaluated (-20 to -5 dB SNR) with the largest improvements occurring at lower SNRs (-20 and -15 dB). The standard error intervals of both modified speech forms do not overlap with the standard error intervals of original speech at -20 and -15 dB, suggesting that the differences in scores at these noise levels are significant. The standard error bars of the two forms of modified speech signals overlap at all SNRs evaluated, suggesting that the scores for modified speech signals are not different. Based on these results, the algorithm with the unvoiced speech booster does not provide additional speech intelligibility enhancement over the algorithm without the unvoiced speech booster.

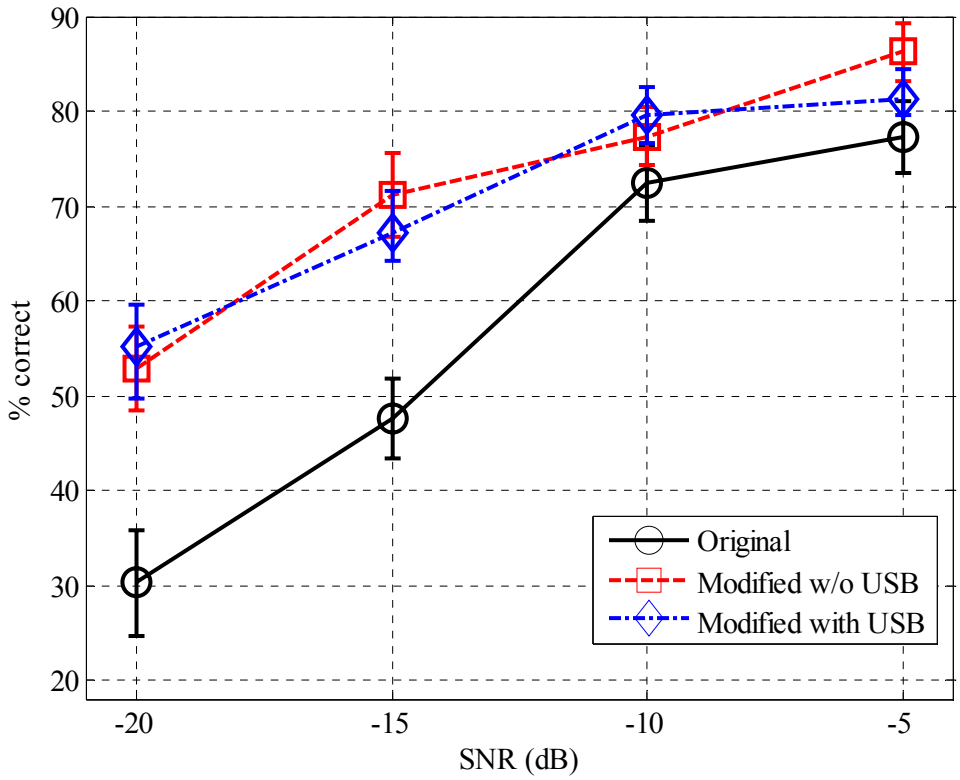

Figure 38: Mean percent word correct scores and standard error for original and modified speech created with and without the unvoiced speech booster (USB).

In Figure 39, the differences in mean percent correct word scores (percent correct scores for modified speech minus percent correct scores for original speech) obtained using the MFCC-based algorithm without USB are compared to improvement in intelligibility obtained using the STE-based method. The standard errors for these differences are presented in Table 4. The scores presented were obtained from two different experiments and are for SNRs (-15 and -5 dB) that were common in the two experiments conducted to evaluate the STE- and MFCC-based algorithms. Differences in percent word correct scores, while small, suggest that modified speech created using the MFCC-transitivity function is more intelligible than transient speech created using the STE-transitivity function.

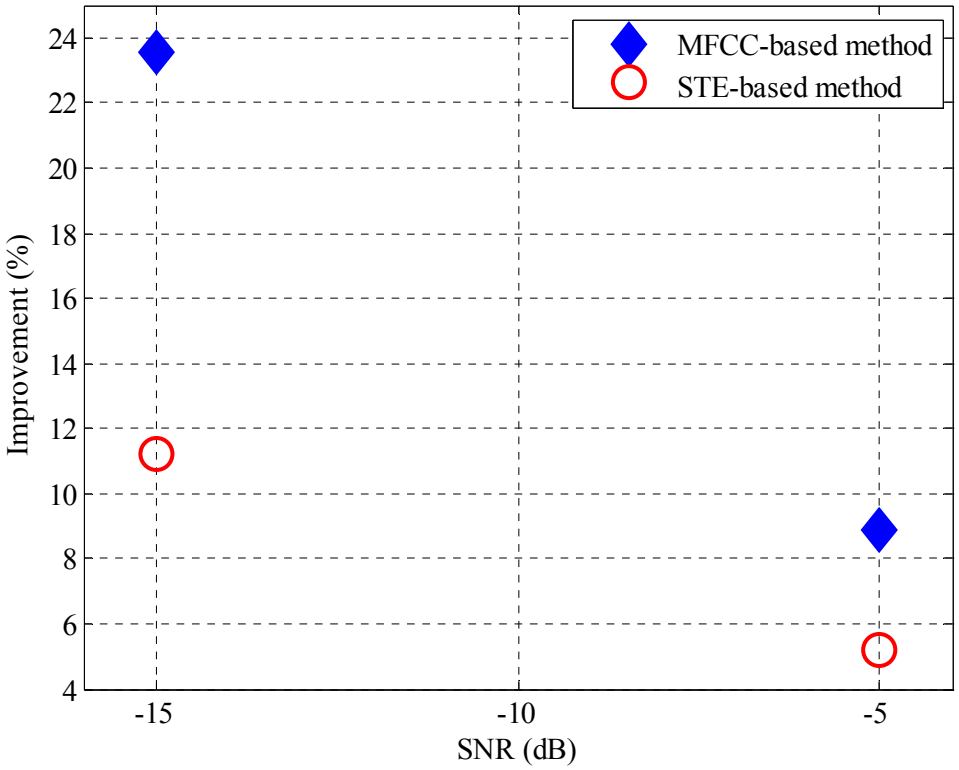

Figure 39: Comparison of intelligibility improvements obtained with MFCC-based algorithm to improvements obtained with STE-based algorithm.

Table 4: Differences in mean percent correct word scores and standard errors for the differences for MFCC-based algorithm and STE-based algorithm.

Method	SNR (dB)	Mean difference	Standard error
MFCC-based	-15	23.5500	5.0100
	-5	8.8900	3.7000
STE-based	-15	11.20	8.6178
	-5	5.20	7.0883

5.4 SUMMARY

Psycho-acoustic experiments to select the best parameters (pre-processing filter, weight parameter and enhancement factor) and to evaluate the speech intelligibility enhancement algorithm using these parameters were described. Pre-processing using HPF_0, a 50^{th} order highpass filter with a cutoff frequency of 700 Hz, resulted in modified speech with the highest intelligibility and as such HPF_0 was selected as the best pre-processing filter. In the experiment to select the best weight parameter, speech intelligibility increased with increase of the weight parameter and the highest intelligibility was obtained with a weight parameter of $\alpha = 1.0$ (the largest effective value α can have). Consequently, a values of $\alpha = 1.0$ was considered the best.

In the experiment to select the best enhancement factor, values of $\beta = 12$ and $\beta = 24$ resulted in the highest intelligibility. There was an unexpected dip in intelligibility at both -5 and -20 dB SNR when $\beta = 18$. The consistency of the dip at both SNR levels suggests that the effect

is not just a statistical anomaly and is discussed further in Chapter 6. We selected $\beta = 24$ to use for the final comparisons because that value results in a greater emphasis of the transient component.

Results for the experiment to evaluate the algorithm showed that modified speech created using the algorithm with the MFCC-transitivity function was more intelligible in noise than original speech, especially at high noise levels (-20 and -15 dB SNR) where intelligibility improvement is needed. The results also showed that modified speech created using the algorithm with the MFCC-transitivity function was more intelligible than modified speech created using the algorithm with the STE-transitivity function.

In Figure 40, the differences in mean percent correct word scores obtained with the MFCC-based algorithm are compared to results obtained by Yoo *et al.* and Tantibundhit *et al.*, who also modified speech using transient speech and evaluated intelligibility using the MRT [2] [4]. The standard errors for the differences in mean percent correct word scores for the different versions of modified speech are presented in Table 5. The differences in percent word correct scores show that the intelligibility of modified speech created using the MFCC-transitivity function closely matches the intelligibility of Yoo's modified speech and are better than the intelligibility of Tantibundhit's modified speech. Additionally, our method can be implemented to run in real-time while the methods of Yoo *et al.* and Tantibundhit *et al.* cannot. These factors make our method attractive for applications that require enhancement of speech intelligibility in noisy environment, like mobile communication.

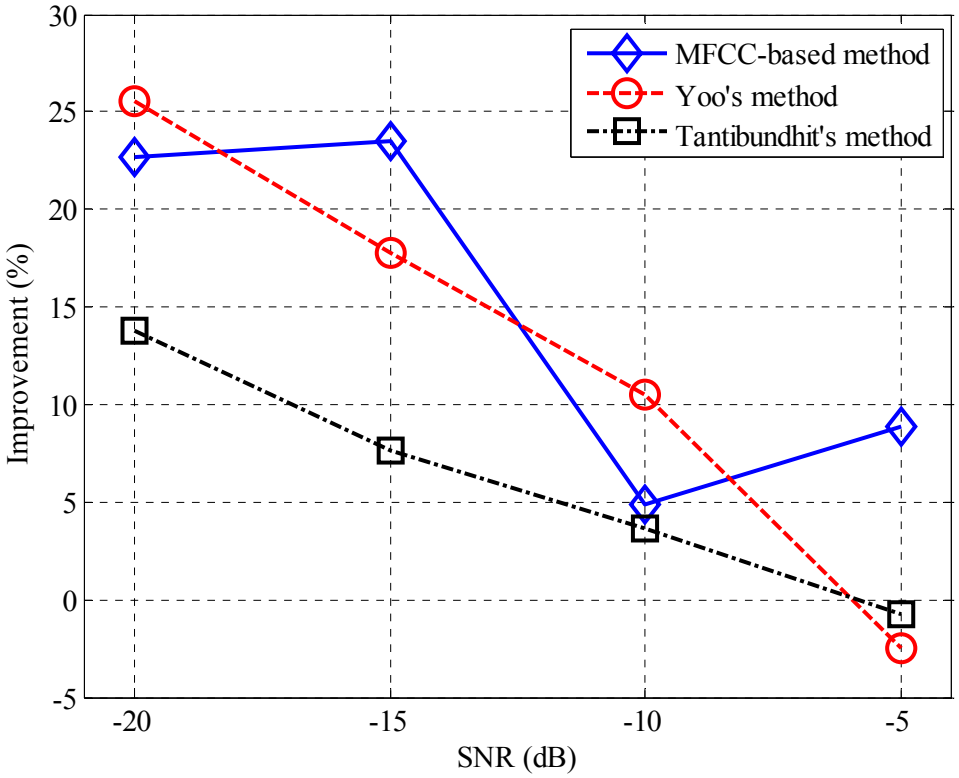

Figure 40: Comparison of intelligibility improvements.

Table 5: Mean difference in percent correct word scores and standard errors for the differences for our MFCC-based modified speech, Yoo's modified speech and Tantibundhit's modified speech.

Method	SNR (dB)	Mean difference	Standard error
Our MFCC-based method	-20	22.6690	6.8300
	-15	23.5500	5.0100
	-10	4.8900	3.9900
	-5	8.8900	3.7000
Yoo's method	-20	25.5000	2.2312
	-15	17.8000	3.6784
	-10	10.5000	5.6081
	-5	-2.5000	1.8995
Tantibundhit's method	-20	13.8200	6.2986
	-15	7.6400	3.3890
	-10	3.6400	4.8091
	-5	-0.7300	4.3749

6.0 DISCUSSION

We introduced an algorithm for extraction of transient speech that can be implemented to run in real-time. The algorithm decomposes a speech signal into several sequences of wavelet coefficients using the forward wavelet packet transform, characterizes the rate of change and adjusts the wavelet coefficients based on how fast they are changing and synthesizes a transient speech signal using the inverse wavelet packet transform. Transient speech was used to create modified speech by amplifying and adding it to the original speech and then adjusting the energy of the modified speech signal so that it equals that of original speech. Wavelets provides subband decomposition that allows the detection of transients occurring at different times in different frequency bands and reduces the amount of quasi-steady-state activity that would be identified as transient. For the characterization of the rate of change of wavelet coefficients, a function that we called the transitivity function was developed. This function is large and positive when the wavelet coefficients have a rapidly changing frequency or amplitude and near zero when the wavelet coefficients are in steady-state.

Two definitions for the transitivity function, one based on the short-time energy (STE) of wavelet packet coefficients and the other on Mel-frequency cepstral coefficients (MFCC) of wavelet packet coefficients, were formulated. Although MFCCs are traditionally applied directly to speech, applying them to wavelet coefficients is reasonable because the wavelet filters of most packets are not strictly narrowband as they have side lopes with significant amplitudes. The side

lopes are required to satisfy the perfect reconstruction property of wavelets. As examples, Figure 41 shows the magnitude frequency response of the wavelet function for 6 of the 16 packets (packets 2, 5, 6, 9, 10 and 13) for the Daubechies-18 mother wavelet. The differences in magnitude between the main lope and the biggest side lope for these packets are 23.5 dB for packet 2, 23.5 dB for packet 5, 10.7 dB for packet 6, 10.7 dB for packet 9, 25.6 dB for packet 10 and 21.1 dB for packet 13. These side lopes pass enough energy of speech for the wavelet coefficients to be considered not strictly narrowband. The side lopes result in wavelet coefficients that retain some speech-like character and are readily recognized as speech. Applying MFCCs to wavelet coefficients is also reasonable because use of MFCC results in transient and modified speech signals with higher intelligibility than use of the short-time energy.

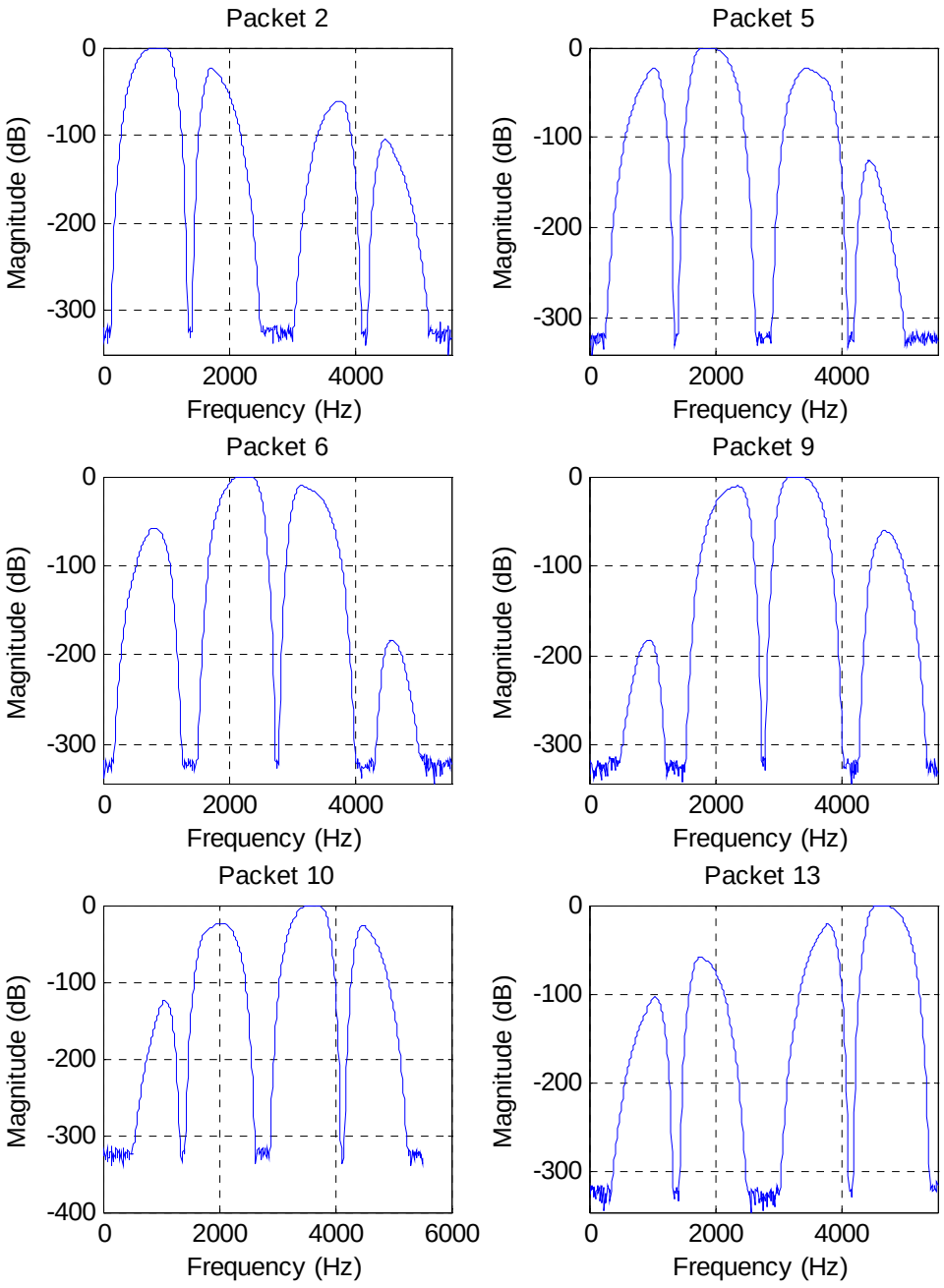

Figure 41: Illustration of wavelet filters (frequency magnitude response of packet wavelet function) for 6 of 16 packets for the Daubechies-18 mother wavelet.

The transient extraction method was applied to a wide range of speech material, and transient and modified speech obtained using the algorithm were compared to transient and modified speech obtained using the algorithms of Yoo *et al.* and Tantibundhit *et al.* and to processed speech obtained using the methods of Villchur, Skowronski *et al.* and Gordon-Salant. To facilitate the comparisons, three indices (P, R and Q) were developed. Index P was used to compare the effect of a speech modification/processing method on a particular region of speech. Index R was developed to characterize the extent to which consonants are emphasized relative to vowels in a speech signal. Index Q was developed to characterize the extent to which the onsets and offsets of formants are emphasized compared to steady segments in a speech signal. These indices were very useful in the comparisons because they quantify differences in speech signals that are difficult to show using spectrograms, spectra and time-domain waveforms.

A disadvantage of these indices, especially R and Q, is that manual placement of time-frequency masks is required for their computation. Also, only words with relatively steady vowel segments and with 2nd, 3rd and 4th formants that were roughly coincident in time were used in the computation of index Q. The placement of masks is time-consuming and the restriction on choice of words limits the material that can be used. However, using an automated method to place the masks and using words that may produce ambiguous identification of formant onsets and offsets could introduce errors to a method (computation of P, R and Q) that is intended to evaluate another method (the transient extraction algorithm). Manual placement of masks was used to minimize these errors.

Comparison of transient speech extracted using our algorithm to transient speech components of Yoo *et al.* and Tantibundhit *et al.* showed that our algorithm can extract transient

speech signals that are similar to both Yoo's and Tantibundhit's transient speech components by adjusting the weight parameter. The comparisons also showed that our transient speech emphasizes onsets and offsets of formants similar to the transient components of Yoo *et al.* and Tantibundhit *et al.*.

Comparison of our modified speech to the modified speech signals of Yoo *et al.* and Tantibundhit *et al.*, and the processed speech signals of Skowronski *et al.*, Villchur and Gordon-Salant showed that the relative emphasis of consonants provided by our modified speech can be increased by use of the unvoiced speech booster.

The transient extraction algorithm includes a pre-processing stage to reduce the energy of the first formant and a weight parameter α to control the amount of quasi-steady-state energy in the transient speech signal. The creation of modified speech using the extracted transient speech involves use of an enhancement factor β. All these parameters (pre-processing filter, α and β) influence the intelligibility of the modified speech signal. The algorithm can also include an unvoiced speech booster for increasing the energy of unvoiced speech. Psycho-acoustic experiments with the modified rhyme test (MRT) were used to select these parameters and to evaluate the intelligibility of modified speech with and without the unvoiced speech booster. The major purpose of these experiments was to evaluate the sensitivity of the algorithm to these parameters and to be sure that particularly disadvantageous parameter values were not used.

The best value for the parameters were selected one at a time using experiments in the order pre-processing filter, weight parameter, enhancement factor, with the later experiments using the best parameter values from the earlier experiments. This design of experiments assumes that the selection of the best value for one parameter has minimal influence on of the selection of the best value for another parameter and does not identify globally optimal

parameter values. Without this assumption, selection of the best parameters would have required testing a very large combination of parameters. Since the MRT protocol supports testing of only 12 stimulus conditions for each subject, the time required to test each subject is 1.5 hours, and there was limited availability of subjects and time, conducting such an extensive range of experiments would be prohibitive. Also, the algorithm performance did not seem to depend critically on parameter values, i.e. changes in performance with parameter values were gradual and extreme precision in selecting values does not appear to be necessary. The parameters values determined as the best (a 50^{th} order highpass filter with a cutoff frequency of 700 Hz for pre-processing (HPF_0), weight parameter of $\alpha = 1.0$ and an enhancement factor of $\beta = 24$), while not optimal, are good and we do not expect a major improvement in intelligibility (of say > 10 %) with different values. Some fine tuning may provide a few percent improvements in speech intelligibility, but we would be surprised to see much more.

Incorporation of the unvoiced speech booster to the algorithm was evaluated. Although emphasis of consonants can alone improve speech intelligibility as was shown by [9] [10] [21], this process does not result in further intelligibility improvements over emphasis of transient speech.

Intelligibility improvements over original speech obtained with the MFCC transitivity function were greater than the improvements obtained with the STE-transitivity function. This suggests that the ability of MFCC transitivity function to capture frequency changes with constant energy (e.g. diphthongs) is important for transient extraction and for enhancement of speech intelligibility.

Evaluation of the final form of the algorithm, which uses the best parameter values, showed that the intelligibility of modified speech was greater than the intelligibility of original

speech, especially at high noise levels (-20 and -15 dB SNR) where intelligibility improvement is needed. This suggests that emphasis of transient speech can enhance speech in noise. This enhancement method can be applied to any speech communication system where the speaker is in a noise free environment and the listener is in a noisy environment. Such scenarios are encountered during communication between control tower and ground support at an airport, during battle field communications between command center and soldiers, in public address systems, while listening to AM/FM radio in a car, during cellular phone communications in a loud restaurant, etc.

Index Q, which can be interpreted as representing the amount of transient speech in a speech signal, can be related to speech intelligibility. In the experiment to select the best enhancement factor β, modified speech was obtained by adding transient speech to original speech (Equation (20)). When original speech $x[n]$ is considered as a combination of quasi-steady-state $w[n]$ and transient speech $y[n]$, i.e.

$$x[n] = w[n] + y[n] \tag{26}$$

Equation (20) can be written as

$$z[n] = \rho(w[n] + y[n] + \beta y[n]) \tag{27}$$

$$z[n] = \rho(w[n] + \mu y[n]) \tag{28}$$

where $\mu = \beta + 1$. In this form, $z[n]$ can be interpreted as a combination of quasi-stead-state and transient speech.

When $\mu = 0$, $z[n]$ (modified speech) is composed entirely of quasi-steady-state speech and when $\mu = \infty$, $z[n]$ is composed entirely of transient speech. When $\mu = 1$, $z[n]$ is composed of a balance of transient and quasi-steady-state speech that produces original speech, i.e. $z[n]$ is original speech. This interpretation of the formation of modified speech can be used to evaluate the role of transient speech on speech intelligibility and to relate index Q to intelligibility. We expect intelligibility to grow with the amount of transient speech until an "optimal" amount is reached and then to decrease, as the artifact noise reduces its quality, with further increase in the amount of transient speech. Pure transient speech has lower quality than both original and modified speech which reduces its intelligibility.

Using index Q to quantify the amount of transient speech in a signal, the increase in the amount of transient speech in a speech signal (using the list of test words described in Chapter 4) as μ increases is shown in Figure 42.

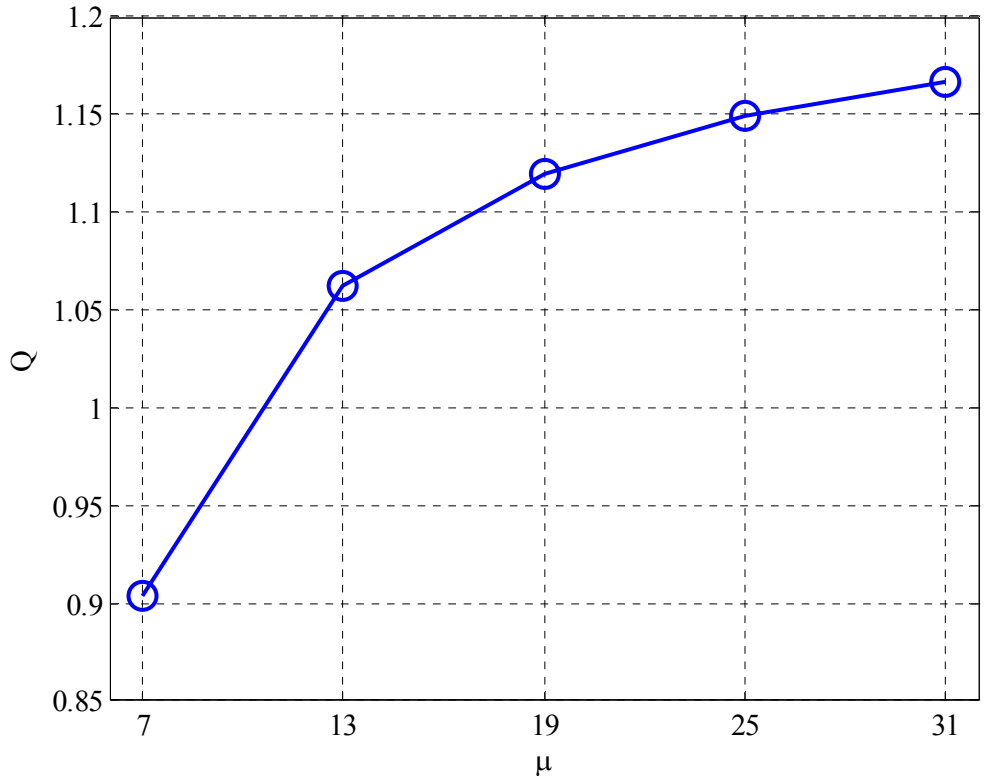

Figure 42: Index Q as a function of μ.

Figure 43 shows the relationship between intelligibility and the amount of transient speech in a speech signal (represented by Q) at -20 and -5 dB SNR. Speech intelligibility is low when the amount of transient speech in a speech signal is very low ($\mu = 7$) and when it is very high ($\mu = 31$). When the amount of transient speech in a signal is low, the benefit of transient emphasis is too small to affect speech intelligibility positively and when the amount of transient is high, artifact noise is introduced to modified speech, which reduces speech intelligibility. Values of μ greater than 31 seemed to reduce speech intelligibility during informal listening test. There was an unexpected dip in intelligibility at both -5 and -20 dB SNR when $Q = 1.12$ ($\mu = 19$). The consistency of the dip at both SNR levels and the fact that the scores obtained with this experiment (45.6 % at -20 dB and 82.8 % at -5 dB) match the scores obtained in the

prior experiment to select the best weight parameter (48.9 % at -20 dB and 84.4 % at -5 dB) using the same values of weight parameter and enhancement factor (α = 1, $\beta = \mu - 1$=18) suggests that the effect is not just a statistical anomaly.

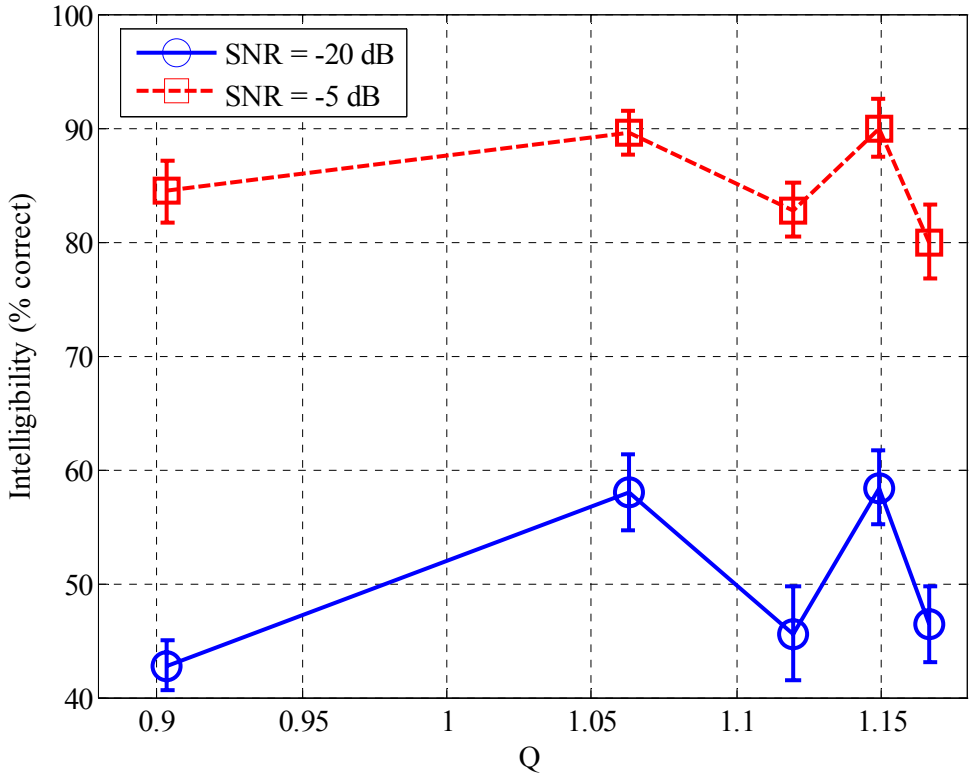

Figure 43: Speech intelligibility as a function of the amount of transient speech in a speech signal (Q)

Transient speech includes onset and offset of formants and consonants. The average value of index Q for our modified speech was greater than zero (Q = 0.98) while the average value of index R was less than zero (R = -0.2), indicating that our algorithm emphasizes onset and offset of formants relative to steady-state segments of formants and slightly de-emphasizes consonants relative to vowels. This suggests that emphasis of onsets and offsets of formants is

more important to enhancement of speech intelligibility than emphasis of consonants. This suggestion is consistent with average values of indices Q and R for the modified speech of Yoo *et al.* and Tantibundhit *et al.*. Yoo's modified speech emphasizes onset and offset of formants (Q = 0.59) and de-emphasizes consonant (R = -0.49) similar to our method. Tantibundhit's modified speech emphasizes both onset and offset of formants (Q = 1.93, R = 1.88), however, it provides less enhancement of speech intelligibility than both our modified speech and Yoo's modified speech. Further understanding of the importance of the emphasis of the different parts of speech on intelligibility may be useful in the design of speech enhancement systems and hearing aids.

7.0 CONCLUSION

Studies have shown that emphasis of transient speech can improve the intelligibility of speech in background noise, but methods to demonstrate this improvement have either identified transient speech manually or proposed algorithms that cannot be implemented to run in real-time. An algorithm for extraction of transient speech that can be implemented to run in real-time has been described. The algorithm decomposes a speech signal into several sequences of wavelet coefficients using the forward wavelet packet transform, characterizes the rate of change and adjusts the wavelet coefficients based on how fast they are changing and synthesizes a transient speech signal using the inverse wavelet packet transform. Transient speech was used to create modified speech by amplifying and adding it to the original speech and then adjusting the energy of the modified speech signal so that it equals that of original speech. Wavelets provides subband decomposition that allows the detection of transients occurring at different times in different frequency bands and reduces the amount of quasi-steady-state activity that would be identified as transient. For the characterization of the rate of change of wavelet coefficients, a function that we called the transitivity function was developed. This function is large and positive when the wavelet coefficients have a rapidly changing frequency or amplitude and near zero when the wavelet coefficients are in steady-state. Two definitions for the transitivity function, one based on the short-time energy (STE) of wavelet packet coefficients and the other on Mel-frequency cepstral coefficients (MFCC) of wavelet packet coefficients, were formulated and evaluated

experimentally. The MFCC-based transitivity function resulted in transient and modified speech with higher intelligibility than the STE-transitivity function.

To facilitate comparison of our transient and modified speech to speech processed using methods proposed by other researchers to emphasize transients, we developed three indices. The indices are used to characterize the extent to which a speech modification/processing method emphasizes (1) a particular region of speech, (2) consonants relative to, and (3) onsets and offsets of formants compared to steady formant. These indices are very useful because they quantify differences in speech signals that are difficult to show using spectrograms, spectra and time-domain waveforms.

The algorithm includes parameters (pre-processing filter, weight parameter and enhancement factor) which when varied influence the intelligibility of the extracted transient speech. The best values for these parameters were selected using psycho-acoustic testing. The incorporation of a method that automatically identifies and boosts unvoiced speech into the algorithm was evaluated and showed that this method does not result in additional speech intelligibility improvements. Measurement of speech intelligibility in background noise using psycho-acoustic experiments showed that the intelligibility of speech modified with the algorithm that utilizes any of the transitivity functions is higher than the intelligibility of original speech, especially at high noise levels (-20 and -15 dB SNR) where enhancement of intelligibility is needed. This suggests that emphasis of transient speech can enhance speech in noise. Additionally, unlike previously proposed algorithms, our algorithm extracts transient speech much more efficiently and can be implemented to run in real-time.

8.0 FUTURE RESEARCH WORK

The design of the transient extraction algorithm assumed that we had access to the original speech signal before noise was added to it. The algorithm also focuses on enhancing speech intelligibility without degrading speech quality. It is important to note that speech quality and intelligibility are different. Speech quality relates to how comfortable it is for a listener to listen to a speech utterance. The utterance does not necessarily have to convey meaning. Intelligibility relates to the ability of a speech utterance to convey meaning, that is, whether the listener can correctly identify words being spoken. The extension of the algorithm for extraction of transient speech from noisy speech and for enhancement of both speech intelligibility and speech quality should be evaluated. A possible extension that may improve speech quality is the incorporation of a method that improves speech quality such as spectral subtraction or active noise reduction to our algorithm.

Three indices were developed to facilitate the comparisons of speech signals modified or processed using different methods. Index R was developed to characterize the extent to which consonants are emphasized relative to vowels in a speech signal. Index Q was developed to characterize the extent to which the onsets and offsets of formants are emphasized compared to steady segments in a speech signal. Incorporating the two indices into one index for measuring the "transient-ness" of speech and automatic placement of time-frequency masks are interesting topic to investigate. A measure of "transient-ness" of speech can be used to provide a better

understanding of the role of transient speech in speech intelligibility. This understanding may be beneficial to the design of speech enhancement algorithms and hearing aids.

Recent studies in auditory research suggest that the outer hair cells (OHCs) implement a nonlinear active process that may play a role in the processing of noisy speech. It has been suggested that this role may be related to the processing of transient speech. Measurement of otoacoustic emissions (OAE), acoustic energy that is generally considered to be produced by OHC, provides a non-invasive method to probe OHC function. Comparison of the response of OHC to transient speech to the response to quasi-steady-state speech could give a better understanding of OHC function. A better understanding of OHC may be used to improve algorithms for extraction of transient speech for enhancement of speech intelligibility for people with normal hearing and for the design of better hearing aids for the hearing impaired. An adaptation of the ILO88 OHC probing system (David Kemp-1989), which uses clicks as stimuli, to a system that can use short duration speech for OHC probing can be useful for studying OHC responses of transient, quasi-steady-state speech and modified speech. A study of response of OHC to different speech stimuli (transient, quasi-steady-state, modified) in order to gain a better understanding of OHC could be conducted.

The transitivity function used to characterize the rate of change of wavelet coefficients, has peaks during transitions between speech sounds and hence it could be useful in phoneme (speech sound) segmentation - an important pre-processing stage in automatic speech recognition. The application of the transitivity function to phoneme segmentation could be evaluated.

Previously, we used automatic speech recognition (ASR) to evaluate the transient extraction algorithm by comparing the recognition rates of modified speech to those of original

speech. The results showed that emphasis of transient speech does not work for ASR. Emphasis of quasi-steady-state speech (instead of transient speech) may provide robust ASR in noisy environments since ASR models are heavily based on vowels (quasi-steady-state speech). An investigation of the effect of speech modification to emphasize quasi-steady-state speech on ASR should be conducted.

APPENDIX A

WAVELETS

This appendix supplements the description of wavelet packets in Chapter 2 by describing the continuous wavelet transform, multiresolution analysis, the discrete wavelet transform, signal decomposition and reconstruction using wavelets and factors considered in choosing a wavelet function. The descriptions are based on [71] [72] [73] [74] [75] [76] [77] [78] [79] and [80].

A.1 THE CONTINUOUS WAVELET TRANSFORM

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

$$\psi_{b,a}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{29}$$

is an orthonormal basis in the Hilbert space $L^2(R)$, where a, and b are real. L^2 refers to the space of square-integrable signals. The set of functions $\psi_{b,a}(t)$ are generated by translating and dilating the function $\psi(t)$. Parameter a is a scaling parameter. Varying it changes the center frequency and the bandwidth of $\psi(t)$. The time and frequency resolution of the wavelet transform also depend on a. Small values of the scaling parameter a provide good time localization and poor frequency resolution, and the reverse is true for large a. The time delay parameter b produces a translation in time (movement along the time axis). Dividing $\psi(t)$ 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 integers 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 $\psi(t)$ has to satisfy the following properties for it to be a wavelet:

- $\psi(t)$ integrates to zero and it's Fourier transform $\Psi(\omega)$ evaluates to zero at ω = 0 [77]

$$\Psi(\omega = 0) = \int_{-\infty}^{\infty} \psi(t)dt = 0 \tag{30}$$

- $\psi(t)$ has finite energy, i.e. most of the energy of $\psi(t)$ has to be confined to a finite duration

$$\int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty \tag{31}$$

- $\psi(t)$ has to meet the *admissibility condition*, [77] i.e.

$$\int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{\omega} d\omega = C_\psi < \infty \tag{32}$$

where $\Psi(\omega)$ is the Fourier transform of $\psi(t)$. The admissibility condition ensures perfect reconstruction of a signal from its wavelet representation and will be discussed further later in this section.

The wavelet function $\psi(t)$ may be complex. In fact, a complex wavelet function is required to analyze the phase information of signals [75].

The *continuous wavelet transform* (CWT) $W_x(b,a)$ of a continuous-time signal $x(t)$ is defined as [77]

$$W_x(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-b}{a} \right) dt \tag{33}$$

where a, 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, $\psi(t)$, i.e. $W_x(b,a) = \langle x(t), \psi^*_{b,a}(t) \rangle$. Equation (39) 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 [78]

$$W_x(b,a) = \left\langle x(t), \psi^*_{b,a}(t) \right\rangle = x(t) * \psi^*_{b,a}(-t) \tag{34}$$

The CWT expressed as a convolution may be interpreted as the output of an infinite bank of linear filters described by the impulse response $\psi_{b,a}(t)$ over the continuous range of scales a [78].

To recover $x(t)$ from $W_x(b,a)$, the mother wavelet $\psi(t)$ has to satisfy the admissibility condition given in Equation (39). If the admissibility condition is satisfied, $x(t)$ can be perfectly reconstructed from $W_x(b,a)$ as

$$x(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{|a|} W_x(b,a) \psi_{b,a}(t) da db \tag{35}$$

The constant C_ψ is the admissibility constant and is defined in Equation (39).

A.2 MULTIRESOLUTION ANALYSIS AND SCALING FUNCTION

In this Section, the scaling function $\varphi(t)$ will be introduced via a multiresolution analysis. The relationship between the scaling function $\varphi(t)$ and the wavelet function $\psi(t)$ will be discussed. This discussion follows [78].

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:

(a) Ladder property: $...V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2...$

(b) $\bigcap_{j=-\infty}^{\infty} V_j = \{0\}$.

(c) Closure of $\bigcup_{j=-\infty}^{\infty} V_j$ is equal to L^2

(d) 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.

(e) 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 $2^{-j}k$.

(f) 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 $\left\{2^{\frac{-j}{2}}\phi(2^{-j}t-k)\right\}$ is an orthonormal basis of V_j. The function $\phi(t)$ is called the *scaling function* of *multiresolution analysis*.

The scaling function $\phi_{j,k}(t) = 2^{\frac{-j}{2}}\phi(2^{-j}t-k)$ 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 [80];

(a) $\{\psi(t-k)\}$ 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 $\varphi(t)$.

(b) 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. $V_{m+1} = V_m \oplus W_m = V_0 \oplus W_0 \oplus W_1 \oplus \cdots \oplus W_m$.

(c) $L^2 = V_0 \oplus W_0 \oplus W_1 \oplus \ldots$

Using the scaling function $\varphi(t)$ and the wavelet function $\psi(t)$, 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 [71]

$$x(t) = \sum_{k=-\infty}^{\infty} c(j,k)\phi_{J,k}(t) + \sum_{j=0}^{\infty}\sum_{k=-\infty}^{\infty} d(j,k)\psi_{j,k}(t) \tag{36}$$

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 [71].

A.3 THE DISCRETE WAVELET TRANSFORM

The CWT does not offer a practical representation of the continuous-time signal $x(t)$. 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 [78].

The discrete wavelet transform (DWT) is obtained, in general, by sampling the corresponding continuous wavelet transform [78]. 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 for the use of finite impulse response (FIR) filters in the DWT implementation. There are many possible discretizations of the CWT, but the most common DWT uses a dyadic sampling lattice. 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* [75] *or fast wavelet transform* in the DWT implementation. The Mallat algorithm will be discussed in the next Section.

Sampling the CWT using a dyadic sampling lattice results in the discrete wavelet given by

$$\psi_{j,k}(t) = 2^{\frac{-j}{2}}\psi\left(2^{-j}t - k\right) \tag{37}$$

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

A.4 SIGNAL DECOMPOSITION AND RECONSTRUCTION USING WAVELETS

Equation (42) can be expanded as [71]

$$x(t)=\sum_k c(j,k)2^{-\frac{j}{2}}\phi\left(2^{-j}t-k\right)+\sum_k d(j,k)2^{-\frac{j}{2}}\psi\left(2^{-j}t-k\right) \tag{38}$$

In this and subsequent equations, scale $j+1$ is coarser than scale j. If the wavelet function is orthonormal to the scaling function, the level j scaling coefficients $c(j,k)$ and wavelet coefficients $d(j,k)$ can be obtained as:

$$c(j,k)=\left\langle x(t),\phi_{j,k}\right\rangle=\int x(t)2^{-\frac{j}{2}}\phi\left(2^{-j}t-k\right)dt \tag{39}$$

$$d(j,k)=\left\langle x(t),\psi_{j,k}\right\rangle=\int x(t)2^{-\frac{j}{2}}\psi\left(2^{-j}t-k\right)dt \tag{40}$$

The level $j+1$ scaling and detail coefficients can be obtained from the level j scaling coefficients as [71]

$$c(j+1,k)=\sum_m \tilde{h}(m-2k)c(j,m) \tag{41}$$

$$d(j+1,k)=\sum_m \tilde{g}(m-2k)c(j,m) \tag{42}$$

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* [75] and is illustrated in Figure 45. The partial binary tree of Figure 44 is sometimes referred to as a Mallat tree.

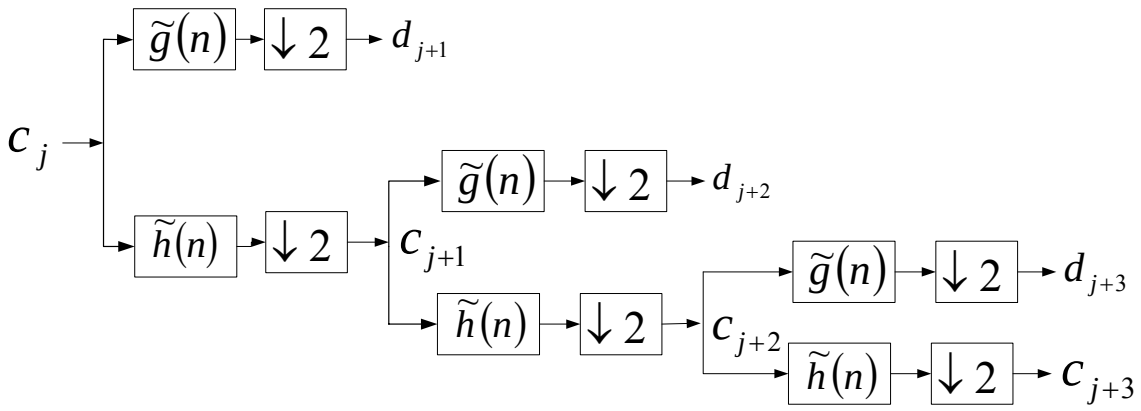

Figure 44: A three-stage Mallat signal decomposition scheme

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 45.

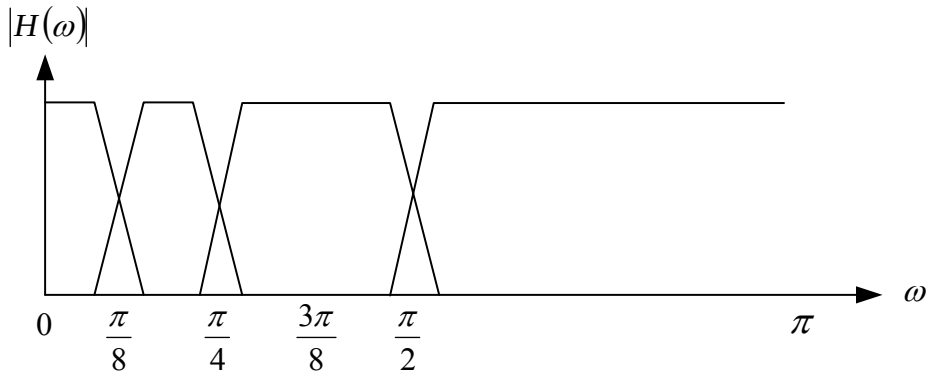

Figure 45: Frequency response for the discrete wavelet transform

As expected, the reconstruction of the level j scaling coefficients from the level $j+1$ wavelet and scaling coefficients is possible. The reconstruction can be achieved by

$$c(j,k) = \sum_m c(j+1,m)h(k-2m) + \sum_m d(j+1,m)g(k-2m) \qquad (43)$$

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 46.

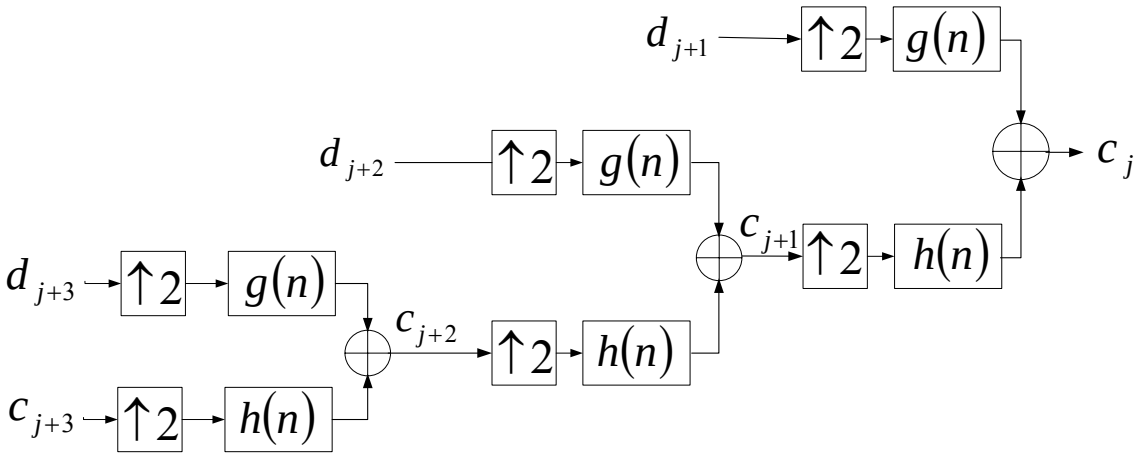

Figure 46: A three-stage Mallat signal reconstruction scheme

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 [75];

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

$g(n) = (-1)^{1-n} h(1-n)$, i.e. H and G are *quadrature mirror filters.*

$|H(\omega = 0)| = 1|$ and $\tilde{h}(n) = O(n^{-2})$ at infinity, i.e. the asymptotic upper bound of $h(n)$ at infinity is n^{-2}.

$|H(\omega)|^2 + |H(\omega + \pi)|^2 = 1$.

A.5 CHOOSING A WAVELET

The Haar wavelet is the first known and the simplest wavelet function and was proposed by Alfred Haar in 1909. Since then, many wavelets have been formulated. The paper 'Where do wavelets come from?-a personal point of view' by Daubechies presents a good historical

perspective on wavelets [73]. 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 [75]. 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 $\psi(t)$. Mallat related the number of vanishing moments and the support size to the wavelet coefficients amplitudes. His results are summarized in this Section.

A.5.1 Vanishing Moments

$\psi(t)$ has p vanishing moments if

$$\int_{-\infty}^{\infty} t^k \psi(t) dt = 0 \quad for \quad 0 \le k < p \tag{44}$$

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

A.5.2 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 $\psi(t)$ has a compact support of size K, then 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 $\psi(t)$.

If $\psi(t)$ has p vanishing moments, then its support size is at least $2p-1$ [75]. 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. But if the number of singularities of $x(t)$ is large, a $\psi(t)$ with a short support size is a better choice.

APPENDIX B

ILLUSTRATION OF DAUBECHIES-18 WAVELET

This appendix illustrates, for a Daubechies-18 wavelet, the following function

(a) Scaling function

(b) Wavelet function

(c) Impulse response and magnitude frequency response of lowpass and highpass decomposition filters associated with the Daubechies-18 wavelet function. The decomposition lowpass and highpass filters are used as a prototype type filters from which the filters $H_k(z)$, $0 \le k \le K-1$, are derived in the wavelet decomposition structure of Figure 13.

(d) Impulse response and magnitude frequency response of lowpass and highpass reconstruction filters associated with the Daubechies-18 wavelet function. The reconstruction lowpass and highpass filters are used as a prototype type filters from which the filters $G_k(z)$, $0 \le k \le K-1$, are derived in the wavelet reconstruction structure of Figure 13.

The Daubechies-18 wavelet was used for wavelet packet decomposition in the transient extraction algorithm because it offers a good balance of frequency selectivity and computation efficiency.

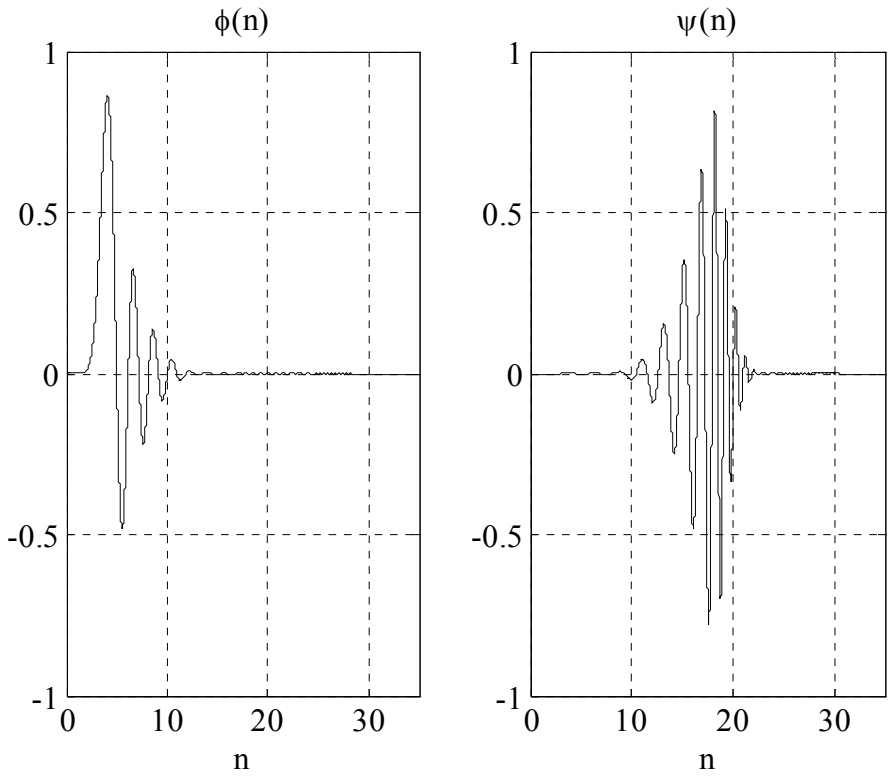

Figure 47: Scaling $\phi(n)$ and wavelet function $\psi(n)$ for Daubechies-18 wavelet

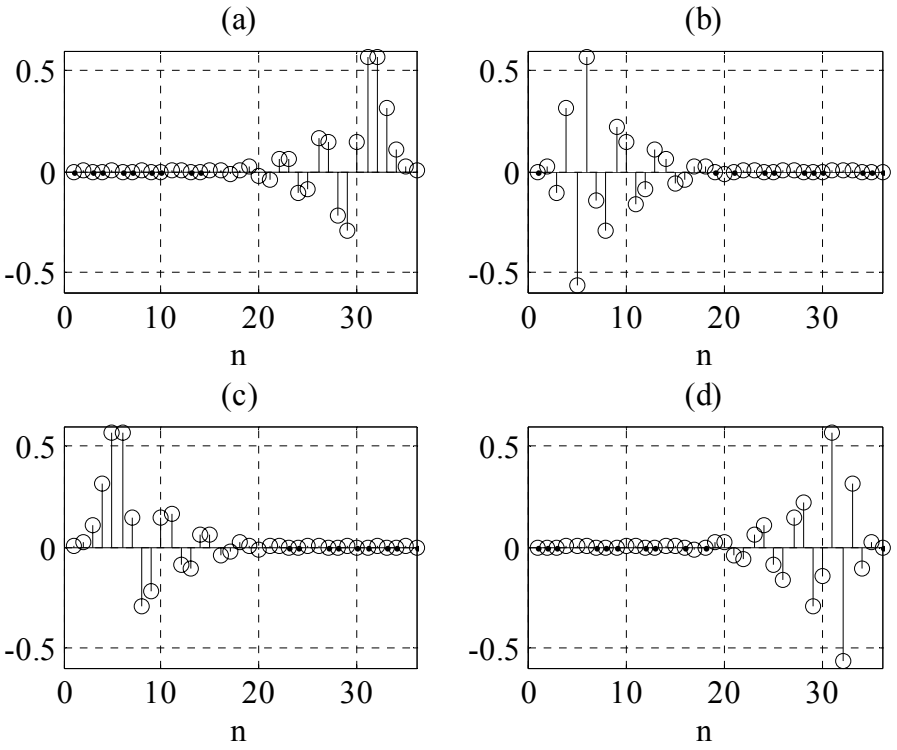

Figure 48: Impulse responses for (a) lowpass decomposition filter, (b) highpass decomposition filter, (c) lowpass reconstruction filter and (d) highpass reconstruction filter for Daubechies-18 wavelet.

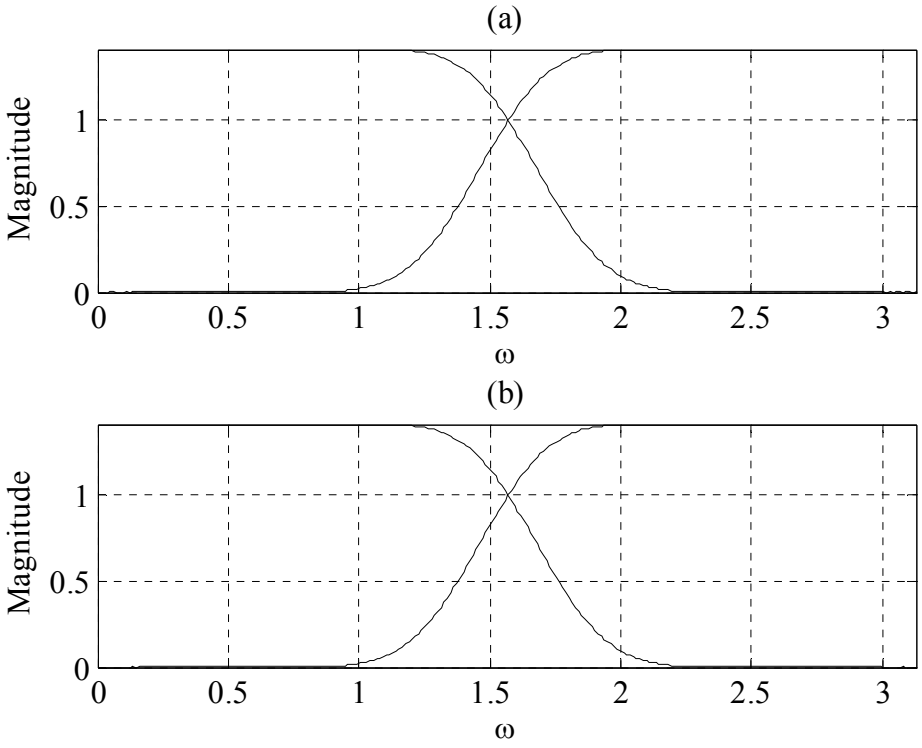

Figure 49: Magnitude frequency responses for (a) lowpass and highpass decomposition filters and (b) lowpass and highpass reconstruction filters.

APPENDIX C

SELECTION OF A VOICED/UNVOICED METHOD FOR THE UNVOICED SPEECH BOOSTER

The transitivity function has larger peaks for transitions into and out of high energy formants than for transitions associated with low energy events such as unvoiced consonants. The incorporation of an unvoiced speech booster to the transient extraction method, to increase the peaks of the transitivity function that correspond to unvoiced consonants, was investigated. Two voiced/unvoiced detection methods, described in Chapter 2, were considered for use with the unvoiced speech booster method – a short-time energy/short-time average zero-crossing rate–based method and a short-time autocorrelation function–based method. The selection of the voiced/unvoiced detection methods and its parameters are described in here.

For the short-time energy/short-time average zero-crossing rate–based method, the short-time energy and the short-time average zero-crossing rate were both computed using a window duration of 25 ms and a window overlap of 10 ms and were normalized to have a maximum value of one. The short-time autocorrelation function–based method was computed using a window duration of 20 ms with no window overlap, and the autocorrelation was normalized so that $R_n(k=0)=1$.

To evaluate the two methods, voiced and unvoiced segments of ten words from the modified rhyme test list [12] [13] were identified manually by the author using listening tests and an International Phonetic Alphabet (IPA) chart for English [87], shown in Figure 50. Then the two methods were used to automatically identify voiced and unvoiced segments. The thresholds for discriminating between voiced and unvoiced segments and the window duration and overlap for the two voiced/unvoiced detection methods were adjusted to obtain the best results for each method. For the short-time energy/short-time average zero-crossing rate–based method, a windowed segment of speech was considered voiced if the short-time energy was greater than 0.4 and the short-time average zero-crossing rate was less than 0.5. For the short-time autocorrelation function–based method, a windowed speech segment was considered to be voiced speech if it included an autocorrelation function peak with a value greater than 0.3 at $k > 0$.

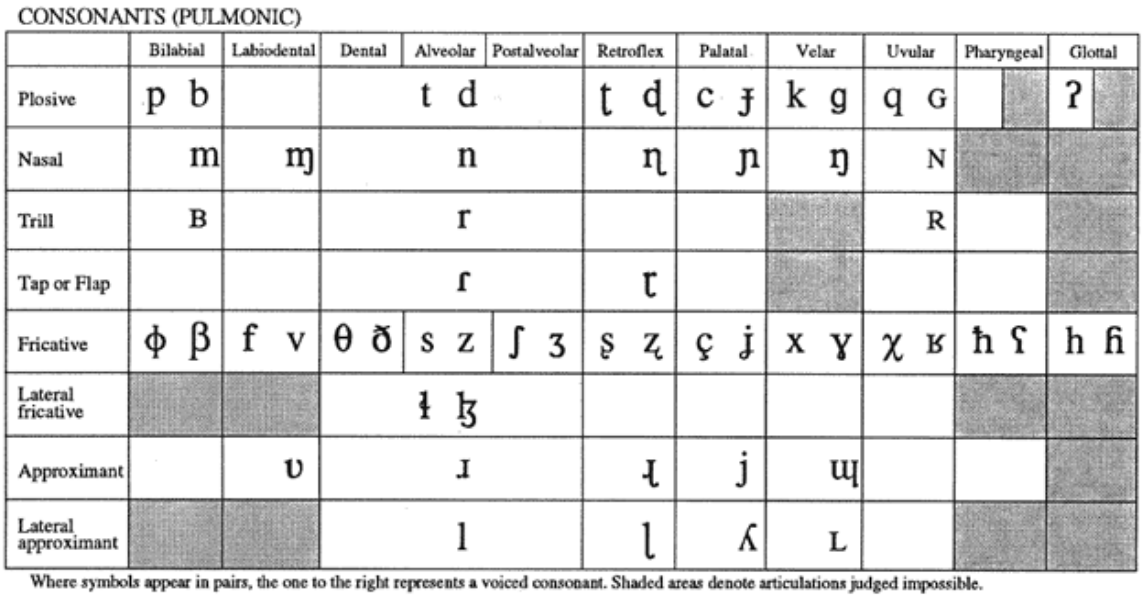

Figure 50: The international phonetic alphabet (IPA) chart for English consonants [87].

Figure 51 shows time-domain plots of the 10 words, the manually made voiced/unvoiced decision and the automatically made voiced/unvoiced decision by the two methods. The short-time autocorrelation-based voiced/unvoiced detection method is correct more often and was selected for evaluation.

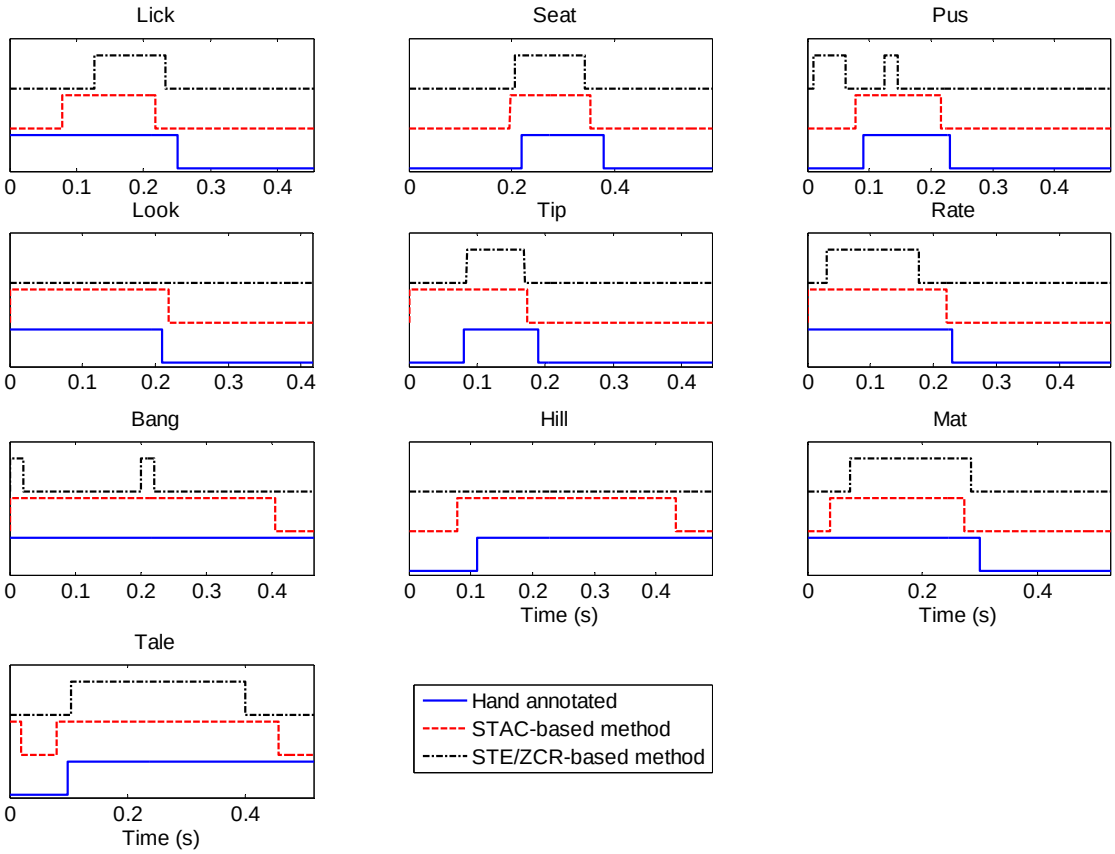

Figure 51: Selection of a voiced/unvoiced detection method. Each plot compares manually identified voiced segment to voiced segments automatically identified using the short-time autocorrelation (STAC) function-based method and the short-time energy/zero-crossing rate (STE/ZCR)-based method. High indicates voiced and low indicates unvoiced.

BIBLIOGRAPHY

[1] S. Yoo, *Speech Decomposition and Enhancement*. Department of Electrical and Computer Engineering, University of Pittsburgh, PhD Dissertation, 2005.

[2] S. Yoo, J. R. Boston, J. D. Durrant, K. Kovacyk, S. Shaiman, El-Jaroudi, and C. C. Li, "Speech signal modification to increase intelligibility in noisy environments," *The Journal of the Acoustical Society of America,* vol. 122, pp. 1138-1149, August 2007.

[3] C. Tantibundhit, *Speech Enhancement Using Transient Speech Components*. Department of Electrical and Computer Engineering, University of Pittsburgh, PhD Dissertation, 2006.

[4] C. Tantibundhit, J. R. Boston, C. C. Li, J. D. Durrant, S. Shaiman, K. Kovacyk, and El-Jaroudi, "New signal decomposition method based speech enhancement," *Signal Processing,* vol. 87, pp. 2607-2627, November 2007.

[5] M. D. Skowronski and J. G. Harris, "Applied Principles of Clear and Lombard Speech for Automated Intelligibility Enhancement in Noisy Environments," *Speech Communication,* vol. 48, pp. 549-559, May 2006.

[6] M. Raghavan, M. D. Skowronski, and J. G. Harris, "Enhanced energy redistribution speech intelligibility algorithm with real-time implementation," *The Journal of the Acoustical Society of America,* vol. 116, p. 2482, November 2004.

[7] A. R. Jayan, P. C. Pandey, and P. K. Lehana, "Automated Detection of Transition Segments for Intensity and Time-Scale Modification for Speech Intelligibility Enhancement," *IEEE International Conference on Signal Processing, Communications and Networking,* pp. 63 - 68, January 2008.

[8] H. Rogers, *The Sounds of Language*. Essex, England: Pearson Education Limited, 2000.

[9] W. Strange, J. J. Jenkins, and T. L. Johnson, "Dynamic specification of co-articulated vowels," *The Journal of the Acoustical Society of America,* vol. 75, pp. 695-705, September 1983.

[10] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE,* vol. 67, pp. 1586-1604, December 1979.

[11] V. Ramamoorthy, N. S. Jayant, R. V. Cox, and M. M. Sondhi, "Enhancement of ADPCM speech coding with backward-adaptive algorithms for postfiltering and noise feedback," *IEEE Journal on Selected Areas in Communications,* vol. 6, pp. 364-382, February 1988.

[12] A. S. House, C. E. Williams, H. M. L. Hecker, and K. D. Kryter, "Psychoacoustic speech tests: A modified rhyme test," *United States Air Force Technical Documentary Report No. ESD-TDR-63-403,* June 1963.

[13] A. S. House, C. E. Williams, H. M. L. Hecker, and K. D. Kryter, "Articulation-testing methods: Consonantal differentiation with a closed-response set," *The Journal of the Acoustical Society of America,* vol. 37, pp. 158-166, January 1965.

[14] C. Mackersie, A. Neuman, and H. Levitt, "A comparison of response time and word recognition measures using a word-monitoring and closed-set identification task," *Ear and Hearing,* vol. 20, pp. 140-148, April 1999.

[15] K. D. Kryter, "Methods for the Calculation of the Articulation Index," *The Journal of the Acoustical Society of America,* vol. 34, pp. 1689-1697, November 1962.

[16] J. B. Allen, "Consonant Recognition and the Articulation Index," *The Journal of the Acoustical Society of America,* vol. 117, pp. 2212-2223, April 2005.

[17] M. R. Weiss, E. Aschkenasy, and T. W. Parsons, *Study and development of the INTEL technique for improving speech intelligibility*: Nicolet Scientific Corp., Final Report NSC-FR/4023, 1975.

[18] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 27, pp. 113-120, April 1979.

[19] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 4, pp. 208-211, April 1979.

[20] H. T. Hu and C. Yu, "Adaptive noise spectral estimation for spectral subtraction speech enhancement," *IET Signal Processing,* vol. 1, pp. 156-163, September 2007.

[21] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. Singapore: John Wiley and Sons, 2002.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. ASSP-32, pp. 1109-1121, December 1984.

[23] G. H. Ding, T. Huang, and B. Xu, "Suppression of additive noise using a power spectral density MMSE estimator," *IEEE Signal Processing Letters,* vol. 11, pp. 585-588, June 2004.

[24] N. W. D. Evans, J. S. D. Mason, W. M. Liu, and B. Fauve, "An Assessment on the Fundamental Limitations of Spectral Subtraction," *IEEE International Conference on Acoustics, Speech and Signal Processing,* vol. 1, pp. I-145 - I-148, May 2006.

[25] Y. Hu and P. C. Loizou, "A Comparative Intelligibility Study of Speech Enhancement Algorithms," *IEEE International Conference on Acoustics, Speech and Signal Processing,* vol. 4, pp. IV-561 - I-V564, April 2007.

[26] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing,* p. 334341, July 2003.

[27] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing,* vol. 11, pp. 700-708, 2003.

[28] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. ASSP-33, pp. 443-445, 1985.

[29] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters,* vol. 9, pp. 12-15, January 2002.

[30] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Transactions on Speech and Audio Processing,* pp. 779-807, 2001.

[31] S. Kamath and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *IEEE International Conference on Acoustics, Speech and Signal Processing,* vol. 4, May 2002.

[32] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Transactions on Speech and Audio Processing,* pp. 56-67, January 2004.

[33] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," *IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 629-632, 1996.

[34] A. M. Liberman and F. S. Cooper, "In Search of the Acoustic Cues," in *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*, A. Valdman, Ed. Mouton, Netherlands: The Hague, 1972, pp. 329-338.

[35] R. K. Potter, G. A. Kopp, and G. H. C., *Visible Speech*. New York, NY: Van Nostrand, 1947.

[36] M. Joos, *Acoustic Phonetics*. Baltimore, MD: Linguistic Society of America, 1948.

[37] A. M. Liberman, D. P. C., F. S. Cooper, and L. J. Gerstman, "The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants," *Psychological Monographs: General and Applied,* vol. 68, pp. 1-13, 1954.

[38] G. Fant, *Speech Sounds and Features*. Cambridge, MA: MIT Press, 1973.

[39] A. M. Liberman, K. S. Harris, H. S. Hoffman, P. C. Delattre, and F. S. Cooper, "Effect of Third-Formant Transitions on the Perception of the Voiced Stop Consonants," *The Journal of the Acoustical Society of America,* vol. 30, pp. 122-126, February 1958.

[40] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the Speech Code," *Psychological Review,* vol. 74, pp. 431-461, November 1967.

[41] I. B. Thomas and R. J. Niederjohn, "Enhancement of Speech Intelligibility at High Noise Levels by Filtering and Clipping," *Journal of the Audio Engineering Society,* vol. 16, pp. 412-415, October 1968.

[42] I. B. Thomas and R. J. Niederjohn, "The Intelligibility of Filtered-Clipped Speech," *Journal of the Audio Engineering Society,* vol. 18, pp. 299-303, June 1970.

[43] I. B. Thomas and W. J. Ohley, "Intelligibility Enhancement through Spectral Weighting," *Conference on Speech Communication and Processing,* pp. 360-363, June 1972.

[44] R. J. Niederjohn and J. H. Grotelueschen, "The Enhancement of Speech Intelligibility in High Noise Levels by High-Pass Filtering Followed by Rapid Amplitude Compression," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. ASSP-24, pp. 277-282, August 1976.

[45] R. J. Niederjohn and J. H. Grotelueschen, "Speech Intelligibility Enhancement in a Power Generating Noise Environment,," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. ASSP-26, pp. 378-380, August 1978.

[46] I. B. Thomas and A. Ravindran, "Preprocessing of an Already Noisy Speech Signal for Intelligibility Enhancement," *The Journal of the Acoustical Society of America,* vol. 49, p. 133, January 1971.

[47] E. Villchur, "Signal processing to improve intelligibility in perceptive deafness," *The Journal of the Acoustical Society of America,* vol. 53, pp. 1646-1657, November 1973.

[48] A. Ramasubramanian, K. L. Payton, and A. H. Costa, "A comparison between conventional and wavelet based amplitude compression schemes," *IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis,* pp. 161-164, October 2006.

[49] J. G. Harris and M. D. Skowronski, "Energy redistribution speech intelligibility enhancement, vocalic and transitional cues," *The Journal of the Acoustical Society of America,* vol. 112, p. 2305, November 2002.

[50] P. S. Chanda and S. Park, "Speech Intelligibility Enhancement Using Tunable Equalization Filter," *IEEE International Conference on Acoustics, Speech and Signal Processing,* vol. 4, pp. 613-616, April 2007.

[51] ANSI-S3.5-1997, "Methods for the Calculation of the Speech Intelligibility Index," *American National Standards Institute,* 1997.

[52] S. Gordon-Salant, "Recognition of Natural and Time/Intensity altered CVs by Young and Elderly Subjects with Normal Hearing," *The Journal of the Acoustical Society of America,* vol. 80, pp. 1599-1607, December 1986.

[53] V. Hazan and A. Simpson, "The Effect of Cue-Enhancement on Intelligibility of Nonsense Word and Sentence Material in Noise," *Speech Communication,* vol. 24, pp. 211-226, June 1998.

[54] P. Tallal and M. Piercy, "Defects of Non-Verbal Auditory Perception in Children with Developmental Aphasia," *Nature,* vol. 241, pp. 468-469, February 1973.

[55] P. Tallal, S. L. Miller, G. Bedi, G. Byma, X. Wang, S. S. Nagarajan, C. Schreiner, W. M. Jenkins, and M. M. Merzenich, "Language Comprehension in Language-Learning Impaired Children Improved with Acoustically Modified Speech," *Science,* vol. 271, pp. 81-84, January 1996.

[56] J. C. R. Licklider and I. Pollack, "Effects of Differentiation, Integration, and Infinite Peak Clipping upon the Intelligibility of Speech," *The Journal of the Acoustical Society of America,* vol. 20, pp. 42-51, January 1948.

[57] A. Rao and R. Kumaresan, "On Decomposing Speech into Modulated Components," *IEEE Transactions on Speech and Audio Processing,* vol. 8, pp. 204-254, May 2000.

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

[59] Q. Zhu and A. Alwan, "On the Use of Variable Frame Rate Analysis in Speech Recognition,," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 3, pp. 1783-1786, June 2000.

[60] P. Le Cerf and D. Van Compernolle, "Frame and Frame Dimension Reduction Techniques for Automatic Speech Recognition," *IARP/IEEE Conference on Image, Speech and Signal Analysis,* pp. 717-720, August 1992.

[61] P. Le Cerf and D. Van Compernolle, "A New Variable Frame Rate Analysis Method for Speech Recognition," *IEEE Signal Processing Letters,* vol. 1, pp. 185-187, December 1994.

[62] K. L. Brown and V. R. Algazi, "Characterization of Spectral Transitions with Applications to Acoustic Sub-word segmentation and Automatic Speech Recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 1, pp. 104-107, May 1989.

[63] T. F. Quateiri and R. B. Dunn, "Speech Enhancement Based on Auditory Spectral Change," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 1, pp. 257-260, May 2002.

[64] Y. M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *IEEE Transactions on Signal Processing,* vol. 36, pp. 1943 - 1954 September 1991.

[65] G. Fairbanks, "Test of phonemic differentiation: The rhyme test," *The Journal of the Acoustical Society of America,* vol. 30, pp. 596-600, July 1958.

[66] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Upper Saddle River, NJ: Prentice Hall, Inc., 1978.

[67] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 1, May 2001.

[68] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Signal Processing,* vol. ASSP-28, pp. 357-366, August 1980.

[69] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," *International Symposium on Music Information Retrieval,* 2000.

[70] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: PTR Prentice Hall, Inc., 1993.

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

[72] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992.

[73] I. Daubechies, "Where do Wavelets Come From?-A Personal Point of View," *Proceedings of the IEEE,* vol. 84, pp. 510-513, April 1996.

[74] M. Jansen, *Noise Reduction by Wavelet Thresholding*. New York, NY: Springer-Verlag, 2001.

[75] S. G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis Machine Intelligence,* vol. 11, pp. 674-693, July 1989.

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

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

[78] A. Teolis, *Computational Signal Processing with Wavelets*. Boston, MA: Birkhauser, 1988.

[79] P. P. Vaidyanathan and I. Djokovic, "Wavelet transform," in *Mathematics for Circuits and Filters*, W. K. Chen, Ed. Boca Raton, FL: CRC Press LLC, 2000, pp. 131-216.

[80] G. G. Walter, *Wavelets and Other Orthogonal Systems with Applications*. Boca Raton, FL: CRC Press Inc., 1994.

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

[82] D. M. Rasetshwane, J. R. Boston, C. C. Li, and J. D. Durrant, "An Index to Measure "Transient-ness" of Speech," *IEEE Workshop on Digital Signal Processing and Signal Processing Education,* pp. 54-59, January 2009.

[83] S. Yoo, J. R. Boston, J. D. Durrant, K. Kovacyk, S. Karn, S. Shaiman, El-Jaroudi, and C. C. Li, "Relative energy and intelligibility of transient speech components," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 1, pp. 69-72, March 2005.

[84] F. E. Musiek and W. F. Rintelmann, *Contemporary Perspectives in Hearing Assessment*: Allyn and Bacon, 1999.

[85] ANSI-S3.6-1996, "American National Standard Specification for Audiometers," *American National Standards Institute,* 1996.

[86] D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hayerman, R. Hetu, J. Kei, C. Lui, J. Kiessling, M. Nasser Notby, N. H. A. Nasser, W. A. H. E. Kholy, Y. Nakanishi, H. Oyer, R. Powell, D. d. Stephens, R. Meredith, T. Sirimanna, G. avartkiladze, G. Frolenkov, S. Westerman, and C. Ludvigsen, "An International Comparison of Long-Term Average Speech Spectra," *The Journal of the Acoustical Society of America,* vol. 96, pp. 2108-2120, October 1994.

[87] IPA, *Handbook of the International Phonetic Association*. Cambridge, UK: Cambridge University Press, 1999.