SPEECH DECOMPOSITION AND ENHANCEMENT

by

Sungyub Yoo

BS, Soonchunhyang University, 1995

MS, University of Pittsburgh, 1998

Submitted to the Graduate Faculty of

School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

SCHOOL OF ENGINEERING

This dissertation was presented

by

Sungyub Yoo

It was defended on

June 29, 2005

and approved by

Ching-Chung Li, Professor, Electrical and Computer Engineering

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

Heung-no Lee, Assistant Professor, Electrical and Computer Engineering

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

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

SPEECH DECOMPOSITION AND ENHANCEMENT

Sungyub Yoo, PhD

University of Pittsburgh, 2005

The goal of this study is to investigate the roles of steady-state speech sounds and transitions between these sounds in the intelligibility of speech. The motivation for this approach is that the auditory system may be particularly sensitive to time-varying frequency edges, which in speech are produced primarily by transitions between vowels and consonants and within vowels. The possibility that selectively amplifying these edges may enhance speech intelligibility is examined.

Computer algorithms to decompose speech into two different components were developed. One component, which is defined as a tonal component, was intended to predominately include formant activity. The second component, which is defined as a non-tonal component, was intended to predominately include transitions between and within formants.

The approach to the decomposition is to use a set of time-varying filters whose center frequencies and bandwidths are controlled to identify the strongest formant components in speech. Each center frequency and bandwidth is estimated based on FM and AM information of each formant component. The tonal component is composed of the sum of the filter outputs. The non-tonal component is defined as the difference between the original speech signal and the tonal component.

The relative energy and intelligibility of the tonal and non-tonal components were compared to the original speech. Psychoacoustic growth functions were used to assess the intelligibility. Most of the speech energy was in the tonal component, but this component had a significantly lower

maximum word recognition than the original and non-tonal component had. The non-tonal component averaged 2% of the original speech energy, but this component had almost equal maximum word recognition as the original speech.

The non-tonal component was amplified and recombined with the original speech to generate enhanced speech. The energy of the enhanced speech was adjusted to be equal to the original speech, and the intelligibility of the enhanced speech was compared to the original speech in background noise. The enhanced speech showed higher recognition scores at lower SNRs, and the differences were significant. The original and enhanced speech showed similar recognition scores at higher SNRs. These results suggest that amplification of transient information can enhance the speech in noise and this enhancement method is more effective at severe noise conditions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank you to my advisor, Dr. J. R. Boston, for his guidance during my Ph.D study. I also thanks to Drs. Ching-Chung Lee, Amro A. El-Jaroudi, Heung-no Lee, and John D. Durrant for taking their time to serve on my thesis committee and for their suggestions. I also thank you to Dr. Susan Shaiman for her suggestions.

I would also like to thank you the speech research group at University of Pittsburgh; Paul, Daniel, Kristie, Bob, and Nazeeh. Special thanks goes to Wonchul and the many other friends I made during my years at the University of Pittsburgh.

I particularly owe my success to my wife, Jeymyung, for being extremely patient and supportive for the past years. I would like to thanks to my family, Myungchul, Hwaja, Sooyeon, Hanmoo, two latest additions Seoyoung and Seoyoon, Woonhwa, Hyunjoon, and Minkyung for their support.

Finally, I thank you God for letting me have all these chances and wonderful times at Pittsburgh.

1.0 INTRODUCTION

1.1 DECOMPOSITION AND ENHANCEMENT OF SPEECH

The goal of this study is to investigate the roles of steady-state speech sounds and transitions between these sounds on the intelligibility of speech. Computer algorithms to decompose speech into a "quasi-steady-state" component and a "transition" component are developed, and the energy and intelligibility of the components are compared to the original speech. The quasi-steady-state component includes signal energy with frequency content and amplitude that are relatively constant over short time periods (minimum 10-20 msec.). The transition component represents changes in frequency content and amplitude between quasi-steady-state components. A method to enhance the intelligibility of speech in noise, based on the speech decomposition, is developed, and the intelligibility of original and enhanced speech in background noise is examined by psychoacoustic tests.

Speech sounds can be classified as vowels or consonants. Vowels are voiced sounds that are characterized by the size and shape of the vocal tract [1], [2], [3]. Changing the size and shape of the vocal tract produces a change in frequency content of speech sounds. The frequency region where the speech energy is concentrated is called a formant [4]. Consonants are usually generated by a narrowing or by complete obstruction of region of the vocal tract [1], [2], [3].

Quasi-steady-state components are the dominant characteristic of vowels. Although consonants are predominantly brief transients, some include quasi-steady-state components as well. These quasi-steady-state parts of consonants are called hubs. Since the onset and offset of

speech sounds are inherently transient, both vowels and consonants contain transient events. The transitions are observed between vowels and consonants and within vowels (*e.g.* diphthong). The articulators cannot move instantly from one position to another, and initial portions of formants often show brief frequency shifts that, for a given vowel, may differ among different consonant-vowel combinations [1]. Consequently, the transient energy is included in both vowels and consonants. Conventional vowel-consonant classifications and concepts of spectral composition de-emphasize this transition information.

Most human sensory systems are sensitive to abrupt changes in stimuli. If the auditory system shows the same characteristics in the frequency domain, it would probably be particularly sensitive to time-varying frequency edges that reflect transition components in speech. Although these transitions represent a small proportion of the total speech energy compared to quasi-steady state portions of both vowels and consonants, they may be critical to the perception of speech by humans.

Traditional methods of studying the auditory system and speech intelligibility have emphasized frequency-domain techniques, a perspective that also has dominated concepts of speech intelligibility [5], [6], [7], [8], [9]. While it is generally recognized that voicing and steady vowel sounds are largely low frequency and that consonants are dominated by higher frequencies, no single cutoff frequency uniquely separates them. Information on transitions between and within vowel sounds is even more difficult to isolate using fixed-frequency filters, as this information is inherently dynamic and can be rather broad band.

In this project, an algorithm to emphasize transition components in speech is developed in order to investigate the role of these components in speech intelligibility. The algorithm decomposes speech into two components. One component is intended to predominately include

quasi-steady-state formant activity representing primarily vowels and hubs of consonants, and it is referred to as the "tonal" component. The second component is intended to emphasize transitions between vowels and consonants and within vowels, and it is defined as the "non-tonal" component. We compare the energy and intelligibility of the tonal and non-tonal components to the original speech. We expect the intelligibility of the tonal and non-tonal speech components to be different, and we suggest that the non-tonal component, since it emphasizes the transition information in speech, may be critical to the perception of speech.

We expect the tonal component to contain most of the energy of original speech and the non-tonal component to contain less energy. Thus, noise would affect the non-tonal component more than it affects the tonal component. To enhance speech, the non-tonal component is amplified and recombined with the original speech. The energy of the enhanced speech is adjusted to be equal to that of the original speech. The intelligibility of the enhanced speech in noise is compared to the intelligibility of the original speech in noise.

Our approach to speech decomposition is to first highpass filter (at 700 Hz) the speech signal to remove most of the voicing energy. Then three time-varying filters whose center frequencies and bandwidths are controlled to pass most of the energy in the three largest formant components in the signal are used. Each center frequency and bandwidth is estimated using frequency modulation (FM) and amplitude modulation (AM) information of each formant component. The tonal component is composed of the sum of the filter outputs. It is subtracted from the original speech signal to yield the non-tonal component. The decomposition can be viewed as removing as much of the quasi-steady-state energy from the original speech signal as possible, while maintaining reasonable intelligibility in the remaining speech signal. That is, the

energy of the tonal component is maximized, while keeping reasonable intelligibility in the non-tonal component.

1.2 OUTLINE

This dissertation is organized as follows. Relevant literature for this study is summarized in Chapter 2. Chapter 3 describes the decomposition method, which is based on time-varying filters, and explains how the center frequency and bandwidth of the time-varying filters are determined. Synthetic tone signals are presented to illustrate the proposed decomposition algorithm. The characteristics of the time-varying filters are described by analyzing their response to synthetic chirp signals, and software modifications for processing long speech samples and issues related to computation times are discussed. Results for real speech samples are presented in Chapter 4, and the effects of parameter variations on filter performance are described. Psychoacoustic evaluations of the intelligibility of the speech components and of the enhanced speech in noise are presented in Chapter 5. Finally, implications of this study and future research areas are discussed in Chapter 6.

2.0 BACKGROUND

A general background on speech and previous studies on speech processing techniques are reviewed in this chapter. The nature of speech intelligibility and methods to test speech intelligibility are also briefly described.

2.1 STRUCTURE OF SPEECH

The characteristics of speech formants, which dominate the tonal component, and transitions between and within formants, which dominate the non-tonal component, are described in this section. The effects of noise on speech intelligibility are also discussed.

2.1.1 Formants and Vowels

A formant can be defined as a natural mode or resonance of the vocal tract [1], [2], [3], [4]. There are an infinite number of formants for speech, but only the lowest three or four formants are typically considered in practice. Each formant can be characterized by the formant frequency, which represents the frequency content (center frequency) of a certain formant.

A formant is generally seen as a peak in the acoustic spectrum of a speech sound [1], [2]. The vocal tract can be viewed as an energy modifier, and its transfer function can be characterized by the first three or four formants. The vocal tract does not generate sound energy but modifies sound energy provided by a source of sound.

In some views, vowels are considered simple sounds to analyze and describe because of the quasi-steady-state acoustic pattern in vowel sounds [2]. Each vowel sound has different formant frequencies.

2.1.2 Transitions

The acoustic characteristics of consonants are more complicated and more difficult to describe than vowel sounds. A momentary narrowing or obstruction of the vocal tract can produce a stop consonant (in English /*p b t d k g*/). Transition sounds are associated with changes of the vocal tract configuration between vowels, such as the change from a stop to a relatively open position for the following vowel.

A speech sound can not go to the next sound without producing transition events. The transition from consonant to vowel or vowel to consonant is related to changing or shifting of the formants [2]. These changes of formants are caused by the changes of the resonating cavities of the vocal tract, which can be characterized by frequency changes of the formants. The transitions have information on the place of stop articulations and voicing features of the vocal tract. The transitions also contain information on timing of the articulatory changes because the acoustic changes have the same duration as the articulatory changes [2]. The transitions are probably important acoustic cues for speech intelligibility. However, it is not easy to detect or measure the transitions because of the variability in their durations, rates of changes, and start and end points.

2.1.3 Effects of Noise

Speech intelligibility is degraded when speech is corrupted by noise. In some situations, an otherwise clearly audible sound can be masked by another sound. For example, conversation

at a bus stop can be completely impossible if a loud bus is driving past. This phenomenon is called masking. A quieter sound is masked if it is made inaudible in the presence of a louder sound. The presence of noise can cause masking of all or part of the speech. Once the noise energy has increased above an effective level, a noise increment results in a corresponding increase in speech threshold. In general, the intelligibility of consonants, which have lower energy, is more affected by a given noise level than the intelligibility of vowels, which have higher energy [10].

2.2 ANALYSIS OF SPEECH

Attempts to decompose speech into different components have been reported, but these decompositions were primarily performed for speech coding or production of synthetic speech. Speech decomposition for speech enhancement or to study the relation between components of speech and overall speech intelligibility has been rarely described. Many investigators have addressed the problem of identifying the start and end of phonemes or word segments for automated speech recognition, but only a few studies have focused specifically on transition components in speech. In this section, literature on speech decomposition, time-varying filtering, tracking algorithms, and noise reduction are reviewed. Background on the tracking algorithm that is used in this study is described, and the advantages and disadvantages of the various approaches are addressed.

2.2.1 Periodic and Aperiodic Decomposition

Yegnanarayana *et al*. proposed an iterative algorithm to identify and separate periodic and aperiodic components of speech signal [11]. They considered a speech signal to be the output of a vocal tract excited by pulses of quasi-periodic (periodic component) and random (aperiodic component) sequences. The decomposition was performed on an approximation to the excitation signal of the vocal tract (residual signal) rather than on the speech signal directly. Linear prediction analysis was performed on speech data and the residual signal was obtained by passing the speech signal through the inverse filter calculated from these linear prediction coefficients. Voiced and unvoiced parts of the residual signal were determined. The voiced part was decomposed into periodic and aperiodic parts by identifying frequency regions of harmonic and noise components of the voiced part. A first approximation to the aperiodic component was obtained from the signal corresponding to the noise frequency region, and then the aperiodic component in the harmonic frequency region was estimated by an iterative algorithm. The estimated aperiodic component was subtracted from the residual signal to obtain the periodic component.

The periodic and aperiodic components of the residual signal were passed through an all-pole filter of the vocal tract, which was derived from the linear prediction analysis, to generate the corresponding components of the speech signal. A synthetic voiced segment generated by a formant synthesizer with white Gaussian noise was tested to demonstrate the capability of their decomposition algorithm. The authors mentioned some computational problems to implement the algorithm in real-time and suggested possible applications to speech synthesis and production of voice with desired source characteristics.

This study primarily investigated the modeling and decomposition of speech signals for speech synthesis or speech production. The authors were not particularly concerned with speech perception or intelligibility.

2.2.2 Wavelet Decompositions

Daubechies and Maes proposed a nonlinear squeezing of the continuous wavelet transform for estimating the modulated components of speech and the parameters characterizing them [12]. First, they noted the similarities between the cochlea and the natural wavelet transform. They pointed out the disadvantages of the discrete wavelet transform (at low frequency, poor time resolution and at high frequency, poor frequency resolution) and focused on reassigning or weighting the important (desired) components in the time-frequency plane in order to remedy this blurring in time and frequency resolutions. They transformed the original time-scale plane to a time-instantaneous frequency plane by reassigning contributions with the same instantaneous frequencies to the same bin, weighted by amplitude. As a result, the reassigned (synchrosqueezed) wavelet transform showed an improved resolution of the speech signal. After reassigning, the speech structures were well identified. From the reassigning representation, they determined more clearly the central frequency of speech signal.

Daudet and Torresani proposed a decomposition method including tonal, transient, and stochastic components of an audio signal [13]. They applied a modulated discrete cosine transform to extract the tonal component, which is a locally stationary signal, and defined the non-tonal component as the difference between original and tonal components. The tonal component contained most of the energy of the signal. A discrete wavelet transform was then used to obtain a transient component, which exhibited rapid variations, from the non-tonal

component. The stochastic component was defined as the difference between the non-tonal and transient components. They illustrated the decomposition results using musical sounds. The transient component captured most of the rapid activity (attack) in a musical sample. The stochastic component was composed of a relatively white noise component with much smaller dynamic range than the tonal or transient components, although the stochastic component still included some tonal quality. They applied these decompositions to audio signal encoding, using a musical sound. The three components were estimated and individually encoded for high sound quality with high compression ratio. The musical sound was originally recorded at 16 bit per sample, and the encoding of the tonal and transient components using their method required about 0.167 bits per sample and 0.8 bits per sample respectively.

These studies primarily investigated modeling, feature extraction, and decomposition of audio and speech signals for coding purposes. They were not particularly concerned with speech perception or intelligibility.

2.2.3 Identifying Transition Segments

Zhu and Alwan focused on transitional information in speech recognition. They suggested that computing frames every 10 msec was not sufficient to represent transitional information, and they proposed a simple method, using variable frame rates, to detect transitions [14]. The frame size was constant, but the overlap between windows was increased (resulting in windows being applied more frequently) when speech models showed that the speech was changing rapidly. The rate of change was described by calculating an energy-weighted Euclidean distance between Mel-frequency cepstral coefficients in consecutive frames. The Euclidean distance increased when the speech was rapidly changing. If the Euclidean distance exceeded a

certain threshold, the overlap was increased, so that the algorithm effectively detected the rapidly changing speech (transitions). They showed that the variable frame rate speech processing improved the performance of automated recognition of noisy speech.

Yu and Chan proposed a transient model for speech coding [15]. They used an 8^{th} order all-pole filter with random noise excitation to model the unvoiced part of speech. The voiced part was modeled by low frequency harmonic components. The transitions were characterized by the onset time and growth rate of each harmonic component of the transient speech segment. They applied the transient model to a 2.4 kbps speech coder to improve the quality at the transition region of the speech signal. They performed an informal listening test and reported that the speech quality of the proposed coder is preferable, especially at the transition parts.

Zhao *et al.* proposed a method to model and detect spectral transitions for applications of speech or speaker recognition [16]. They investigated the detection of the spectral transition based on time-frequency analysis using models of transitions by linear and quadratic frequency modulation signals. They applied two different detection methods, the Randon-Wigner transform and Randon-Ambiguity transform, and concluded that the detection of spectral transitions helped in the modeling of correlations among parameters of speech frames both in time and in frequency.

These studies were concerned with the detection and modeling of speech transitions for automatic speech recognition and for speech coding. The importance of transitional information to perceptual problems and the relations between energy and intelligibility of the various speech components were not discussed in these studies.

2.2.4 AM and FM Separation and Time-varying Filters

Voelcker suggested that representing signals by sums of sinusoidal components as in Fourier analysis was not appropriate for a time localized description of signals [17]. He proposed a unique way to represent signals as products rather than sums by modeling complex-valued signals as polynomials in complex time. By extending his idea, Rao and Kumaresan developed a method to represent a speech signal as a product of components [18]. They proposed a pole-zero model of a signal by considering periodic extensions of the signal and an algorithm to decompose the speech signal into modulated components. Their motivation for time-domain processing came from a study of the auditory periphery, as opposed to LPC/cepstral analysis motivated by vocal-tract models. They suggested that the human auditory system may be sensitive to modulations. If so, characterizing signals by these modulations may reveal new insight into the nature of speech signals and speaker-specific information. They also suggested that the modulations could be used as AM/FM features in applications such as computer processing of speech. They used a bank of adaptive filters to track each formant component of a speech signal. Each tracked formant component was then decomposed into minimum and all-phase parts, from which the envelope (AM) and instantaneous frequency (FM) information respectively of a tracked formant component were estimated.

Quatieri *et al.* proposed a filter to estimate sine-wave AM and FM information [19]. Their estimation was based on the transformation of FM into AM by filters that were motivated by the hypothesis that the human auditory system uses a transformation of FM into AM for the identification of sine-wave FM. The basic idea of the transformation was that the filter output could be estimated by sweeping the instantaneous frequency (FM) through the filter's frequency range. They described an AM-FM separation algorithm that used two distinct filters. The two

filters had piecewise-linear spectral shapes and were closely overlapped in frequency. The AM and FM components were calculated by utilizing the differences of amplitude envelopes of the two filter outputs. The particular cases of AM-FM decomposition using Gaussian, gamma tone, and auditory filters derived from measured auditory-nerve tuning curves were presented. They emphasized that their algorithm was simply one possible candidate mechanism for auditory FM demodulation. They mentioned that the demodulation assumed that the filter shapes were constant. In the cochlea, however, there may be a fast-acting automatic gain control that can provide nonlinear compression in the main part of the first pass band (the tip of the filter), while leaving the gain in the low-frequency portion of the filter (the tail) unaffected. They suggested that this nonlinear compression could both cause fluctuations in the auditory system output for a low-frequency sinusoid with constant amplitude and reduce the fluctuations in an AM-FM modulated tone. They emphasized that this nonlinearity should be incorporated to understand the complexity of the auditory system.

Boashash and White proposed a method to estimate instantaneous frequency and to design an automatic time-varying filter for non-stationary signals [20]. Their purpose was to reduce noise. The automatic time-varying filter was based on two-dimensional windowing in the time-frequency plane around the estimated instantaneous frequency of the signal, followed by signal synthesis using the Wigner-Ville distribution. The instantaneous frequency of the input signal was estimated by peak magnitude estimation in the short time Fourier transform, and the two-dimensional time-varying windows were designed from the instantaneous frequency and bandwidth. The two-dimensional window was applied to the Wigner-Ville distribution of the input signal, and the output signal was reconstructed from the synthesized Wigner-Ville distribution. The choice of the estimation algorithm for instantaneous frequency determined the

performance of the method. The instantaneous frequency estimation and time-varying filtering were evaluated by applying them to an FM signal with additive noise. The major disadvantages, however, were that the performance of the proposed method depended on the estimation quality of instantaneous frequency and bandwidth.

Francos and Porat suggested a new approach to the design of time-frequency filter banks for non-stationary noisy signals [21]. Multi-component signals were represented by the minimum cross entropy time-frequency distribution, and time-varying filters were applied to the distribution. Each filter processed one component of the signal according to its specific time-frequency support. The outputs of the time-frequency filters were a set of signal components. Their estimation of instantaneous frequency and design of time-varying filters were described as follows. First, they located the peak energy of the minimum cross entropy time-frequency distribution followed by the estimation of time support of the ridge containing the peak energy in the time-frequency plane. They estimated the instantaneous frequency of the located component by a least-square-fit method, and the phase information was estimated from the instantaneous frequency by integration. They de-chirped the multi-component signal by the estimated phase information. As a result, the located component was translated to the low frequency region of the spectrum. They isolated this component by a lowpass filter and estimated its amplitude information. They designed the time-varying filters from the estimated instantaneous frequency and bandwidth information and then repeated the above procedures until all components were processed. They used a synthetic example with white Gaussian noise to illustrate their algorithm. One advantage of their algorithm was that it could identify multi-component signals. However, their algorithm still largely depended on the estimation quality of instantaneous frequency and bandwidth.

Nie, Stickney, and Zeng proposed a method to extract slowly varying amplitude and frequency modulations from speech signals for cochlear implants [22]. A speech sound was divided into fixed sub-bands by a bank of bandpass filters. The amplitude modulations were extracted by full-wave rectifications of the sub-band signals, followed by a lowpass filter. A pair of orthogonal sinusoidal signals at the center frequencies of the sub-bands was used to remove the center frequencies from the sub-band signals, and then the instantaneous frequencies (frequency modulations) were calculated from the in-phase and out-of-phase signals. They pointed out possible inaccurate modulation extractions due to the specific setting of band number and center frequency. They recognized the importance of frequency modulation and suggested to use both amplitude and frequency modulations in order to improve performance of the cochlear implant for noisy speech. They encoded the amplitude and frequency modulations in a limited number of frequency bands to generate synthetic sounds, and the intelligibility of these synthetic sounds was compared to synthetic sounds generated by amplitude modulation only. They conducted psychoacoustic tests with normal hearing listeners and showed improvement in recognition scores when both amplitude and frequency modulations were encoded. They concluded that frequency modulation cues were critical for speech recognition in noise and suggested their use in cochlear implants.

The decomposition proposed in this thesis is intended to produce a "quasi-steady-state" component and a "transition" component. The latter may provide important cues for speech perception [23], [24], [25], [26]. Rao and Kumaresan [18] did not model speech as a sum of AM and FM components. They extracted waveforms for dominant spectral energies (formants) and calculated AM and FM for these waveforms. Specifically, they captured the positive modulations of slowly-varying and dominant energy components on speech with finer time-localized details

of the signal. One advantage of their algorithm is that they provided intuitively reasonable estimates of the AM and FM of the tracked formant. Their estimates of modulations always guaranteed positive AM and FM, and unlike conventional AM and FM decomposition methods, the bandwidths of their estimated modulations did not exceed the bandwidths of the original bandlimited speech when the modulations of bandlimited speech were estimated.

Rao and Kumaresan's algorithm does not require the assumption of a constant filter center frequency, as does the method suggested by Quatieri *et al.* [18]. For a signal with a stationary spectrum (e.g. tone), a fixed bank of bandpass filters may be enough to extract smooth modulations. For non-stationary signals (e.g. speech), bandpass filters, such as Rao and Kumaresan adopted, whose center frequencies are slowly-varying with time to be centered roughly at dominant spectral content may be required to extract the modulations of slowly-varying components. In this study, Rao and Kumaresan's algorithm will be used to identify AM and FM information of "quasi-steady-state" formant components from speech signals. Details of their algorithm are described in Chapter 3. A time-varying tracking filter is developed from the AM and FM information of the formant.

2.2.5 Basis for Tracking Filter Bandwidth

Li *et al.* suggested a perceptual time-frequency algorithm for noise reduction in hearing aids [27]. Their approach provides the basis for the bandwidth of the tracking filters developed in this thesis. They emphasized that existing hearing aids provide little improvement in intelligibility of the signal when background noise is present. This paper presented an integrated approach to the design of a digital hearing aid, based on a wavelet transform, as well as the formation of a temporal and spectral psychoacoustic model of masking. With this model, a

perceptual time-frequency subtraction algorithm was developed to simulate the masking phenomenon and reduce noise in a single input system. A reference noise signal was estimated during quiet periods, and this noise information was used to calculate a weighting function that was computed according to the (signal+noise)-to-noise ratio in each critical band (auditory filter). This weighting function was used to suppress the wavelet coefficients. The weighting function had three different regions: noise masking region in which noise power was strong enough to make the speech inaudible; signal-noise region in which both noise and speech were audible; signal masking region in which the speech was so strong that the noise was inaudible. In the noise masking region, the weighting function was set to zero to minimize the noise signal. The weighting function had a maximum value to pass as much speech as possible in the signal masking region. In the signal-noise region, the weighting function was increased with increasing speech power. The characteristics of the weighting function for the perceptual time-frequency subtraction algorithm are illustrated in Figure 1. For comparison purposes, the transfer function of conventional spectral subtraction, in which the estimated filter is based on the SNR in each frequency band, is also illustrated. Results showed that the use of the perceptual time-frequency subtraction algorithm yielded an improvement in speech quality (increasing noise reduction gain), especially in unvoiced portions. Additionally, the noise component during periods of silence was attenuated significantly.

In the formant-tracking algorithm developed here, the weighting function developed by Li *et al*. will be used to define the bandwidth of the tracking filters. In general, the steady-state portions of formants contain higher energy and the transients contain lower energy. The higher energy is expected to involve more harmonics and to be distributed over broader frequency bands. The bandwidth of the time-varying filter was adjusted, based on the energy of the formant

component, to change with time as the formant energy changes with time, so that the filter would pass most of the high formant energy but reject most of the lower energy portion during transitions.

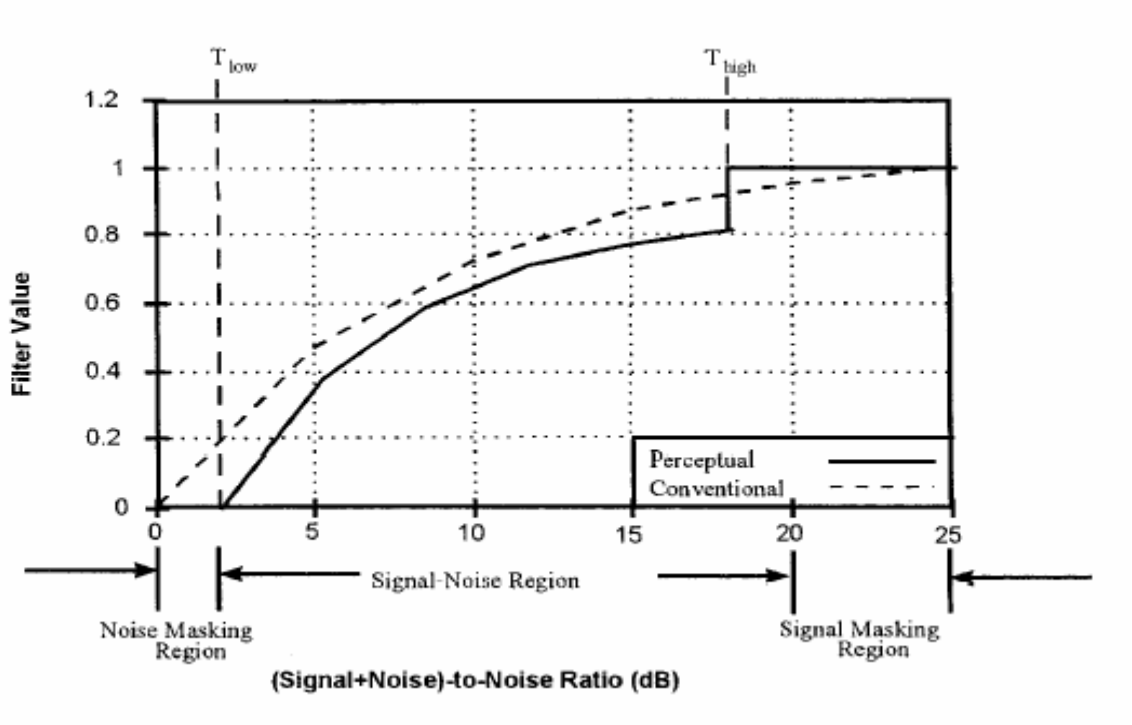

Figure 1: Weighting function to applied perceptual time-frequency algorithm showing the comparison with conventional spectral subtraction. From Li *et al*. [27].

2.3 PSYCHOACOUSTIC TESTS

A psychoacoustic test is an evaluation of auditory perception through audiometric measures requiring voluntary responses from a subject [2], [28], [29]. In general, stimulus sounds are delivered to human ears, and the behavioral responses elicited by these sounds are measured. The stimulus sounds can be controlled by changing their physical parameters (*e.g.* intensities, SNRs, *etc*.). If one-to-one relations between the physical parameters and perception differences exist, the perception characteristics can be quantified by the attributes of the physical

parameters of the stimulus sounds. The purpose of psychoacoustics is to find relations between the physical parameters and perception differences by well designed experiments. Procedures used in this project to evaluate intelligibility of speech, based on [1], [30], [31], and [32], are reviewed in this section.

2.3.1 Intelligibility Test

In order to quantify speech intelligibility, tests based on psychoacoustics are often conducted. Subjects listen to the test words (stimuli) and respond to them. The responses are recorded by the experimenter. Specifically, the subject may write down or speak the words that are heard or choose the closest match from two or more alternatives. The experimenter can control physical parameters of the test words and record how recognition changes due to the parameter changes. For example, the experimenter may vary the intensity of stimulus sounds while asking the subject to identify words that the subject heard and measure the ratio of correct responses as a function of intensity levels. From these results, a psychometric function that shows the percentage of correct responses for different intensities of a stimulus sound can be calculated to establish the relation between intensity levels and intelligibility [1]. The difficulty with intelligibility tests is that the responses are affected by subjects' attentiveness and hearing sensitivity, and test results may vary significantly from subject to subject and from task to task.

The test words can be recorded from one or multiple speakers and presented in quiet or with background noise. Different voice characteristics may cause different results in intelligibility tests. The ability to discriminate speech sounds varies in subjects, and the selection of subject groups depends on the purposes of the experiments.

In intelligibility tests, the number of test words that are correctly recognized by a subject is counted and scored as a percent of correct responses. Mono-syllable consonant-vowel-consonant words can be used in the test [30]. The subject is asked to speak the words that are heard, and the test scores are based on the number of words correctly identified by the subject. The test words can be presented at several intensity levels. A drawback of this test is that direct comparisons of the test scores obtained at different times or different laboratories are difficult because the test scores can gradually increase with repeated testing due to learning. This intelligibility test was used to compare speech intelligibility between original, highpass filtered, tonal, and non-tonal components obtained in this project.

2.3.2 Rhyme Test

Another approach to speech intelligibility testing is the rhyme test. The original rhyme test is a five-alternative closed response test in which the subject is given five word choices in a multiple choice format, and the subject selects one word that most closely matches the word heard [33]. The test was designed to compare performances with different noise canceling microphones. The advantage of this test is that it provides a direct quantitative measure of the intelligibility of a message spoken over any system and requires minimal training of the listeners. In addition, the test stimuli can be repeatedly used with the same listeners with minimal learning effects. Stimulus words are chosen from 250 common mono-syllable words, consisting of 50 sets of five rhyming words. One stimulus word is drawn from each set to form a 50-word test. The rhyming words within a set have a single discriminative feature in the initial consonant (e.g. hot-got-not-pot-lot). Test words can be presented in a quiet background or with background noise. A response sheet shows the 50 sets of rhyming words in order of presentation, and the subject is

asked to mark the words that are heard on the answer sheet. The test scores are calculated from the fraction of the number of words correctly identified.

The modified rhyme test was designed to quantify the performance of voice communication systems to transmit intelligible speech [31]. The test format is similar to the rhyme test described above. The major differences between the two tests are that six alternative words are used instead of five and the test words within sets vary with phonemic elements in word-initial as well as word-final position. Three hundred mono-syllable words are used for test words. This vocabulary consists of 50 sets of six rhyming words each. Twenty five sets differ in the initial word positions, and twenty five sets differ in the final word positions. The attractive feature of this test is the high degree of phonemic balance between the rhyming words, permitting accurate repeated tests. The subject is provided with a response sheet showing the 50 sets of rhyming words in order of presentation. The subject is instructed to mark the word that is heard on the response sheet. The stimulus words are mixed with speech-weighted noise (six different signal-to-noise ratios) before presentation to the subject. The correct and incorrect responses are calculated for each signal-to-noise ratio.

A modification of this word-monitoring task was recently proposed in an effort to improve speech recognition testing sensitivity by incorporating response time measures [32]. The basic idea is that response time can be a supplementary measure to the correct word score. At the beginning of each trial, the target word appears on a computer monitor and remains until all six alternative words are presented. The subjects are instructed to push a button as soon as they hear the target word displayed on the computer monitor. The subjects do not have a second chance to hear the test words. The response time is measured from the end of word presentation to the moment when the subject pushes the button. The test words were presented with speech-

weighted noises at six different signal-to-noise ratios, and correct scores and response times were estimated for each signal-to-noise ratio. The test results showed that response time measures were less sensitive to these different signal-to-noise ratios than were correct word scores. These six different signal-to-noise ratios, however, were selected to show the greatest changes of recognition scores to different signal-to-noise ratios. That is, these six different signal-to-noise ratios were not selected to emphasize the sensitivity of response times to different signal-to-noise ratios). If new signal-to-noise ratios were selected to yield similar recognition scores specifically for the response time experiments, the sensitivity of response times to these new signal-to-noise ratios may be increased. In this study, the word-monitoring task is used to compare speech intelligibility between original and enhanced speech processed by proposed algorithm.

3.0 DECOMPOSITION ALGORITHM

The goal of this study is to decompose speech into tonal and non-tonal components by a bank of time-varying filters and to investigate the intelligibility and energy of the resulting tonal and non-tonal components. The algorithm to be used is described in this chapter. Software modifications for fast computation are also described.

3.1 CONCEPT OF REMOVING FORMANT ENERGY

We assume that a speech signal is a superposition of tonal and non-tonal components as

$$x(t) = x_{ton}(t) + x_{nton}(t) \qquad (3.1)$$

where $x(t)$, $x_{ton}(t)$, and $x_{nton}(t)$ are original, tonal, and non-tonal components of a speech signal, respectively. The estimate of the tonal component is based on the proposed time-varying filters, with center frequencies and bandwidths controlled by the speech signal formants.

We apply three time-varying filters to track the three largest formants. The tracking of formants is implemented using a bank of all-zero filters (AZFs) and dynamic tracking filters (DTFs), in which the center frequency of each of the DTFs tracks a formant of the speech [18]. The FM information of the tracked formants, estimated by linear prediction in the spectral domain, is used to determine the center frequencies of the time-varying filters and to update the pole and zero locations of the AZFs and DTFs. The bandwidths of the time-varying filters are based on the AM information of the tracked formants obtained from the outputs of the DTFs.

The output of each time-varying filter is considered to be an estimate of one of the formants, and the sum of the outputs of the filters estimates the tonal component of the speech. The non-tonal component of the speech signal is estimated by subtracting the tonal component from the original speech signal.

A block diagram of the speech decomposition algorithm, illustrated using two time-varying filters, is shown in Figure 2. The input speech signal is filtered by an AZF and a DTF, and then the FM information (formant frequency) of the output of the DTF is estimated. The estimated formant frequency for the particular formant is used to specify the pole location of the DTF, and the estimated formant frequency from the other filter bank is used to specify the zero location of the AZF. As a result, the AZF suppresses the formant tracked by the other channel so that the DTF follows only one formant of the input speech signal.

The DTF is realized by a difference equation

$$s(n) = r_p e^{j2\pi f(n)} s(n-1) + (1 - r_p) s_I(n) \qquad (3.2)$$

where $s_I(n)$ is the input signal, $s(n)$ is the output of the DTF, and $f(n)$ is the estimated frequency of $s(n)$ [18]. Given a bandwidth of B Hz, the radius of the DTF's single pole, r_p, can be computed as $r_p = \exp(-B\pi / f_s)$, where f_s represents the sampling frequency.

The description of tracking for multiple formants proceeds as follows. Assume there are L formants in a speech signal, and $f_l(n)$ $(l=1,2,\ldots,L)$ represent the individual formants that are to be tracked. To reduce effects of strong neighboring formants, an AZF is applied before the DTF. The zeros of the AZF are adjusted so that a particular DTF effectively filters only one formant. For example, to track the k^{th} formant with frequency trajectory $f_k(n)$, the zeros of k^{th} AZF are located at $f_l(n)$s $(l=1,2,\ldots,L,\ l \neq k)$, using frequency information from the other channels. The center frequency information of the DTFs tracking $f_l(n)$s $(l=1,2,\ldots,L,\ l \neq k)$ are used to determine

the zero locations of the k^{th} AZF. Thus, the k^{th} AZF's output will have only components consisting of the k^{th} formant, and the following k^{th} DTF will track this formant.

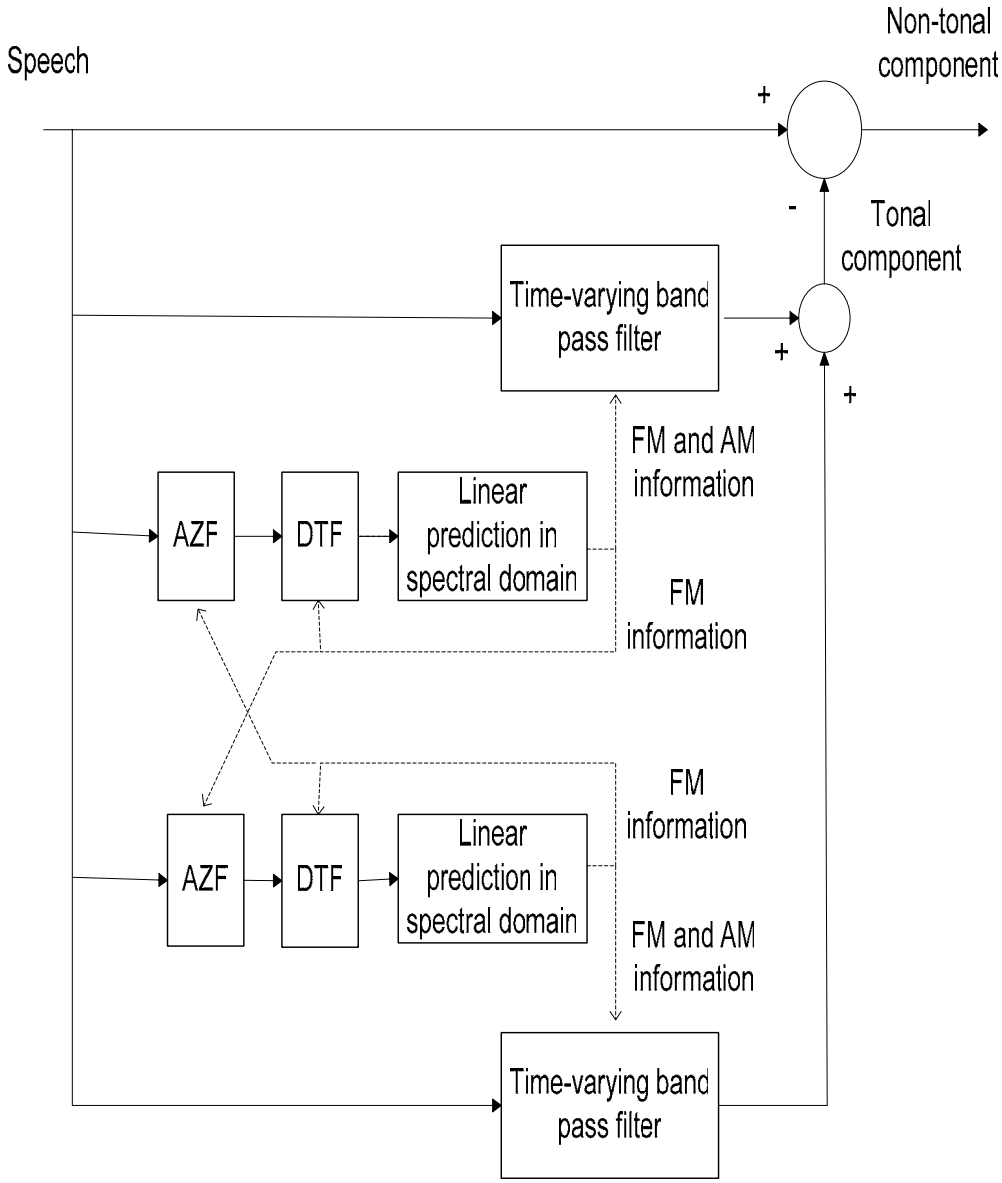

Figure 2: Block diagrams of speech decompositions

The transfer function of the AZF of the k^{th} tracker is

$$H_{AK}(n,z) = K_k(n)\prod_{\substack{l=1 \\ l\neq k}}^{L}(1-r_z e^{j2\pi f_l(n)} z^{-1}) \qquad (3.3)$$

where

$$K_k(n) = \frac{1}{(\prod_{\substack{l=1 \\ l\neq k}}^{L}(1-r_z e^{j2\pi(f_l(n)-f_k(n))})} \qquad (3.4)$$

and r_z is the radius of the AZF's zeroes. The transfer function of the DTF tracking $f_k(n)$ is

$$H_{DK}(n,z) = \frac{1-r_p}{1-r_p e^{j2\pi f_k(n)} z^{-1}} \qquad (3.5)$$

where r_p is as given in Eq. 3.2.

3.2 TRACKING FILTERS

The previous section described the concept of tracking formants using an adaptive filter bank. Each formant of the speech signal is tracked by a combination of an AZF and a DTF. The estimation of envelope (AM) and instantaneous frequency (FM) information from the output of the DTF is described in this section. The method to estimate center frequency and bandwidth of the time-varying filter is presented in Section 3.3.

One method for speech decomposition into AM and FM is to model each individual harmonic component by an AM and an FM component. The AM and FM information can be estimated from the outputs of several narrow filters. The advantage of this approach is that it provides a large number of smooth modulations. The development presented here follows Rao and Kumaresan [18].

A minimum phase signal can be defined as an analytic signal whose log-envelope and phase angle are related by the Hilbert transform. The significance of the minimum phase signal is that all zeros are located inside the unit circle. Hence the minimum phase signal can be completely characterized by its envelope. An analytic signal can be an all-phase signal if its envelope is constant (pure phase signal). The significance of the all-phase signal is that it has a one-sided spectrum and a positive definite instantaneous frequency.

We assume that $s(t)$ is an analytic signal generated from a real finite speech signal by Hilbert transform and filtered by a bank of complex filters. $s_k(t)$ is the output of the k^{th} filter, with a finite bandwidth, B. It is periodic with period $T = 1/B$ sec, and the fundamental angular frequency of $s_k(t)$ can be denoted by $\Omega = 2\pi / T$. Because the spectrum of $s_k(t)$ is concentrated around the center frequency of the k^{th} bandpass filter, $s_k(t)$ can be modeled as a polynomial of sufficiently large degree M in the complex variable $e^{j\Omega t}$

$$s_k(t) = e^{j\omega_t t} \sum_{k=0}^{M} a_k e^{jk\Omega t} \qquad (3.6)$$

where a_k are the complex amplitudes of the sinusoids, ω_t is the nominal carrier frequency of the signal, and $e^{j\omega_t t}$ represents a frequency translation. We may consider that a polynomial of degree M in the complex variable $e^{j\Omega t}$ represents the complex envelope of the signal $s_k(t)$, and this polynomial can be factored into $M=P+Q$ factors, where P and Q are the number of roots inside and outside the unit circle, respectively. Then,

$$s_k(t) = a_0 e^{j\omega_t t} \prod_{i=1}^{P}(1 - p_i e^{j\Omega t}) \prod_{i=1}^{Q}(1 - q_i e^{j\Omega t}) \qquad (3.7)$$

where $p_1, p_2, ..., p_P$ represent the polynomial's roots inside the unit circle and $q_1, q_2, ..., q_Q$ represent the polynomial's roots outside the unit circle: $p_i = |p_i| e^{j\theta_i}$ and $q_i = |q_i| e^{j\phi_i}$. These roots are referred as the complex zeros of the signal $s_k(t)$.

Alternatively Eq. (3.7) can be expressed by grouping the zeros so that the signal can be factored into a minimum phase part that has only envelope information and an all-phase part that has only phase information. The zeros outside the unit circle (q_i) can be reflected to inside the unit circle ($1/q_i^*$) and then canceled by poles at $1/q_i^*$. Therefore, the minimum phase part of the signal is expressed by all the zeros inside the unit circle, and the all-phase part of the signal is expressed by the zeros outside the circle and the poles reflected to inside the unit circle, as

$$s_k(t) = a_0 e^{j\omega_t t} \prod_{i=1}^{P}(1 - p_i e^{j\Omega t}) \prod_{i=1}^{Q}(1 - \frac{1}{q_i^*} e^{j\Omega t}) \frac{\prod_{i=1}^{Q}(1 - q_i e^{j\Omega t})}{\prod_{i=1}^{Q}(1 - \frac{1}{q_i^*} e^{j\Omega t})} \qquad (3.8)$$

---------- ----------

minimum phase part all-phase part

Each factor corresponding to a zero or pole in Eq. (3.8) is referred to as an elementary signal. By expressing the elementary signals as an infinite series, $s_k(t)$ can be represented as a product of a minimum phase signal and an all-phase signal as,

$$s_k(t) = A_c e^{\alpha(t)+\beta(t)+j(\tilde{\alpha}(t)+\tilde{\beta}(t))} e^{j(\omega_c t - 2\tilde{\beta}(t))} \qquad (3.9)$$

---------- ----------

minimum phase part all-phase part

where ω_c is the fundamental frequency (Ω) multiplied by Q plus the arbitrary frequency translation ω_t, $A_c = a_0 \prod_{i=1}^{Q}(-q_i)$, the modulation functions of $s_k(t)$ are given by

$$\alpha(t) = \sum_{k=1}^{\infty}\sum_{i=1}^{P} -\frac{|p_i|^k}{k}\cos(k\Omega t + k\theta_i) \qquad (3.10)$$

$$\beta(t) = \sum_{k=1}^{\infty}\sum_{i=1}^{Q} -\frac{|q_i|^k}{k}\cos(k\Omega t + k\phi_i) \qquad (3.11),$$

and "$\widetilde{\alpha}$" represents the Hilbert transform of α. Eq. (3.9) describes the Hilbert transform relationship between the log-envelope and phase of the minimum-phase signal. The derivative of the phase function of the all-phase signal is always positive and greater than ω_c, in contrast to the derivative of the $s_k(t)$ can have negative values [34]. The AM (envelope) and the FM (positive instantaneous frequency) can be estimated from the separated minimum and all-phase parts of the signal. The first and second exponents in Eq. (3.9) represent minimum and all-phase parts of the analytic signal, respectively. That is, any analytic signal can be represented as a product of two parts; a minimum phase part and an all-phase part (with a positive instantaneous frequency).

To separate the minimum and all-phase parts, the $s_k(t)$ are represented by an all-pole signal model as

$$s_k(t) = \frac{e_k(t)}{h_k(t)} \qquad (3.12)$$

where the error signal, $e(t)$, has a constant (unity) envelope, $h_k(t) = 1 + \sum_{n=1}^{H} h_{nk} e^{jn\Omega t}$ with order H and $\Omega = 2\pi / T$. The decomposition is performed by minimizing the energy of the error signal.

This minimization is accomplished by adjusting the shape of $h_k(t)$ by choosing appropriate coefficients h_{nk} . $h_k(t)$ can be thought of as an output waveform of shape adjustment.

This procedure is similar to the autocorrelation method of linear prediction in terms of flattening the envelope of the input spectrum by error minimization. The major difference between autocorrelation and this procedure is that, in this case, linear prediction is performed on the Fourier coefficients of the signal $s_k(t)$, instead of the signal samples.

The minimization of the energy of the error signal results in $h_k(t)$ being a minimum phase signal whose zeros are inside the unit circle. Since the minimization of the error signal results in an approximation to $s_k(t)$'s envelope, $h_k(t)$ will be an inverse approximation to the minimum phase part of $s_k(t)$ as

$$h_k(t) \approx e^{-(\alpha(t)+\beta(t))} e^{-j(\tilde{\alpha}(t)+\tilde{\beta}(t))} \qquad (3.13)$$

for sufficiently large H.

The residual signal, $e_k(t)$, will be an approximation to the all-phase part of $s_k(t)$

$$e_k(t) \approx Ae^{j(\omega_c t - 2\tilde{\beta}(t))} \qquad (3.14).$$

The FM information, $\omega_c - 2\dot{\tilde{\beta}}(t)$, can be found as the instantaneous frequency of the error signal, $e_k(t)$, which can be computed from the phase difference between neighboring samples. This decomposition procedure, illustrated by the block diagram in Figure 3, is called linear prediction in the spectral domain.

In this section, we described the characteristics of the minimum and all-phase parts of an analytic signal. Linear prediction in the spectral domain, which separates the minimum and all-phase components from an analytic signal, was also described. Each formant of speech can be tracked by the combination of an AZF and a DTF, and the FM (positive instantaneous frequency)

and AM (envelope) information can be estimated from the tracked formant. The output of each DTF is decomposed into minimum phase and all-phase parts by linear prediction in the spectral domain, and a unique FM and AM information are estimated from the all-phase and minimum phase parts, respectively. The AM and FM information is used to determine the bandwidths and center frequencies of a bank of time-varying bandpass filters to identify the tonal component of speech.

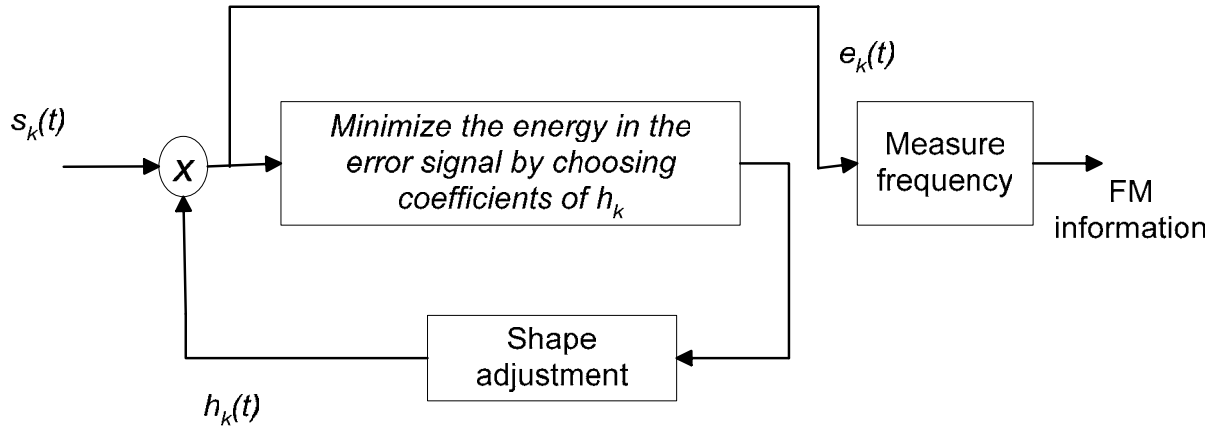

Figure 3: Block diagram of the linear prediction in the spectral domain algorithm. $1/h_k(t)$ is the decomposed minimum phase part and $e_k(t)$ is the decomposed all-phase part. Note that the instantaneous frequency of $e_k(t)$ is positive. From Rao and Kumaresan [18].

3.3 DECOMPOSITION DETAILS

Each time-varying bandpass filter in this bank is a FIR filter with 150 coefficients to provide high frequency resolution between pass and stop bands. The FM information of each

formant (output of DTF) is used as the center frequency of one of the time-varying bandpass filters. Thus, each time-varying bandpass filter will follow the trajectory of one formant of the speech signal.

The bandwidths of the time-varying bandpass filters are determined by the envelope energy of each tracked formant (output of DTF). We assumed that as energy of a formant increases, its bandwidth increases. The basic idea is that the bandwidth of a time-varying bandpass filter following a particular formant should depend on the energy of that formant and should change with time as the energy of the formant changes with time. The concept of bandwidth estimation was developed from the weighting function of Li *et al.* [27]. The decomposition is intended to remove as much of the dominant quasi-steady-state formant energy from the original speech signal as possible, while maintaining reasonable intelligibility in the remaining speech signal. That is, the energy of the tonal component is maximized, while keeping reasonable intelligibility in the non-tonal component. Therefore, if the formant has large energy at a particular time, the time-varying bandpass filter is designed to have a wide bandwidth to successfully pass the wide spread of formant energy. On the contrary, if the formant has small energy at a particular time, the speech is assumed to have significant transient (non-tonal) energy, and the time-varying bandpass filter has a narrow bandwidth to pass only the formant energy.

A maximum bandwidth (B) for the time-varying bandpass filters is selected, and then a function *MBW(t)* for the bandwidth is computed according to the signal-to-noise ratio (SNR) of the tracked formant-energy-to-reference-noise energy. The reference noise is recorded from quiet intervals in the utterance. The SNR is defined as

$$SNR = \frac{s_e(t)}{E[n(t)^2]^{1/2}} \qquad (3.15)$$

where $n(t)$ is the reference noise signal, and $s_e(t)$ is the formant envelope (envelope of each DTF's output), estimated as described in Section 3.2. The $MBW(t)$ is computed as

$$MBW(t) = 0 \quad for\ SNR \leq \alpha$$

$$MBW(t) = 1 - \frac{\alpha}{SNR} \quad for\ SNR > \alpha \qquad (3.16)$$

where α is the bandwidth threshold. The time-varying bandwidth $BW(t)$ is computed as

$$BW(t) = \underline{B} \times MBW(t) \qquad (3.17).$$

The time-varying bandwidths are calculated by the relations between reference noise and formant strengths. That is, the characteristics of the time-varying bandwidths can be described by the SNR of the speech formant energy to the reference noise energy.

The relation of bandwidth to SNR in this approach is illustrated in Figure 4. The SNR is measured by computing an SNR for each short time frame (10 msec.), and the bandwidth is calculated from Eq. (3.17). The bandwidth is set to zero unless the energy in the tracked envelope exceeds the bandwidth threshold. The $MBW(t)$ increases to 1 as SNR increases above the threshold. Zero bandwidth corresponds to the time-varying bandpass filter being "off", and we refer to the filter as being closed. Once SNR exceeds the bandwidth threshold, the bandwidth is increased with increasing SNR, approaching $\underline{B}$ asymptotically. When the bandwidth is non-zero, we refer to the filter as being open. If the energy in the envelope being tracked is large, the time-varying bandpass filter has a wide bandwidth. On the contrary, if the envelope energy is small, the filter has a narrow bandwidth.

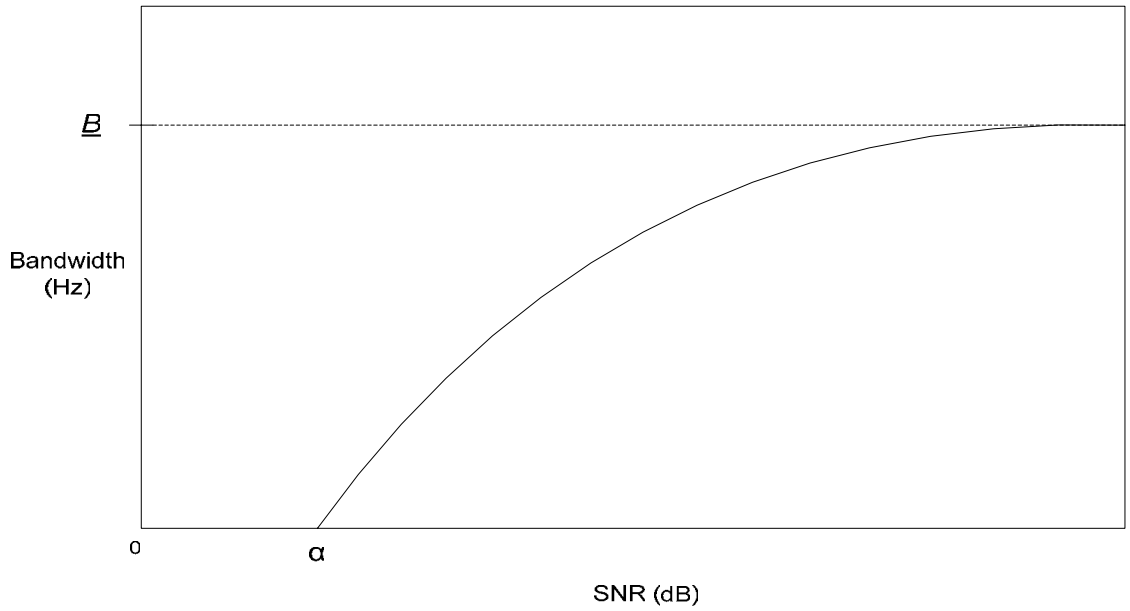

Figure 4: Relation of bandwidth of time-varying bandpass filter to SNR. Based on Li et al. [27]. (see section 2.2.5)

Each time-varying bandpass filter depends on two parameters: the maximum bandwidth (B) and a bandwidth threshold (α) at which the filter is activated. The selection of these parameters is important in the decomposition algorithm. The maximum bandwidth should be large enough to capture most of the energy in the spectral band being tracked but small enough to be restricted to a single band. The activation threshold is based on the ratio of speech to reference noise power in a spectral band. It should be low enough to assure that the filter is active during a sustained sound and high enough to not be active during speech transitions or noise. We examined several maximum bandwidths and bandwidth thresholds for several words, and picked the values 900 Hz for maximum bandwidth and 15 dB for bandwidth threshold as the best tradeoff. The data on which these decisions were based are presented in section 4.3.

The algorithm as described assumes that filter parameters are updated at every sample. The tonal component is dominated by slow-varying formants or "quasi-steady-state" components of speech, and the estimated formant frequency and envelope tend to slowly change with time. A method to improve computation efficiency for tonal estimation is to block speech samples for short time intervals and estimate the formant information by linear prediction in the spectral domain for the first sample in the block. Then, if the formant information does not change within the block, those speech samples within the block have the same formant information. If blocks are small enough for this assumption to be valid, the blocking method provides significant improvement in computation efficiency without affecting the tonal component estimation. Results with different blocking sizes to test this approach are presented in section 3.6.

3.4 ILLUSTRATION

A simple synthetic signal was analyzed to illustrate how the proposed time-varying bandpass filter is formed and how the decomposition algorithm can extract transitional information. The synthetic signal was synthesized at 11.025 kHz and the duration was 127 msec. The signal consisted of three tones with frequencies of 1.5 kHz, 2.8 kHz, and 4.0 kHz and equal amplitudes. The duration of each tone was 53 msec, with linear onset and offset of 7 msec. The synthetic signal was chosen because these three tones and onsets and offsets are similar to the vowel sounds of a simple speech signal.

The number of DTFs in the filter bank was set to 3 to match the number of tones. A white Gaussian waveform was generated and only used as a reference noise signal for SNR calculation. The reference noise signal was not added to the original synthetic signal. The amplitude of the reference noise signal was adjusted so that the SNRs during onsets and offsets changed from 0 to

23 dB. The maximum bandwidth was set to 900 Hz and bandwidth threshold was set to 15 dB SNR.

If the decomposition algorithm is working properly, the tonal component should contain most of the energy of the three tones of the synthetic signal. The non-tonal component should be dominated by onsets and offsets of the tones and contain little energy.

The waveform, spectra, and spectrogram of the synthetic signal are shown in Figure 5. The spectrogram was calculated to describe the time-varying characteristics of the components. The spectrograms were obtained as follows. First, the signal was windowed with a Hanning window with length of 1/10 of the signal, and the spectrum of the windowed signal was estimated by a fast Fourier transform. The estimated spectrum formed one time section of the spectrogram. The window was translated 1 msec and then the above processing was repeated until the sliding window covered the entire synthetic signal. The time-varying characteristics and frequency content of the synthetic signal are effectively described in this spectral plot.

As described in Section 3.1 and 3.2, each tonal component (formant) was tracked by the bank of AZF and DTF, and then FM and AM information of the tracked components was estimated using linear prediction in the spectral domain algorithm. The FM information provided the center frequencies of the time-varying bandpass filters, and the AM information was used to estimate of the bandwidths of the time-varying filters. Figure 6 shows the FM and AM information of the tracked components. The solid, dashed, and dotted lines are associated with the 1st, 2nd, and 3rd time-varying bandpass filters, respectively. As shown in the figures, both FM and AM information properly represent the tonal characteristics of the synthetic signal. The FM information from 0 to 0.030 seconds and 0.097 to 0.127 seconds shows the FM estimation during silent parts.

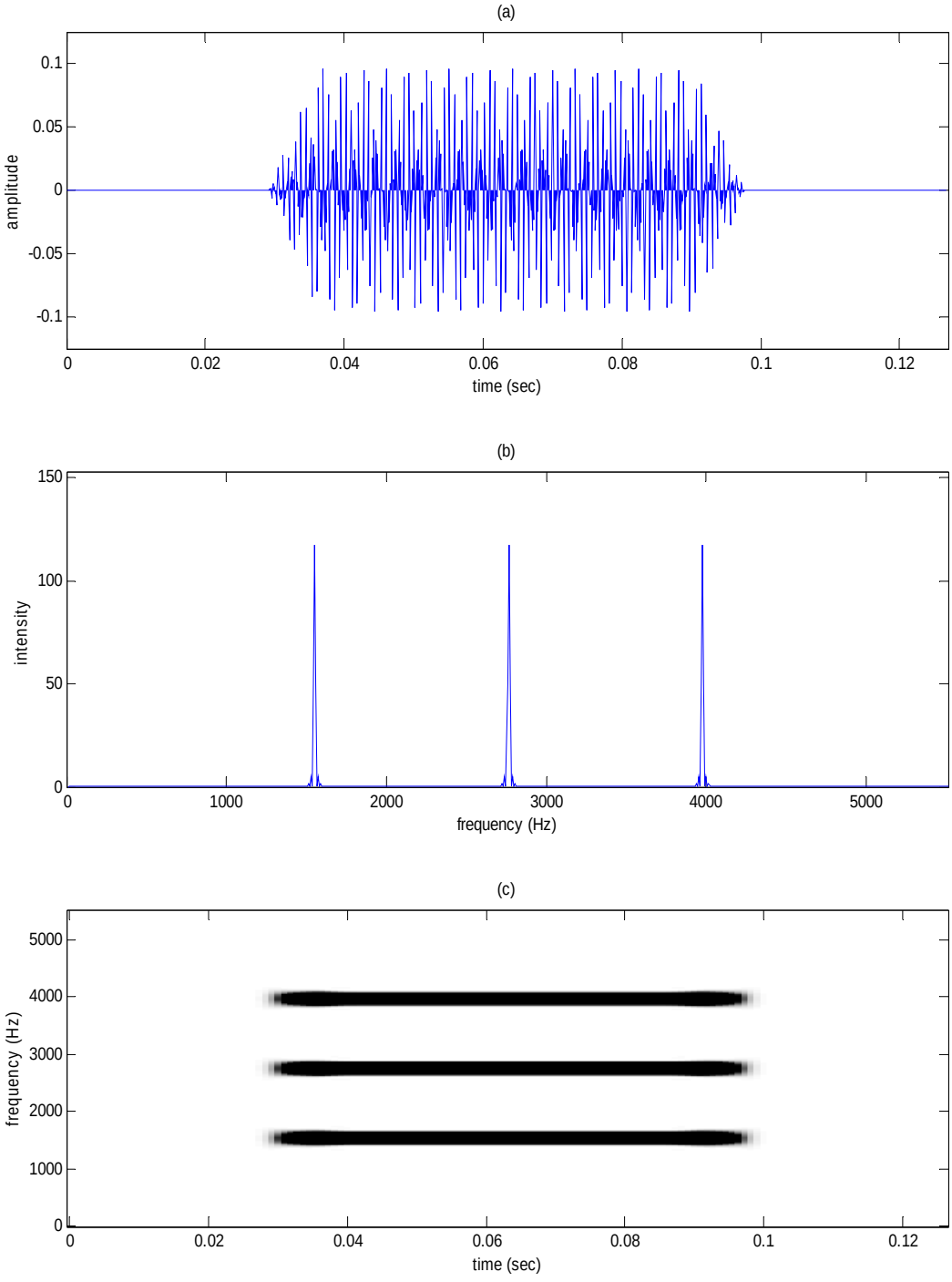

Figure 5: Synthetic signal used to illustrate the algorithm: (a) waveform, (b) amplitude spectrum, and (c) spectrogram

The bandwidths of the time-varying bandpass filters were adjusted following Eq. (3.16) using the SNR values. The SNRs and time-varying bandwidths of each bandpass filter are shown in Figure 7, (a1-3) and (b1-3). The number represents the first, second, and third time-varying bandpass filter, respectively. Each bandwidth is estimated based on the envelope amplitude over a frequency band of the time-varying bandpass filter. When a tonal component is strong, the filter tracking that tonal component has a wide bandwidth, and when the tonal component is weak, the filter has a narrow bandwidth. As illustrated in the figures, the time-varying bandwidths appropriately follow the change of SNRs.

The upper and lower edges of bandwidths of each time-varying bandpass filter are shown in Figure 8 (a). The solid, dashed, and dotted lines are associated with the first, second, and third time-varying bandpass filters, respectively. Each upper and lower edge of bandwidth was calculated for each tracked tonal frequency, estimated from the DTF's output by linear prediction in the spectral domain algorithm. Edges were plotted at plus/minus one-half of the estimated time-varying bandwidth. These upper and lower edges of bandwidths of each time-varying bandpass filter are superimposed on the spectrogram of the original speech in Figure 8 (b). The bandwidths are zero during silent parts of the synthetic signal and gradually opened and closed by increases and decreases in signal energy during onsets and offsets (transitions). The bandwidths are opened enough to pass all tonal components during "quasi-steady-state" parts.

Figure 9 shows the frequency responses of AZF, DTF, and time-varying bandpass filter at a particular moment (0.06 sec. - "quasi-steady-state" part). Note that these plots represent only the one channel that tracks the 1st tone. The AZF suppresses the frequency of adjacent 2nd and 3rd tones, and the center frequency of the DTF is located at the frequency of the 1st tone that the DTF tracks. The center frequency of the time-varying bandpass filter is exactly located at the

frequency of the 1st tone, and the bandwidth is opened enough to pass "quasi-steady-state" energy of the tone.

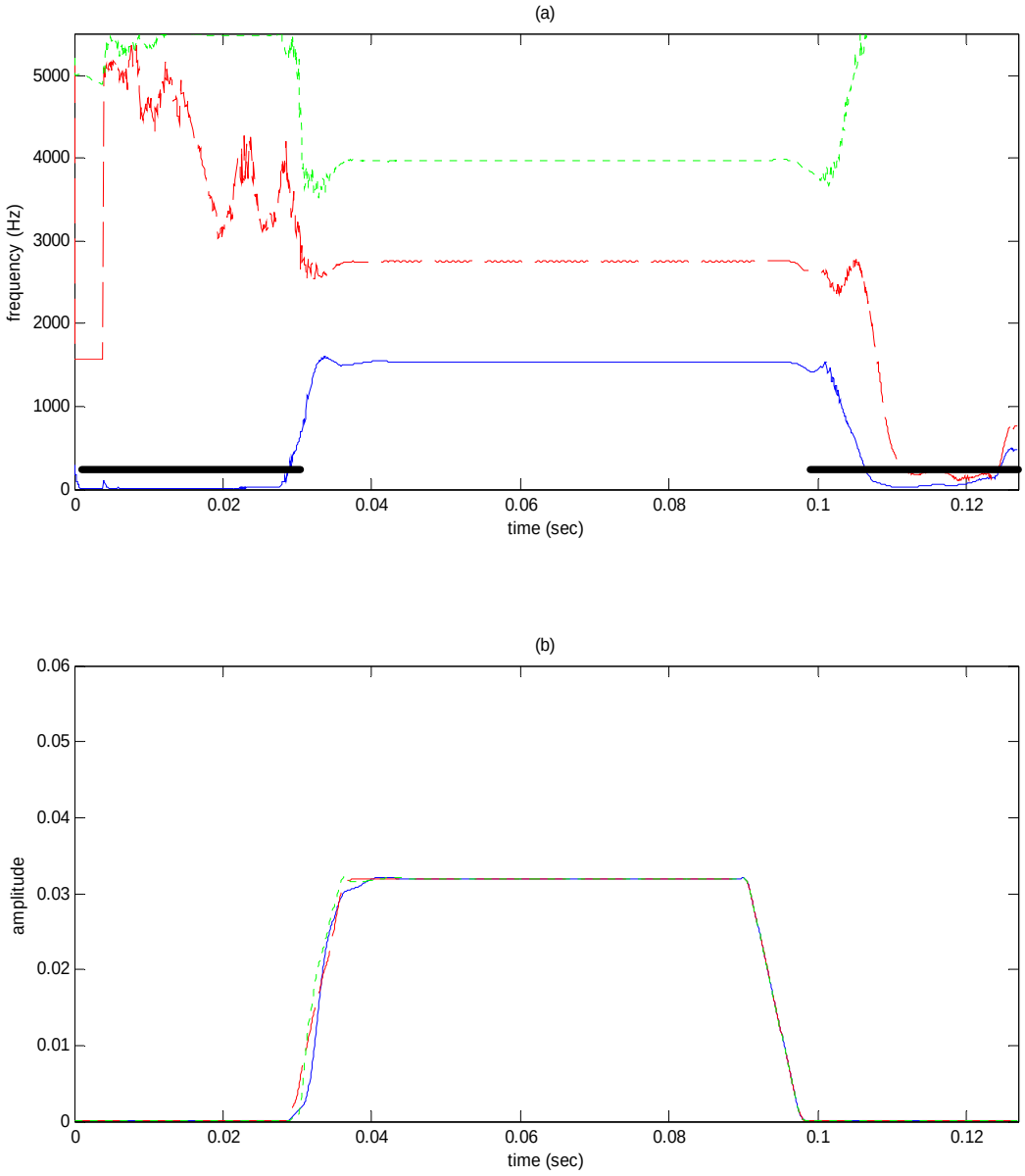

Figure 6: (a) estimated FMs (center frequencies of time-varying bandpass filters) and (b) AMs for a synthetic signal. The solid, dashed, and dotted lines are associated with the 1st, 2nd, and 3rd time-varying bandpass filters, respectively. The bar in (a) indicates the silent part.

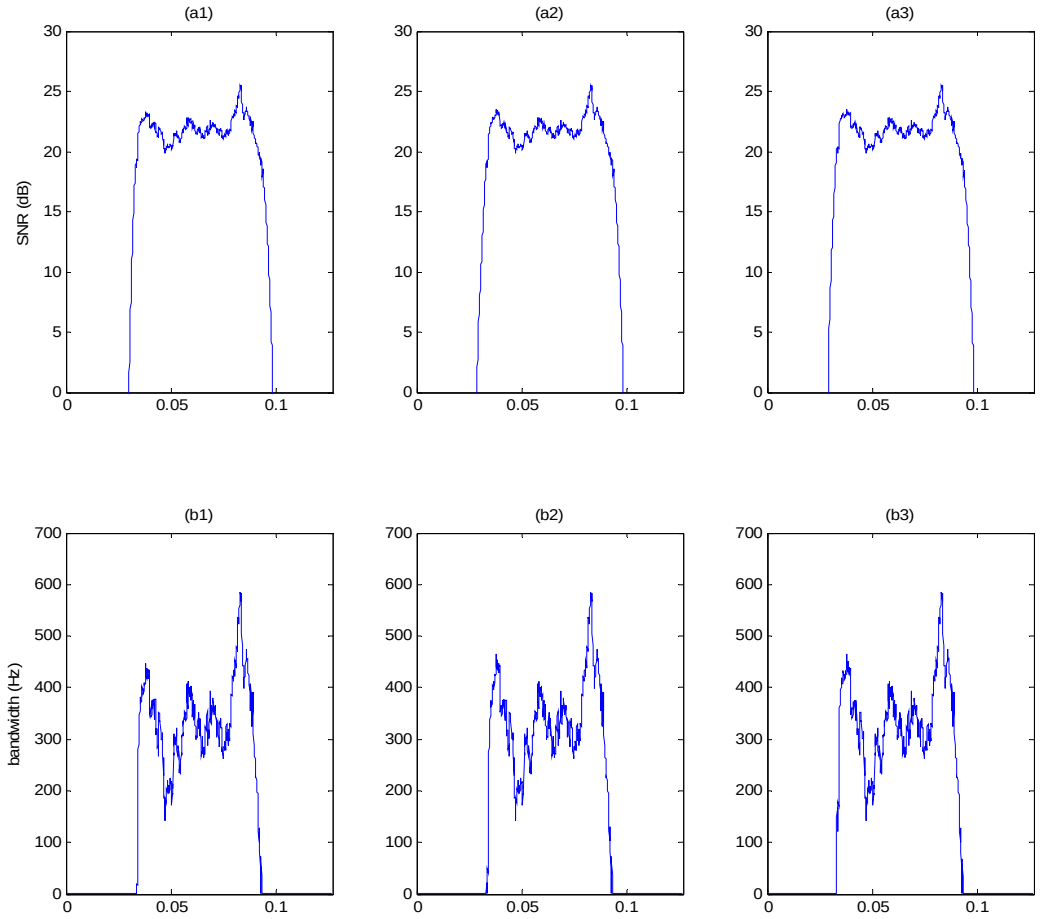

Figure 7: SNRs and time-varying bandwidths of each time-varying bandpass filter for a synthetic signal : (a1-3) SNRs, (b1-3) time-varying bandwidths.

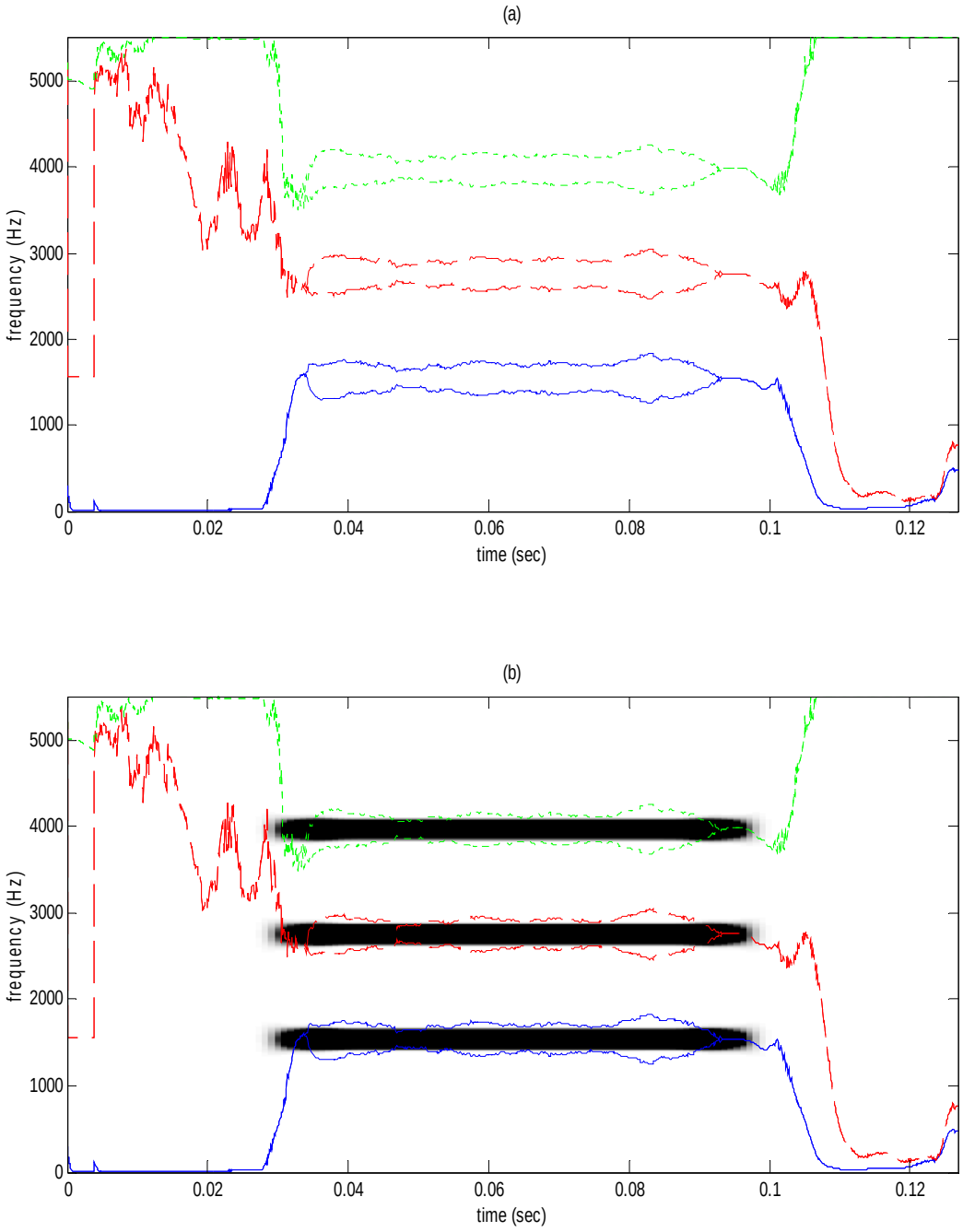

Figure 8: (a) upper and lower edges of time-varying bandwidths and (b) upper and lower edges of bandwidth superimposed on the spectrogram of the original speech. Solid, dashed, and dotted lines are associated with 1st, 2nd, and 3rd time-varying bandpass filters, respectively.

Figure 10 shows the individual outputs of each time-varying bandpass filter. As expected, each filter output includes only the "quasi-steady-state" part of one tone and effectively excludes onset and offset parts (transitions). The waveform, spectra, and spectrogram of the estimated tonal component of the synthetic signal are shown in Figure 11. The steady-state parts of the three tones are effectively passed through the time- varying bandpass filters, and the sum of these filter outputs comprise the tonal component.

The difference between original and tonal components is the non-tonal component. The waveform, spectra, and spectrogram of the estimated non-tonal component of the synthetic signal are shown in Figure 12. The onsets and offsets (transitional components) are appropriately filtered out by the time-varying bandpass filters and shown as the non-tonal component. The tonal and non-tonal components contain 96% and 4% of the total energy of the synthetic signal, respectively.

The results obtained with the synthetic signal demonstrate that the proposed algorithm is able to identify the onset and offset (transition components) of tones. Results with a more complex synthetic signal are given in next section.

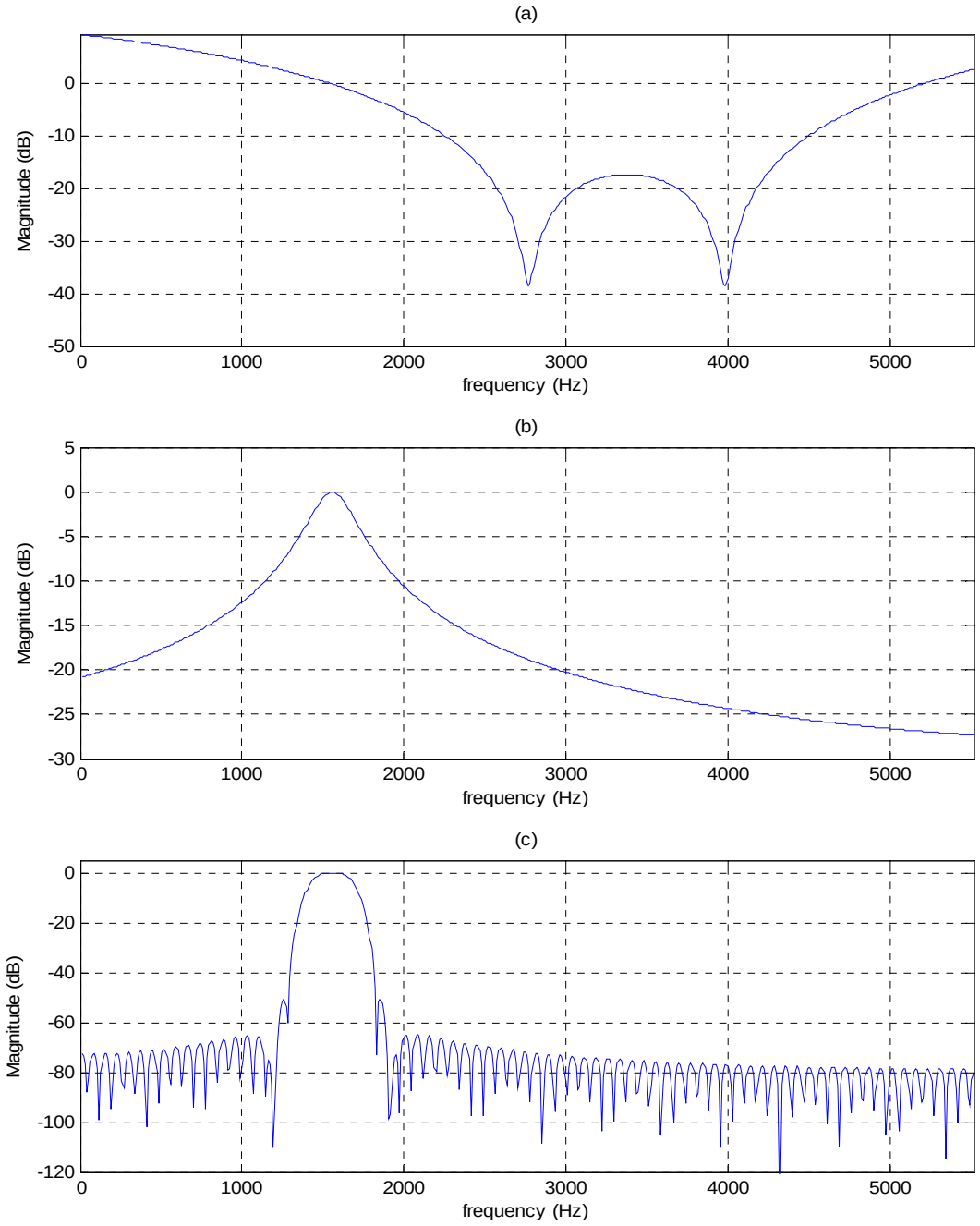

Figure 9: Frequency responses of (a) AZF (b) DTF, and (c) time-varying bandpass filter at 0.06 sec. ("quasi-steady-state" part). Note that these plots represent only the channel that tracks the first tone.

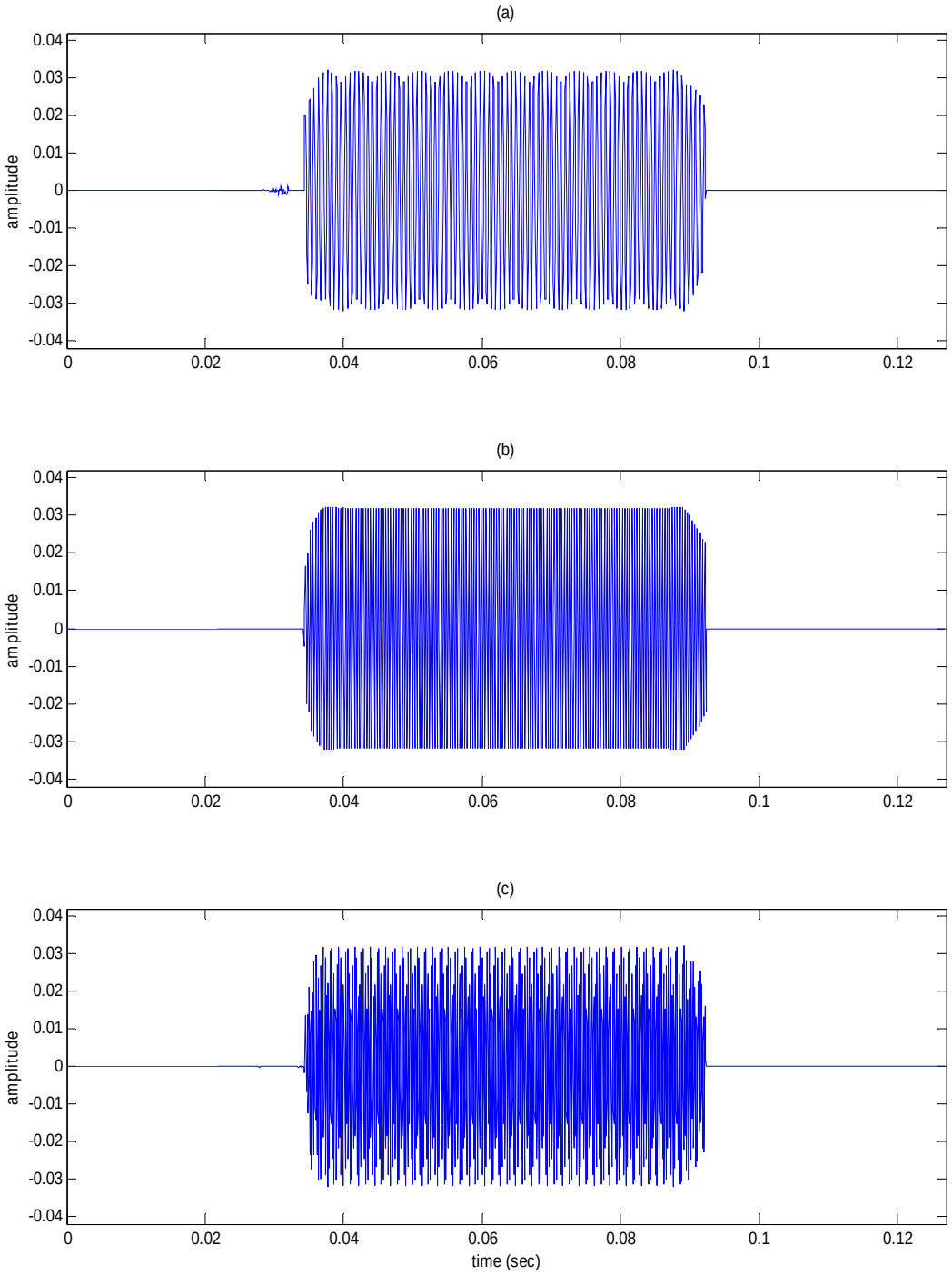

Figure 10: Individual output of each time-varying bandpass filter for a synthetic signal: (a) 1st bandpass filter, (b) 2nd bandpass filter, and (c) 3rd bandpass filter

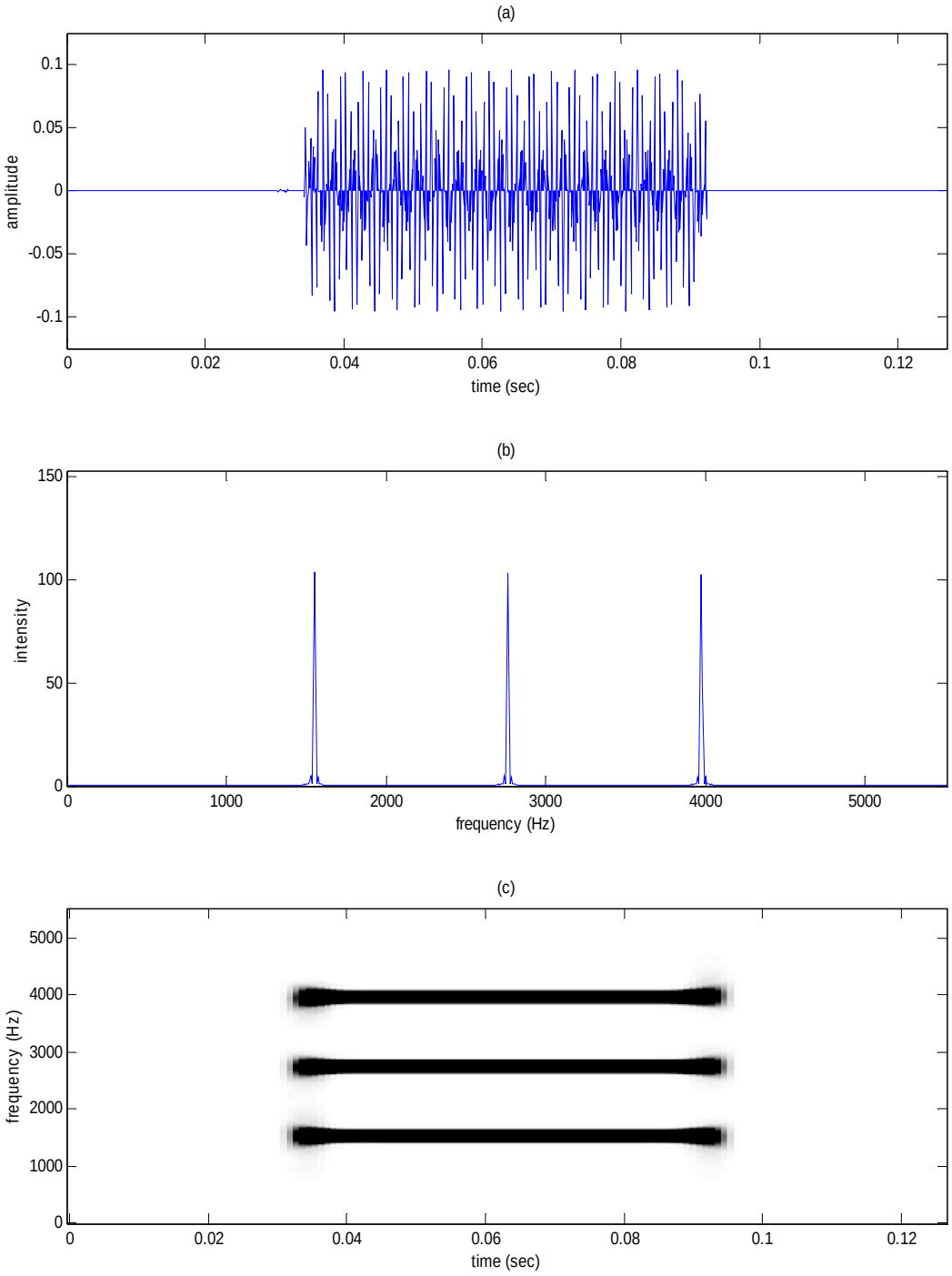

Figure 11: The tonal component of the synthetic signal: (a) waveform, (b) spectra, and (c) spectrogram

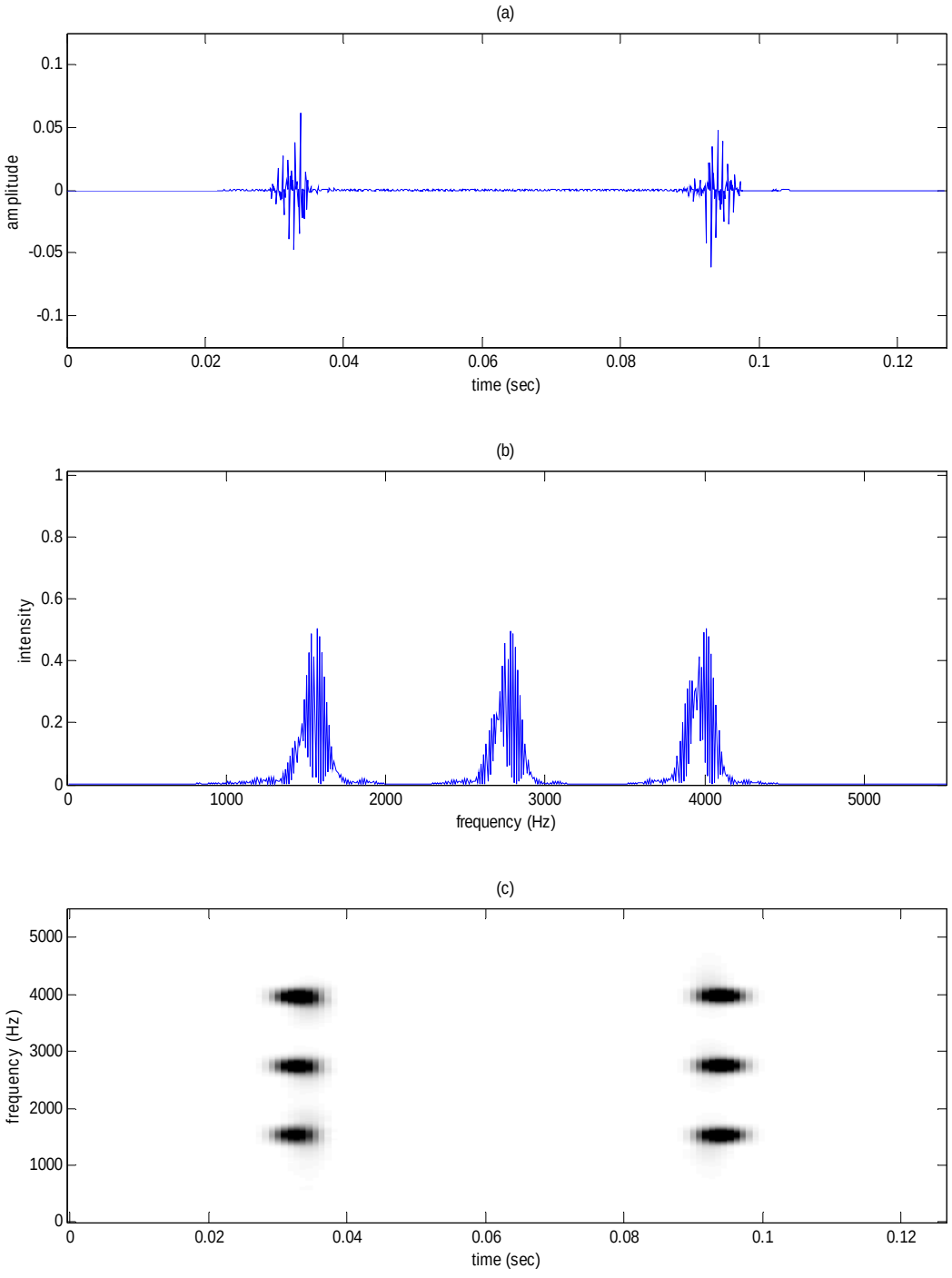

Figure 12: The non-tonal component of the synthetic signal: (a) waveform, (b) spectra, and (c) spectrogram

3.5 FILTER CHARACTERISTICS

The decomposition method used filters with time-varying parameters. Characterizing the filter characteristics is important in describing the decomposition. In this study, the characteristics of the time-varying filters are described by analyzing their response to synthetic chirp signals, known signals with controlled frequency transitions. The chirp rate (Hz/msec) of the synthetic signal was changed by varying either the duration or the frequency change of the chirp. Since the effects of chirp duration and frequency change were not known, both approaches were investigated.

The synthetic chirp signals were intended to represent frequency transitions observed in speech. The time-varying filters should track low chirp rates, so that the chirp portion of the signal is included in the tonal component. The filters should not track high chirp rates, which will be included in the non-tonal component. The purpose of experiments with chirps with varying rates was to determine what rate of frequency transition the filter could follow and how sharply the filters separated tonal and non-tonal components. This study also verified that chirp rate is an appropriate parameter to characterize filter performance.

3.5.1 Synthetic Chirp Signal

The synthetic chirp signal, with structure shown in Figure 13, was sampled at 11025 Hz. The duration was 180 msec. It consisted of three tones (frequencies at F1, F2, and F3), followed by three positive chirps, and then followed by three tones (frequencies at F4, F5, and F6). The duration of each tone+chirp+tone was 140 msec, and each onset and offset was 7 msec.

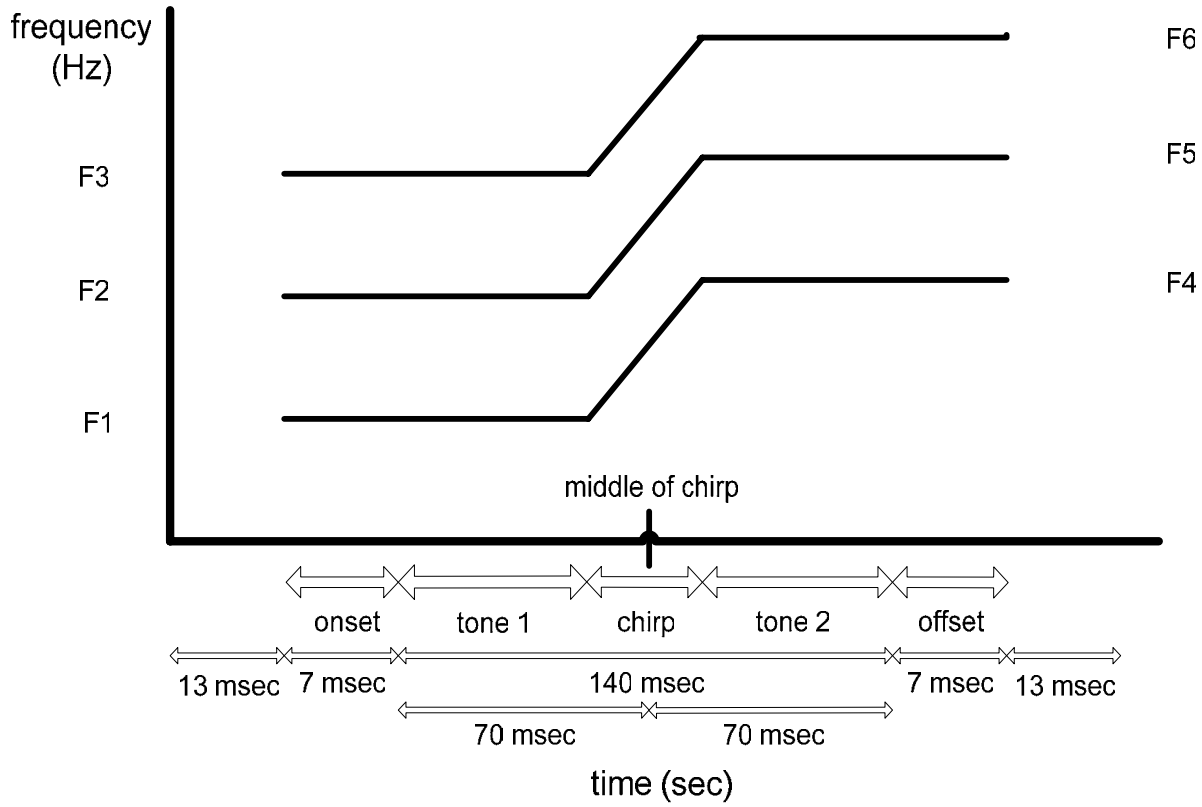

Figure 13: Structure of the synthetic chirp signal

The chirp rate (Hz/msec) of the synthetic signal was changed by varying either the frequency change in the chirp (e.g. fixed chirp duration with different frequency transitions in the chirp) or the chirp duration (e.g. fixed frequency transition in the chirp with different chirp durations). The chirp rates were varied from 24 to 133 Hz/msec.

For the fixed chirp duration with different frequency transitions, the chirp duration was 20 msec and the frequency transitions were varied to produce chirp rates from 24 to 133 Hz/msec.

When the chirp duration was varied, the total length of each tone+chirp+tone was fixed at 140 msec, and the middle of the chirp was always positioned at 90 msec. Therefore, the length

of each tone was varied by the same amount as varying the chirp duration. For example, if the chirp duration was increased by 10 msec, the lengths of the first and last tone were decreased by 5 msec each.

For the experiments with a fixed frequency transition, chirp durations varied from 60 to 11 msec to produce chirp rates from 24 to 133 Hz/msec. Therefore, the chirp intervals ranged from 60-120 msec for 60 msec chirp duration to 84.5-95.5 msec for 11 msec chirp duration.

Synthetic chirp signals, having seventeen different chirp rates, were decomposed into tonal and non-tonal components by the time-varying bandpass filters. The filter characteristics were analyzed by estimating energy of each component in the chirp interval. The energies of the original, tonal, and non-tonal components from 84 to 96 msec were estimated and referred to as chirp energies. The energies of the original, tonal, and non-tonal components from 34 to 46 msec were estimated and referred to as steady-state energies. These chirp and steady-state energies are the energy measurement in a 12 msec window over transition and steady-state intervals respectively, and not measurements of the total energy. The chirp and steady-state energies of the decomposed components were computed as a fraction of the chirp and steady-state energies of the original signal.

We assume that a signal, $x(t)$, is a superposition of tonal, $x_{ton}(t)$, and non-tonal, $x_{nton}(t)$, components as described in Eq. 3.1. Then, the energy of the original signal can be written as

$$\int |x(t)|^2 dt = \int |x_{ton}(t)|^2 dt + \int |x_{nton}(t)|^2 dt + 2\int |x_{ton}(t)x_{nton}(t)| dt \qquad (3.18).$$

That is, the energy of the original signal is the sum of the energy of the tonal and non-tonal components as well as a cross-term $\rho = 2\int |x_{ton}(t)x_{nton}(t)| dt$. If the decomposition is orthogonal, the cross-term is zero. If the tonal and non-tonal components are positively correlated (e.g. positive cross-term), the sum of the relative energies of the tonal and non-tonal components is

less than the energy of the original signal. If the tonal and non-tonal components are negatively correlated (e.g. negative cross-terms), the sum of the relative energy of the tonal and non-tonal components is more than the energy of the original signal.

3.5.2 Analysis Results

Decomposition examples of the synthetic chirp signal are shown in Figures 14 to 16. The synthetic signal consisted of three tones, having frequencies of 574 Hz, 1786 Hz, and 2999 kHz for 60 msec., followed by three positive chirps, having frequencies increasing from 574 Hz to 2514 Hz, 1786 Hz to 3726 Hz, and 2999 Hz to 4939 Hz for 20 msec., and followed by three tones, having frequencies of 2514 Hz, 3726 Hz, and 4939 Hz for 60 msec. The chirp rate for this signal was 97 Hz/msec.

The tonal component is expected to contain the "quasi-steady-state" energy of the synthetic signal, and the non-tonal component should be dominated by onset and offset parts of the tones as well as the chirps between tones. This component should contain relatively little energy of the synthetic signal. The number of DTFs in the filter bank was set to 3 to match the number of tones and chirps. The maximum bandwidth was set to 900 Hz and bandwidth threshold was set to 15 dB SNR. These values were selected for speech analysis based on the results to be presented in section 4.3.

The original, tonal, and non-tonal waveforms decomposed by time-varying bandpass filters are shown in Figure 14, and their corresponding spectrograms are shown in Figure 15. The three tones are effectively extracted by the time-varying bandpass filters as the tonal component. The transitional components (onsets, offsets, and chirps) are appropriately left in the non-tonal component (the difference between the original signal and the tonal component). The tonal and

non-tonal components contain 80% and 20% of the energy of the synthetic signal, respectively. The time-varying characteristics of the decomposed tonal and non-tonal components are illustrated in spectrograms as shown in Figure 15, which demonstrate that tonal and non-tonal components clearly separate the tonal and transitional parts of the signal.

The SNRs and time-varying bandwidths of each bandpass filter are shown in Figure 16, (a1-a3) and (b1-b3). The Arabic number represents the first, second, and third time-varying bandpass filters, respectively. The dashed lines in (a1-a3) represent the 15 dB bandwidth thresholds. The upper and lower edges of bandwidths of each time-varying bandpass filter are shown in Figure 16 (c). The solid, dashed, and dotted lines are associated with the first, second, and third time-varying bandpass filters, respectively.

These upper and lower edges of bandwidths of each time-varying bandpass filter are superimposed on the spectrogram of the original speech signal in Figure 16 (d). The bandwidths are zero during silent parts of the synthetic signal and gradually increase as signal energy increases. The bandwidths are opened enough to pass all "quasi-steady-state" parts but rarely opened during chirps.

The relative chirp energies of tonal components for the fixed frequency transition and the fixed chirp duration are summarized in Table 1 and Table 2, respectively. The relative chirp energies increase with decreasing chirp rates. These results suggest that the time-varying bandpass filters capture more energy when chirps are slowly changing. There are only small differences between constant frequency change and constant chirp duration, suggesting that chirp rate is the most relevant variable to identify the filter characteristics. The chirp energies of the tonal components increased as chirp rate decreased. Changes in relative chirp energy for the

tonal component using constant frequency change and constant chirp duration (dashed line) are plotted together in Figure 17.

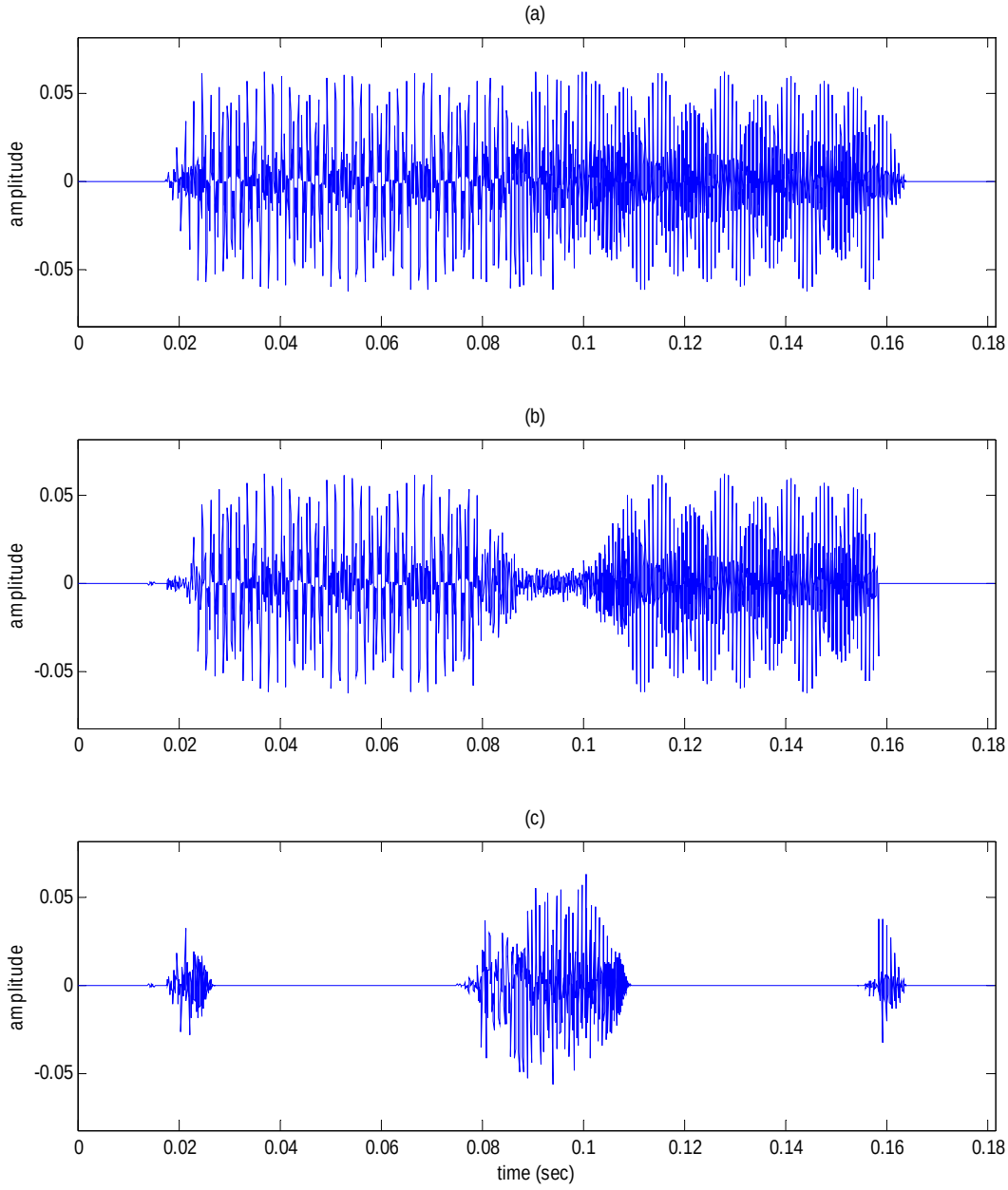

Figure 14: Waveforms of decomposed synthetic chirp signal: (a) original, (b) tonal, and (c) non-tonal components

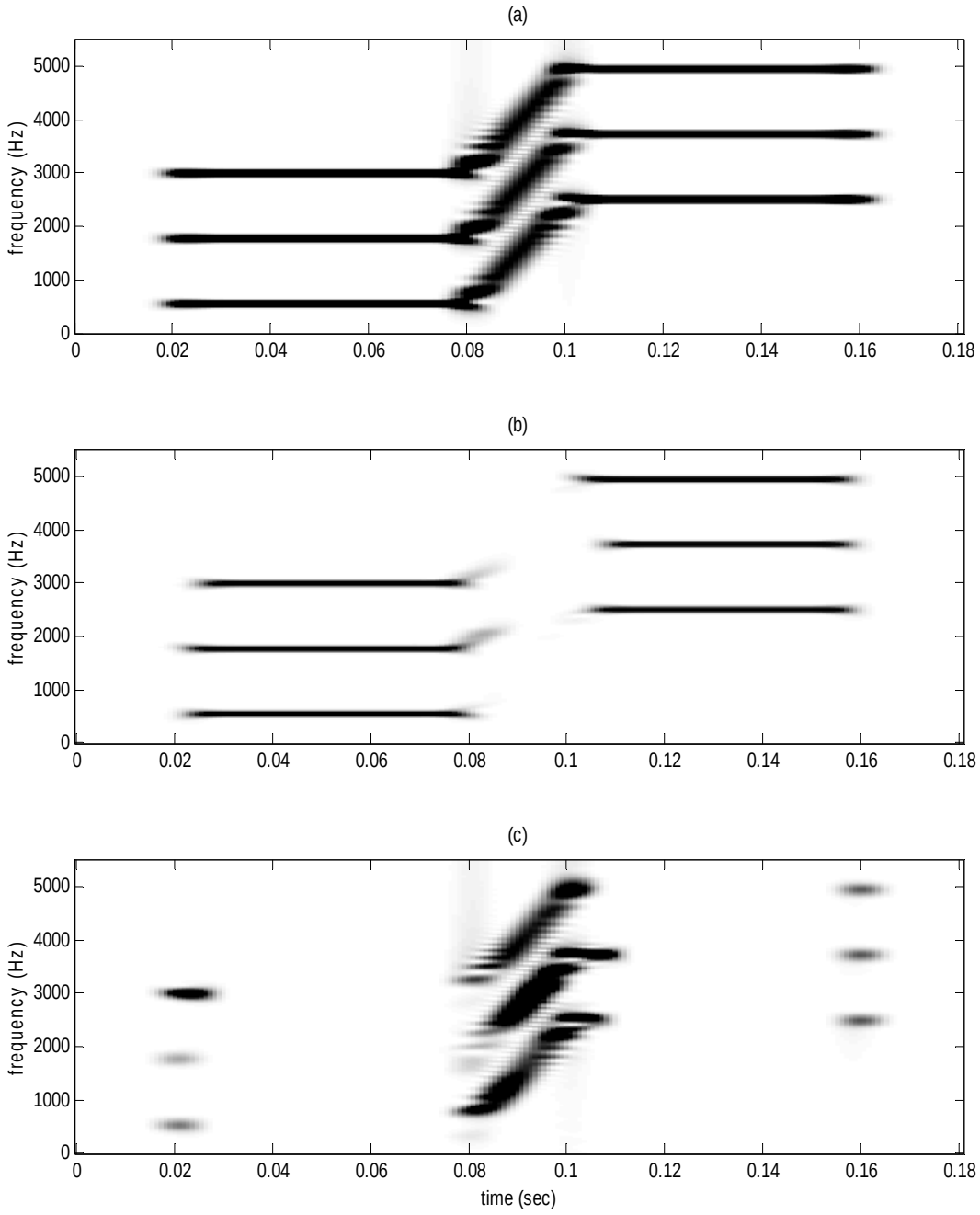

Figure 15: Spectrograms of decomposed synthetic chirp signal: (a) original, (b) tonal, and (c) non-tonal components

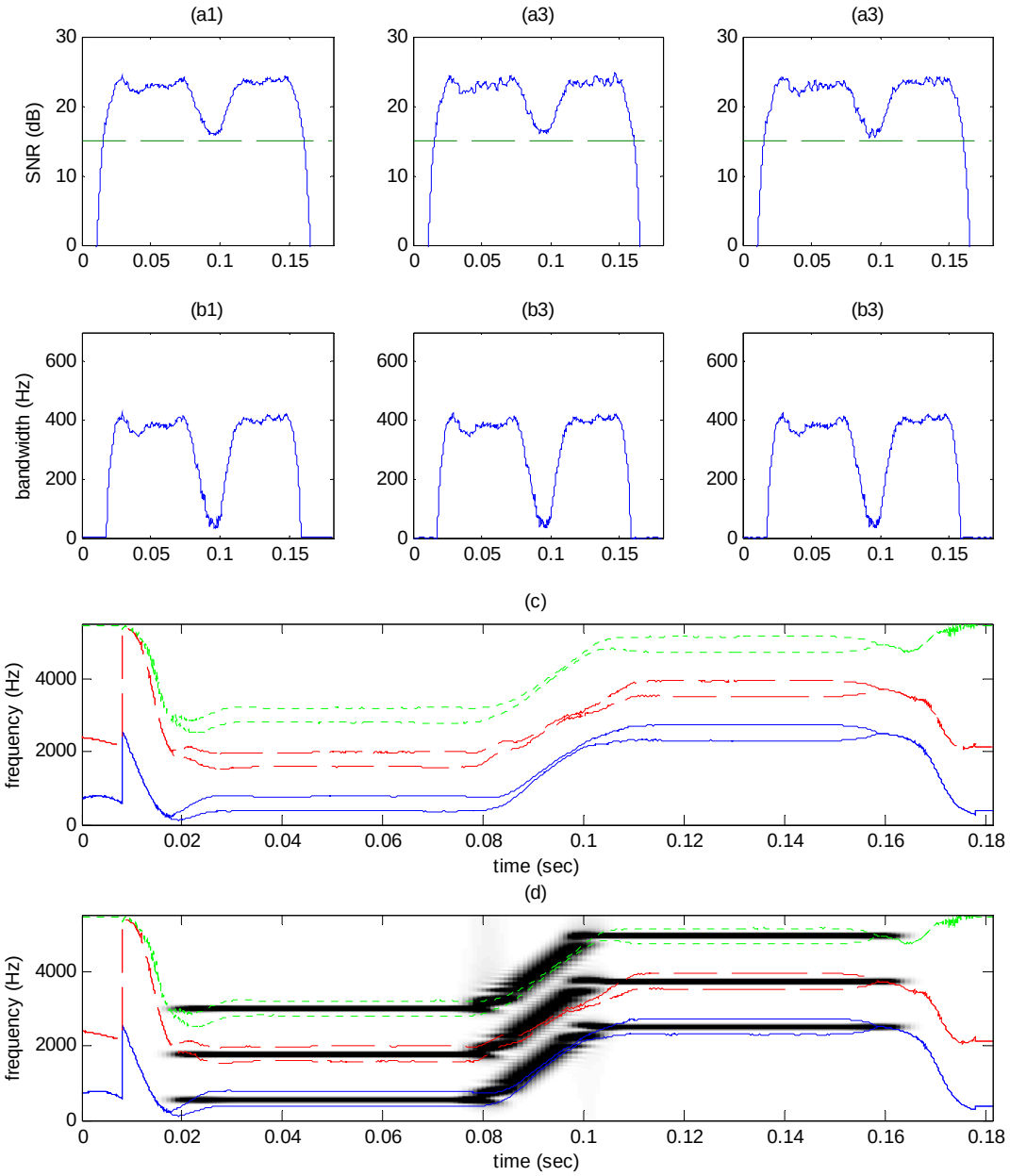

Figure 16: SNRs and time-varying bandwidths of each time-varying bandpass filter for a synthetic chirp signal: (a) SNRs, (b) time-varying bandwidths, (c) upper and lower edges of time-varying bandwidths, and (d) upper and lower edges of time-varying bandwidths plotted with spectrogram. The solid, dashed, and dotted lines are associated with the 1st, 2nd, and 3rd time-varying bandpass filters, respectively.

The chirps can be classified as either tonal or non-tonal components, based on the relative chirp energies. If chirps are slowly changing (slower chirp rates), so that the time-varying bandpass filters capture relatively larger chirp energy (>50%), these chirps are defined as part of the tonal component by the decomposition method. Relative chirp energies are below 50% for chirp rates faster than 73 Hz/msec. Based on this classification method, chirps having chirp rates slower than 73 Hz/msec are classified as tonal components and chirps with faster rates (above 73 Hz/msec) are defined as non-tonal components.

Table 1: Relative chirp energies of tonal components for the fixed frequency transition in chirp and constant chirp duration. Key: E_o: Chirp energy of original synthetic signal, E_t: Chirp energy of tonal component

Chirp rate (Hz/msec)	Chirp duration (msec) for fixed frequency transition (1460 Hz) in chirp	Relative chirp energy in tonal component for fixed frequency transition (1460 Hz) in chirp 100 x (E_t / E_o) (%)
133	11	11
122	12	12
112	13	15
104	14	18
97	15	26
86	17	36
81	18	40
73	20	47
63	23	63
58	25	68
49	30	79
42	35	87
37	40	94
32	45	99
29	50	102
27	55	102
24	60	103

Table 2: Relative chirp energies of tonal components for the fixed chirp duration. Key: E_o: Chirp energy of original synthetic signal, E_t: Chirp energy of tonal component

Chirp rate (Hz/msec)	Frequency transition in chirp (Hz) for fixed chirp duration (20 msec)	Relative chirp energy in tonal component for fixed chirp duration (20 msec) $100 \times (E_t / E_o)$ (%)
133	2660	11
122	2440	15
112	2240	17
104	2080	20
97	1940	28
86	1720	37
81	1620	41
73	1460	47
63	1260	61
58	1160	69
49	980	87
42	840	96
37	740	99
32	640	104
29	580	107
27	540	108
24	480	108

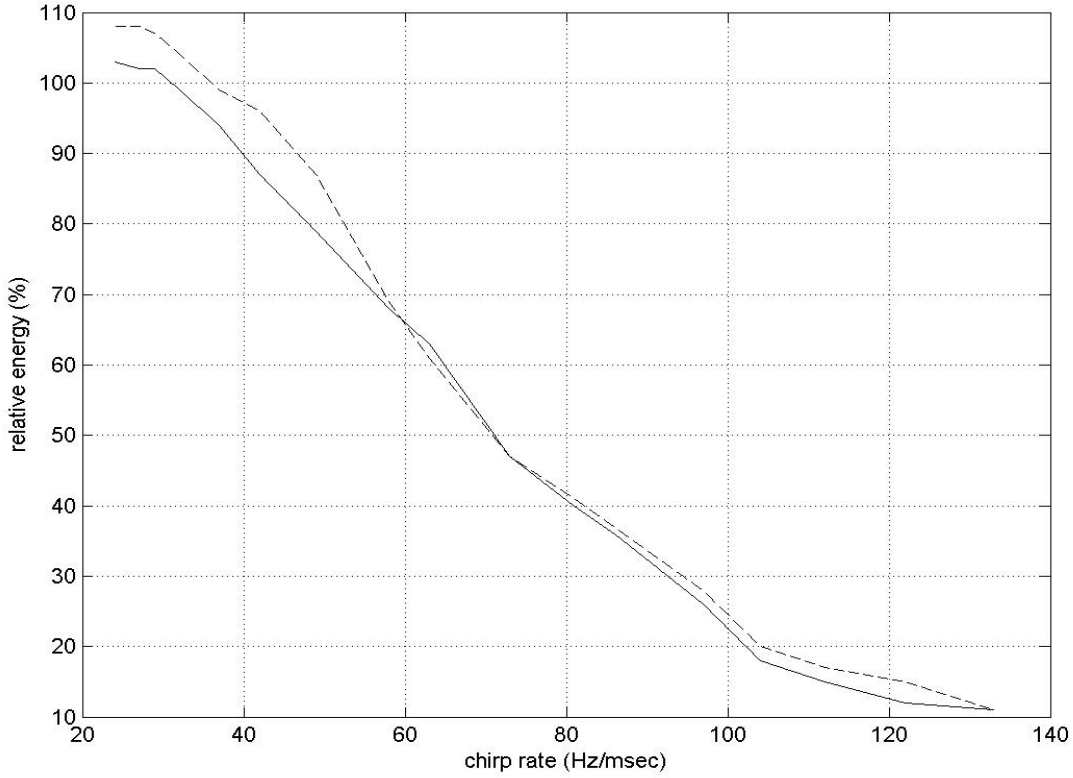

Figure 17: Relative chirp energies of the tonal components for the constant frequency change in chirp (solid) and constant chirp duration (dashed)

The correlation of the cross-term at different chirp rates was examined. The sum of the relative chirp energy (energy measurement in a 12 msec window over transition chirp interval) of the tonal and non-tonal components is illustrated in Figure 18. The solid and dashed lines are associated with the constant frequency change in chirp and constant chirp duration, respectively.

The components have negative correlations at high chirp rates and positive correlations at low chirp rates. For the highest chirp rate, 133 Hz/msec, the sum of the relative chirp energies of the tonal and non-tonal components was 92% for constant frequency change in chirp and 88% for constant chirp duration. For the lowest chirp rate, 24 Hz/msec, the sum of the relative chirp

energies of the tonal and non-tonal components was 106% for constant frequency change in chirp and 109% for constant chirp duration. The relative energy of the cross-term in the steady-state energy (energy measurement in a 12 msec window over quasi-steady-state interval) was also investigated. The cross-term energy was less than ± 0.01% (mean : -0.0002 %).

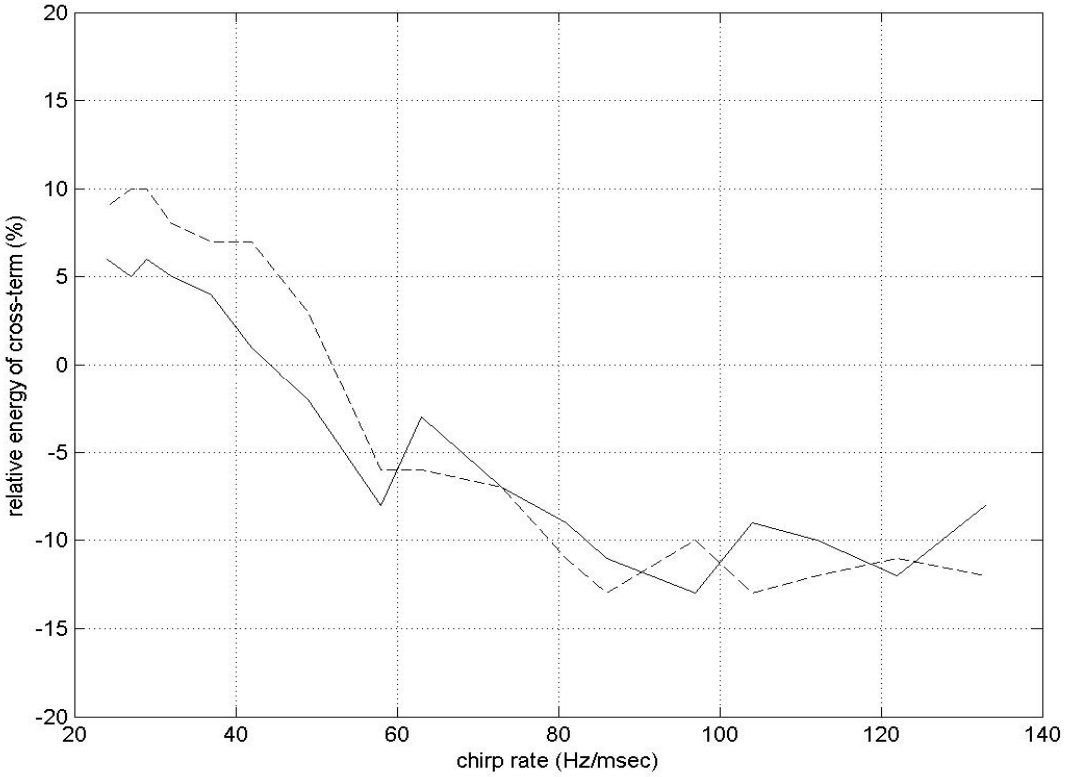

Figure 18: Relative energies of cross-terms for the constant frequency change in chirp (solid) and constant chirp duration (dashed)

The relative energies of these cross-terms were also investigated for real speech samples. Forty three mono-syllable words (twenty three by female speakers and twenty by male speakers) and twelve two-syllable words (spoken by a female and by a male speaker) were examined, as

described in Section 4.2. The relative energies of the cross-terms were less than ± 0.5% (mean : 0.08%) in real speech samples.

Filter characteristics with three negative chirps instead of the three positive chirps were also examined. (Results are summarized in Appendix B.) The structure of the synthetic signal was the same as the positive synthetic chirp signal, except that the first three tones and last three tones were exchanged with each other and the positive chirps were replaced by negative chirps. The same chirp rates were tested, and chirp energy at each chirp rate was estimated. The differences between positive and negative chirp energies were less than 3% in total energy, showing that the algorithm responds the same to positive and negative chirps.

Synthetic chirp signals with an additional 4^{th} tone+chirp+tone were also investigated to determine the effect of a smaller component that is not being tracked. All four tones+chirps+tones had 38 Hz/msec of chirp rates and the chirp durations were fixed at 20 msec. The 4^{th} tone+chirp+tone had 30% of the energy of the other tones+chirps+tones. The 4^{th} tone+chirp+tone was added at either a low frequency (below F1+chirp+F4) or high frequency (above F3+chirp+F6) region. The frequency separations between 4^{th} tone+chirp+tone and F1+chirp+F4 or F3+chirp+F6 were 1 kHz. The tracking filter was not affected by the 4^{th} tone+chirp+tone. (Results are presented in Appendix C.)

3.6 SOFTWARE MODIFICATIONS

The results discussed in section 3.4 and 3.5 involved only synthetic examples, having relatively short durations and identified beginning and ending of sounds. To apply the decomposition algorithm to long speech signals, the speech must be divided into separate segments. To verify that this segmentation does not affect the performance of the algorithm,

several speech samples were analyzed using 0.5 second segments. The energies of the tonal and non-tonal components for five two-syllable and three mono-syllable words were computed using and not using 0.5-second segments, and the energy changes are summarized in Figure 19 (The significance of the 0.9, 1.8, and 2.7 msec. blocks is discussed below.). The relative energies in the tonal and non-tonal components were used to compare the performance of each software modification. Each line represents one word and the average energies across all eight words are indicated by the markers. The windowing method has little effect on the tonal component estimation, as shown in the Figure 19.

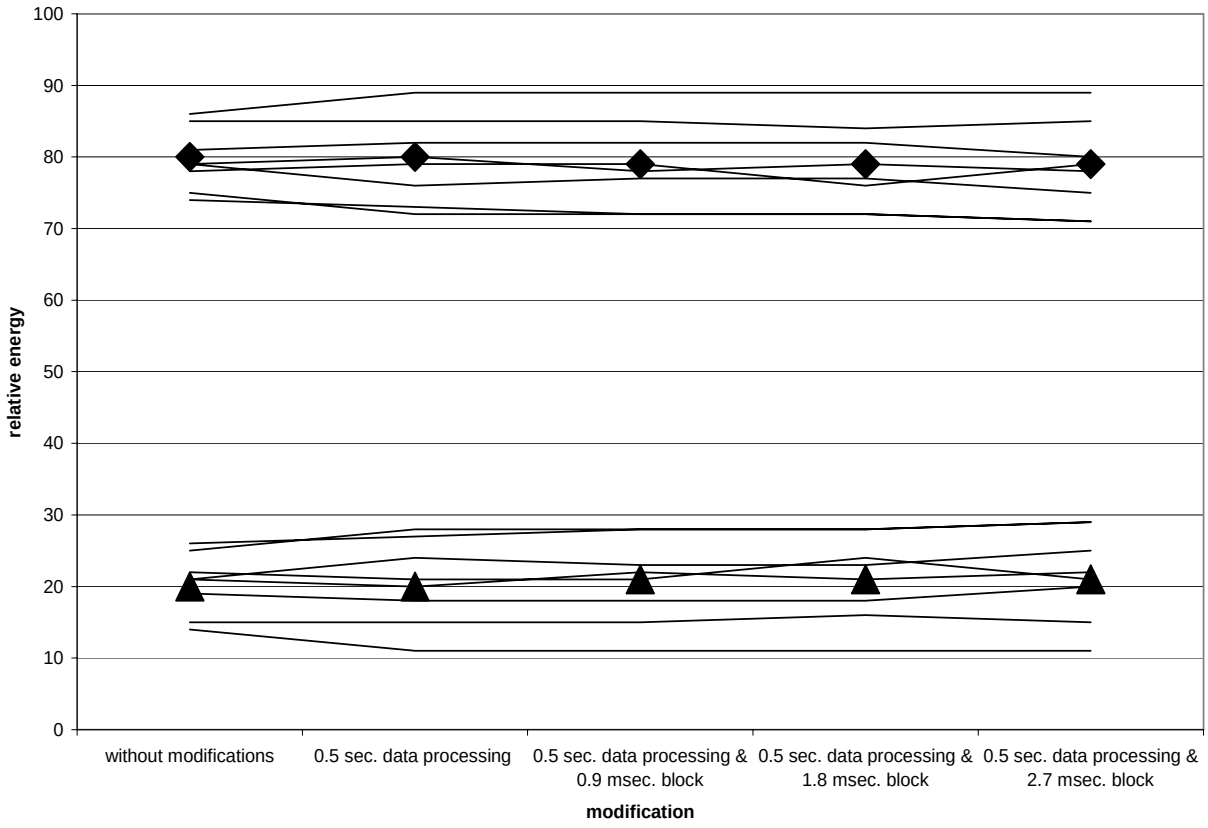

Figure 19: Relative energies in the tonal and non-tonal components with software modifications. The markers represent the average energy of the whole words.

The computation of time-varying filters for the decomposition takes substantial computation time. The software for the speech decomposition is divided into two major parts. The first part consists of calculating the AZFs and DTFs and linear prediction in the spectral domain, and the second part implements the time-varying filtering to estimate the tonal component. The processing times were approximately 314 times real time by a personal computer with 1.6 GHz CPU speeding. Approximately 70% of the total computation time is required for the first part of the algorithm.

As mentioned in the section 3.3, one method to increase computation efficiency is to block speech samples for short time intervals and then estimate the formant information by linear prediction in the spectral domain for the first sample in the block rather than for every sample in the block. To determine an appropriate block size, the synthetic tone signal (same signal as shown in the section 3.4) and synthetic chirp signal (same signal as shown in the section 3.5.1 with 20 msec chirp duration, 580 Hz chirp frequency, and 29 Hz/msec chirp rate) were analyzed using different block sizes. Results are shown in Figure 20. The tonal component estimation is not significantly affected by blocking until 0.9 msec. block. (Energy differences are less than 1.5%.) The energy of the tonal component, however, decreased over 4% and 8% for synthetic tone and synthetic chirp signal for 1.8 msec. block size, respectively. For 2.7 msec. block size, the relative energies of tonal components decreased over 7% and 12% for the synthetic tone and synthetic chirp signals, respectively.

To test this blocking method, the software was modified to apply 10, 20, and 30 data point blocks (0.9, 1.8, and 2.7 msec.). The energy of the tonal and non-tonal components for five two-syllable and three mono-syllable words were computed using the three different block sizes with 0.5 seconds windowing. The energy changes are summarized in the Figure 19. The blocking

method has little effect on the tonal estimation. The energy of the tonal component decreased by an average of 1% as block size increased from 0 to 2.7 msec. The preliminary tests show that this blocking method (10 data point block) can reduce the computation times by approximately 1/6.

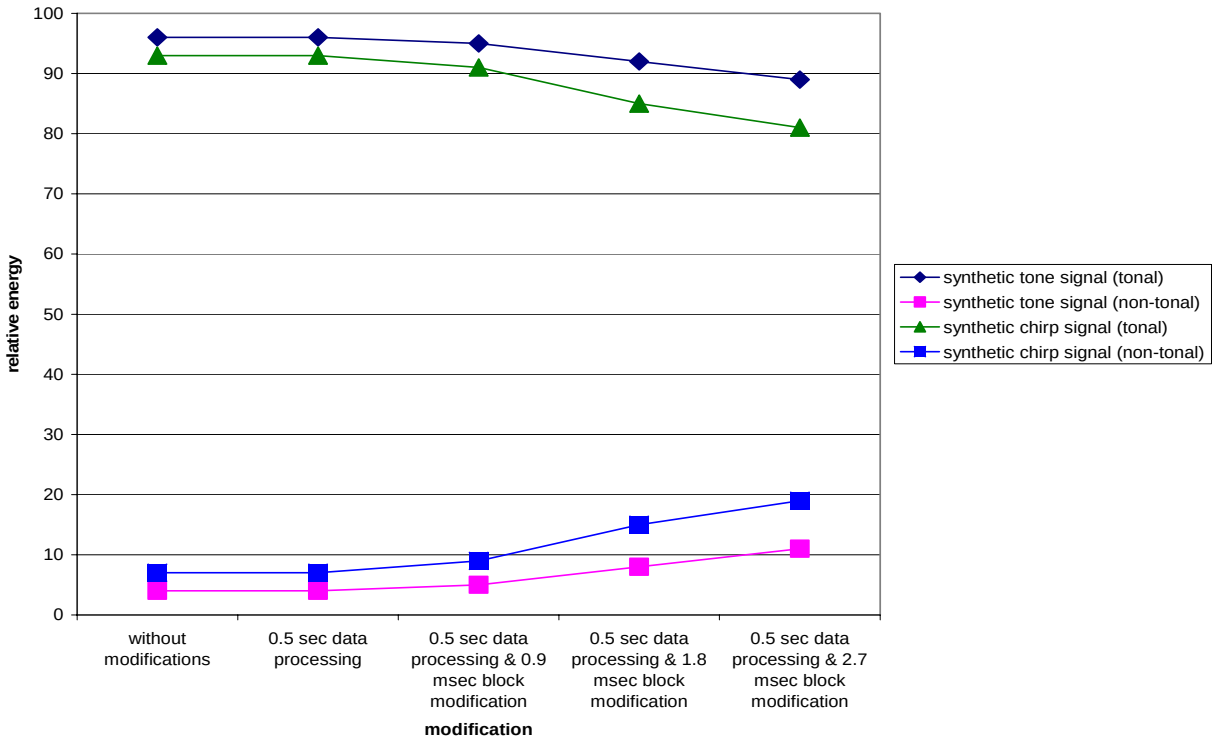

Figure 20: Relative energies in the tonal and non-tonal components of the synthetic tone and synthetic chirp signals with software modifications

One of objectives of this study was to compare the intelligibility of the non-tonal component to the intelligibility of both the original speech and the tonal component. To provide a preliminary test of speech intelligibility, the intelligibility of each component was evaluated subjectively by the author, using a scale from 1 to 5: 1 corresponded to unintelligible and 5 corresponded to the same intelligibility as the original speech. The original speech and each

component were compared by listening through the speaker of a personal computer. Each sound was played by Matlab software (The MathWorks, Inc., USA), and one of the intelligibility levels was assigned to each component.

The intelligibility of the tonal and non-tonal components with data processing using 0.5-second windows and different block sizes was assessed and is shown in the Table 3. As summarized in this table, the intelligibility of the tonal and non-tonal components was not changed by these software modifications. Based on these results, the 0.9 msec. block size was selected as the largest that should be used for speech decomposition. This length had little effect on the test words, and it minimized the effect of blocking on how quickly the filters responded to frequency changes in the test chirps.

Table 3: Relative intelligibility in the tonal and non-tonal components with software modifications

Words \ Modification	Without modification		0.5 sec. data processing		0.5 sec. data processing & 0.9 msec. blocking		0.5 sec. data processing & 1.8 msec. blocking		0.5 sec. data processing & 2.7 msec. blocking	
	Tonal	Non-tonal	Tonal	Non-tonal	Tonal	Non-tonal	Tonal	Non-tonal	Tonal	Non-tonal
Sunset2	1	5	1	5	1	5	1	5	1	5
Cowboy1	1	5	1	5	1	5	1	5	1	5
Cowboy2	2	5	2	5	2	5	2	5	2	5
Headlight1	1	5	1	5	1	5	1	5	1	5
Headlight2	1	5	1	5	1	5	1	5	1	5
Nice1	1	5	1	5	1	5	1	5	1	5
Room1	2	4	2	4	2	4	2	4	2	4
Juice1	1	5	1	5	1	5	1	5	1	5

A long sentence was analyzed to examine possible artifacts due to the software modifications. Examples for a long sentence decomposed by the modified software using a 0.5-second window and 0.9 msec. blocking are illustrated in Figures 21-24. The sentence ("How to feel about changing the time when we began work") was spoken by a male speaker and is approximately 2.76 seconds long. The original, highpass filtered at 700 Hz, tonal, and non-tonal components are shown in Figure 21 (The reason for interest in highpass filtered speech is presented in the next chapter.). The energy in the highpass filtered speech is 6% of the energy in the original speech. The energy in the tonal component is 74% of the energy in the highpass filtered speech (4% of the original speech energy). The remaining 26% of the energy of the highpass filtered speech is in the non-tonal component (2% of the original speech energy). The spectrograms were calculated to describe time-varying characteristics of decomposed components and are shown in Figure 22.

The waveforms and spectrograms of the long sentence from 0.25 to 0.75 seconds are shown in Figures 23-24. Possible artifacts due to the software modifications (windowing) were examined by listening to the sounds and zooming the plot of the signal. The tonal and non-tonal components sounded similar to results obtained with individual words. No artifacts were detected from the examination.

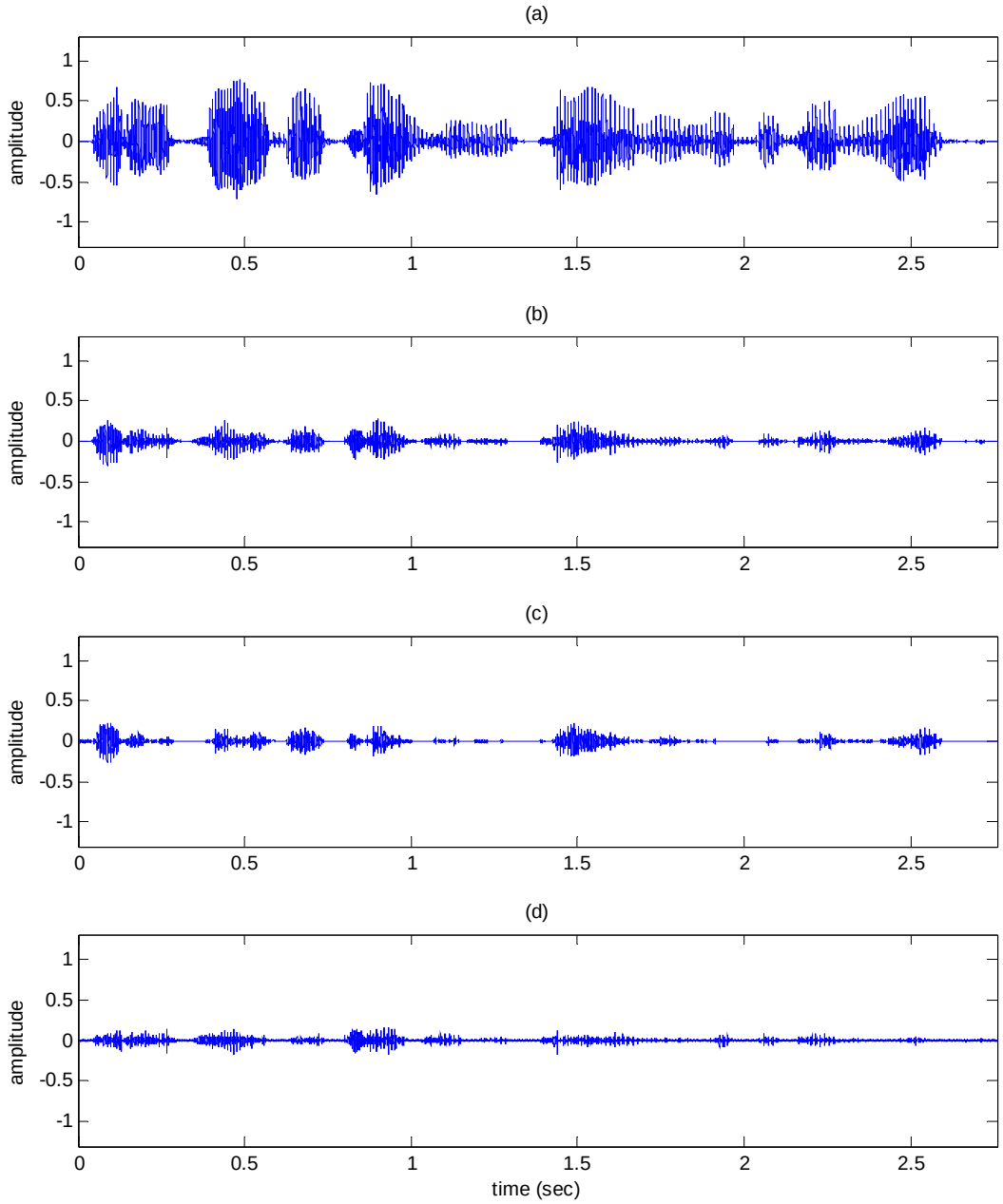

Figure 21: Waveforms of decomposed long speech signal spoken by a male speaker (corresponding to "How to feel about changing the time when we began work"): (a) original, (b) highpass filtered, (c) tonal, and (d) non-tonal components

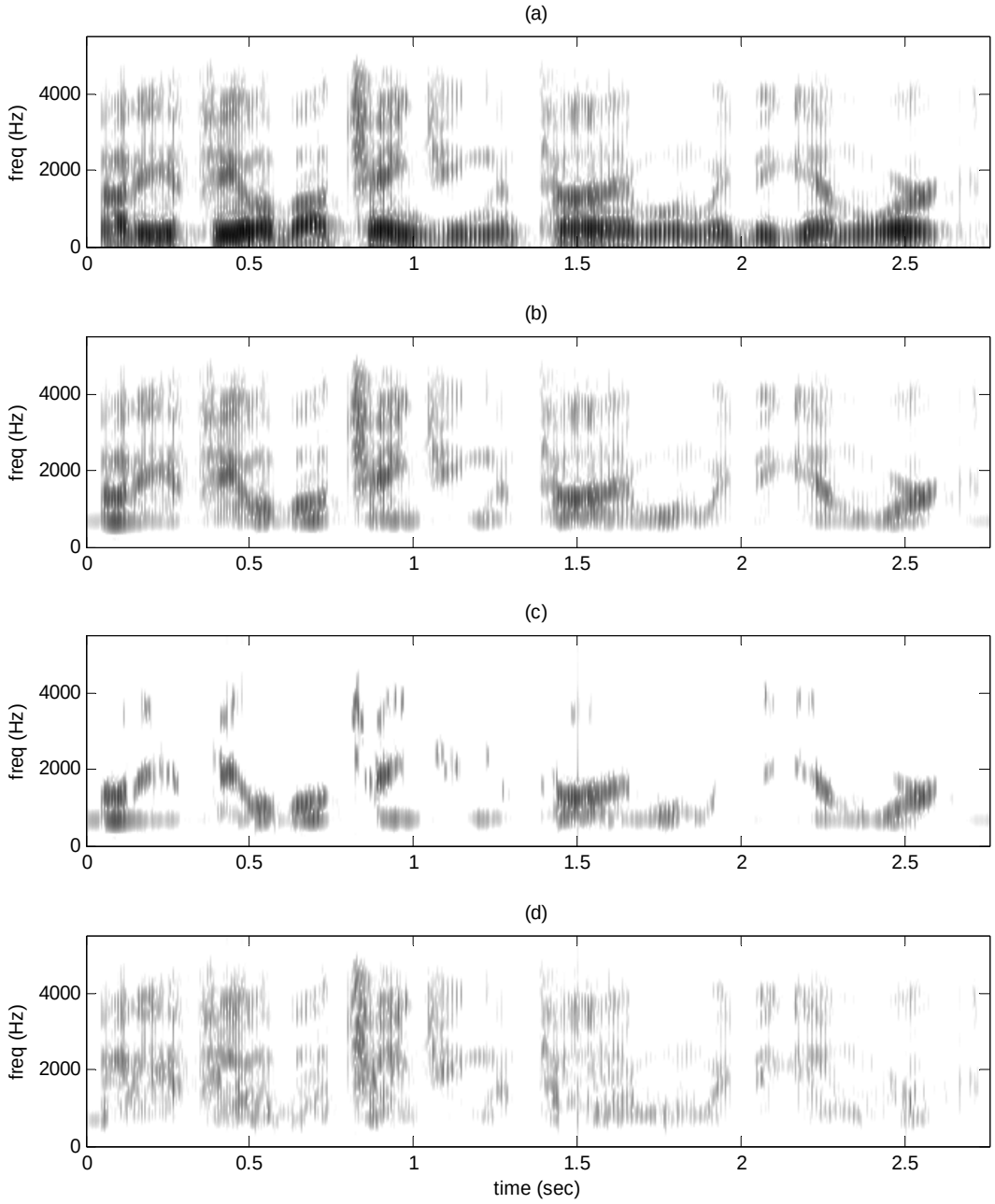

Figure 22: Spectrograms of decomposed long speech signal spoken by a male speaker (corresponding to “How to feel about changing the time when we began work”): (a) original, (b) highpass filtered, (c) tonal, (d) non-tonal components

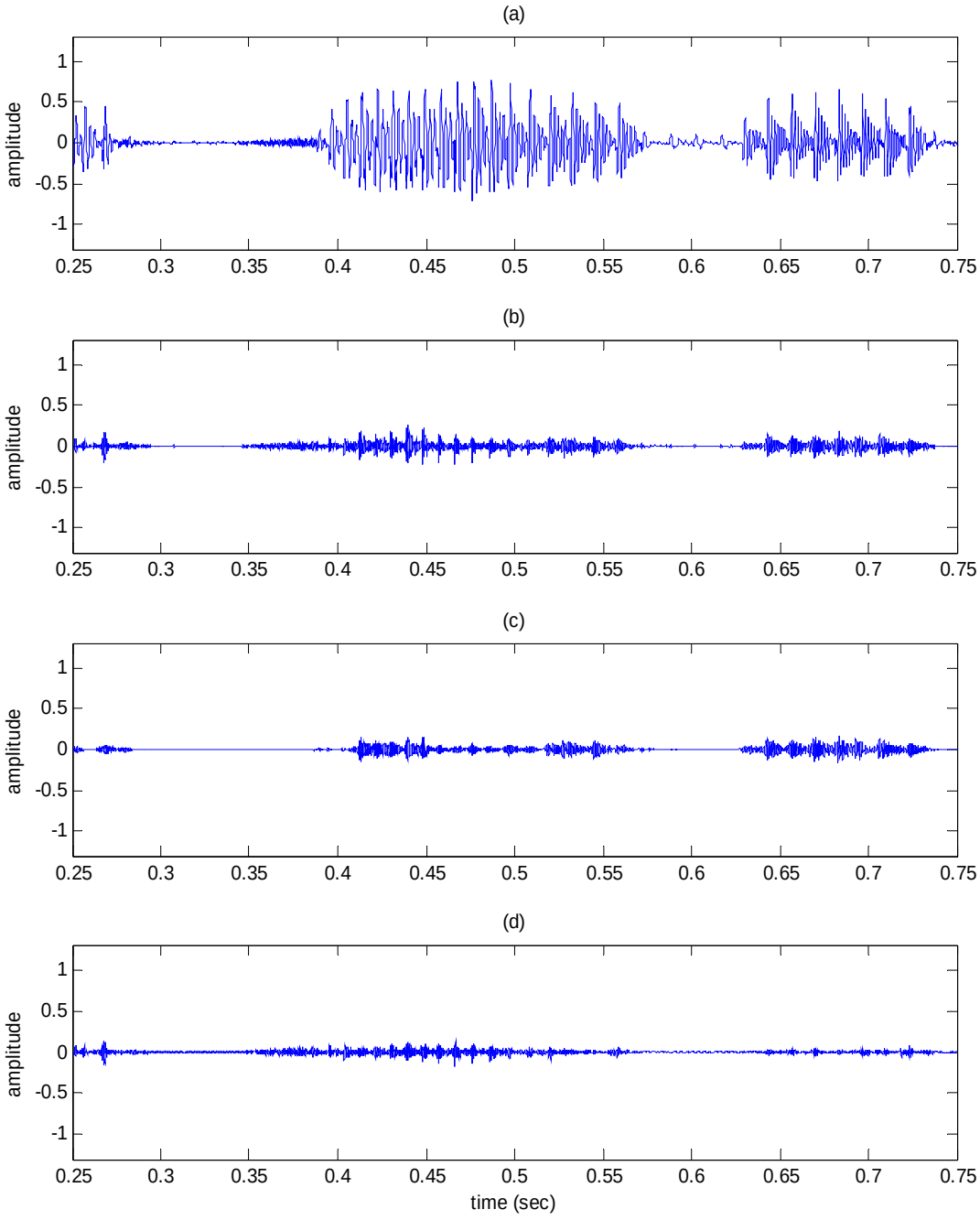

Figure 23: Waveforms of decomposed long speech signal from 0.25 to 0.75 seconds: (a) original, (b) highpass filtered, (c) tonal, and (d) non-tonal components

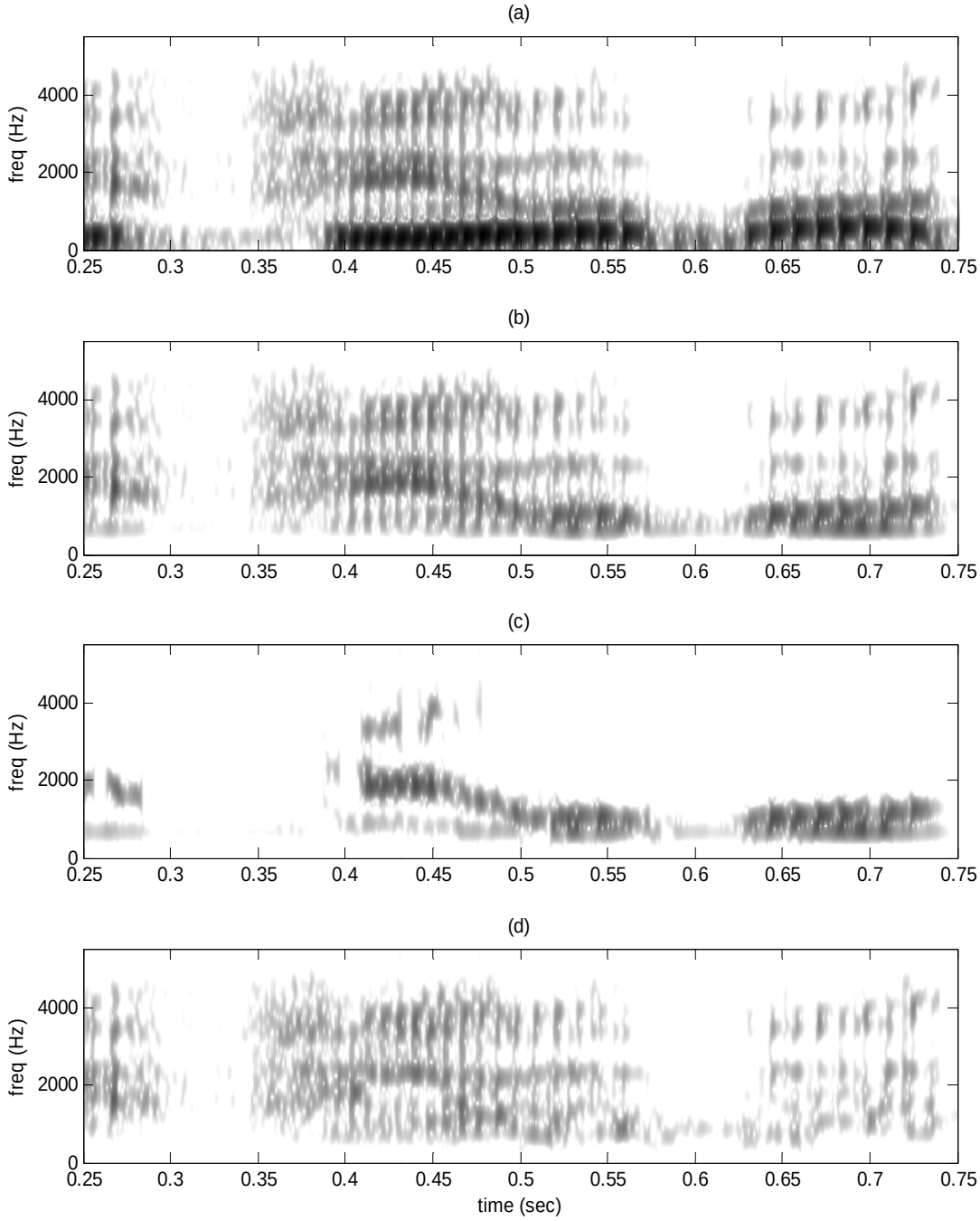

Figure 24: Spectrograms of decomposed long speech signal from 0.25 to 0.75 seconds: (a) original, (b) highpass filtered, (c) tonal, (d) non-tonal components

4.0 PRELIMINARY SPEECH RESULTS

The decomposition of real speech signals with one female and two male speakers was examined. Twelve two-syllable words, spoken by a male and by a female speaker, twenty mono-syllable (Consonant-Vowel-Consonant) words spoken by two male speakers, and twenty three mono-syllable (Consonant-Vowel-Consonant) words spoken by a female speaker (from the audio CDROM that accompanies Contemporary Perspectives in Hearing Assessment, by Frank E. Musiek and William F. Rintelmann, Allyn and Bacon, 1999) were investigated. Speech words were decomposed using the time-varying bandpass filters and results of speech decompositions are described in this chapter. Preliminary intelligibility tests for decomposed components are also presented. These studies were used to select filter parameters.

4.1 DATA PROCESSING DETAILS

Speech signals, sampled at 44.1 kHz, were down-sampled to 11.025 kHz, and highpass filtered with 700 Hz cutoff frequency. Highpass filtering was used because, in unfiltered speech, the first DTF usually tracked a tonal component below 700 Hz. The power near the center frequency of the first tracker was usually large enough to hold the filter open, and the first time-varying bandpass filter effectively functioned as a lowpass filter with approximately a 700 Hz cutoff frequency. The energy of the tonal and non-tonal components obtained from unfiltered speech with four trackers and highpass filtered speech with three trackers showed little difference, and the intelligibility of the tonal and non-tonal components was not changed

between the two methods. Highpass filtering does not affect speech intelligibility [35] but it significantly improved the computational efficiency of the decomposition.

The number of DTFs was set to three because most vowel sounds that have been highpass filtered at 700 Hz are composed of two or three dominant formant components. For each filter, the maximum bandwidth ($\underline{B}$) was set to 900 Hz, and the filter activation threshold (bandwidth threshold) was set to 15 dB. These values were selected based on the results described in section 4.3.

If two adjacent formant components are too close in frequency, the bandwidths of these bandpass filters may overlap in some time intervals, and the outputs of these adjacent filters may contain some energy from the same formant. This overlapping of energy results in the tonal component having too much energy. This situation was avoided by limiting the bandwidth of one of the bandpass filters to avoid overlap between bandwidths. Specifically, if two adjacent bandwidths are close enough to be overlapped, the algorithm increases the low-end bandwidth of the filter tracking the higher formant frequency to prevent overlapping.

Constant reference noise energy, derived from silent parts of a single speech phrase, was used in the preliminary word decompositions. In essence, time-varying bandwidth was computed based on speech signal power.

The intelligibility of original speech, tonal, and non-tonal components was evaluated subjectively by the author as described in section 3.6 (1 corresponded to unintelligible and 5 corresponded to the same intelligibility as the original speech). The energy of each decomposed component relative to the highpass filtered speech was estimated. The amount of relative energy is used as an indicator of how effectively formant information is being removed from the highpass filtered speech while maintaining intelligibility of the non-tonal component. The

conclusions drawn from these subjectively evaluations were verified by psychoacoustic growth functions, as described in chapter 5.

4.2 PRELIMINARY RESULTS

An example of decomposition of a real speech signal spoken by a female speaker is illustrated in Figures 25-28. A mono-syllable word ("Juice", represented phonetically as /*dʒu:*s/) was decomposed into tonal and non-tonal components as described above. The original, highpass filtered, tonal, and non-tonal components decomposed by time-varying bandpass filters are shown in Figure 25. The energy in the highpass filtered speech is 9% of the energy in the original speech. The energy in the tonal component is 78% of the energy in the highpass filtered speech (7% of the original speech energy). The tonal component is dominated by the consonant hub (*/dz/*) at approximately 0.01 to 0.07 seconds, and it also includes some fricative sound (/s/) at around 0.37 seconds. The remaining 22% of the energy of the highpass filtered speech is in the non-tonal component (2% of the original speech energy) and includes energy associated with the onset and offset of the consonant hub at around 0.01 seconds and 0.07 seconds and the beginning and ending of fricative sound at around 0.35 and 0.38 seconds.

Spectrograms of these signals were calculated using the procedures presented in the previous chapter to present time-varying characteristics of decomposed components. As shown in Figure 26, the spectrograms clearly show that most of the sustained consonant hub in "Juice" is included in the tonal component and most of the transition activities, onset and offset of the consonant hub, and beginning and ending of fricative sound are in the non-tonal component.

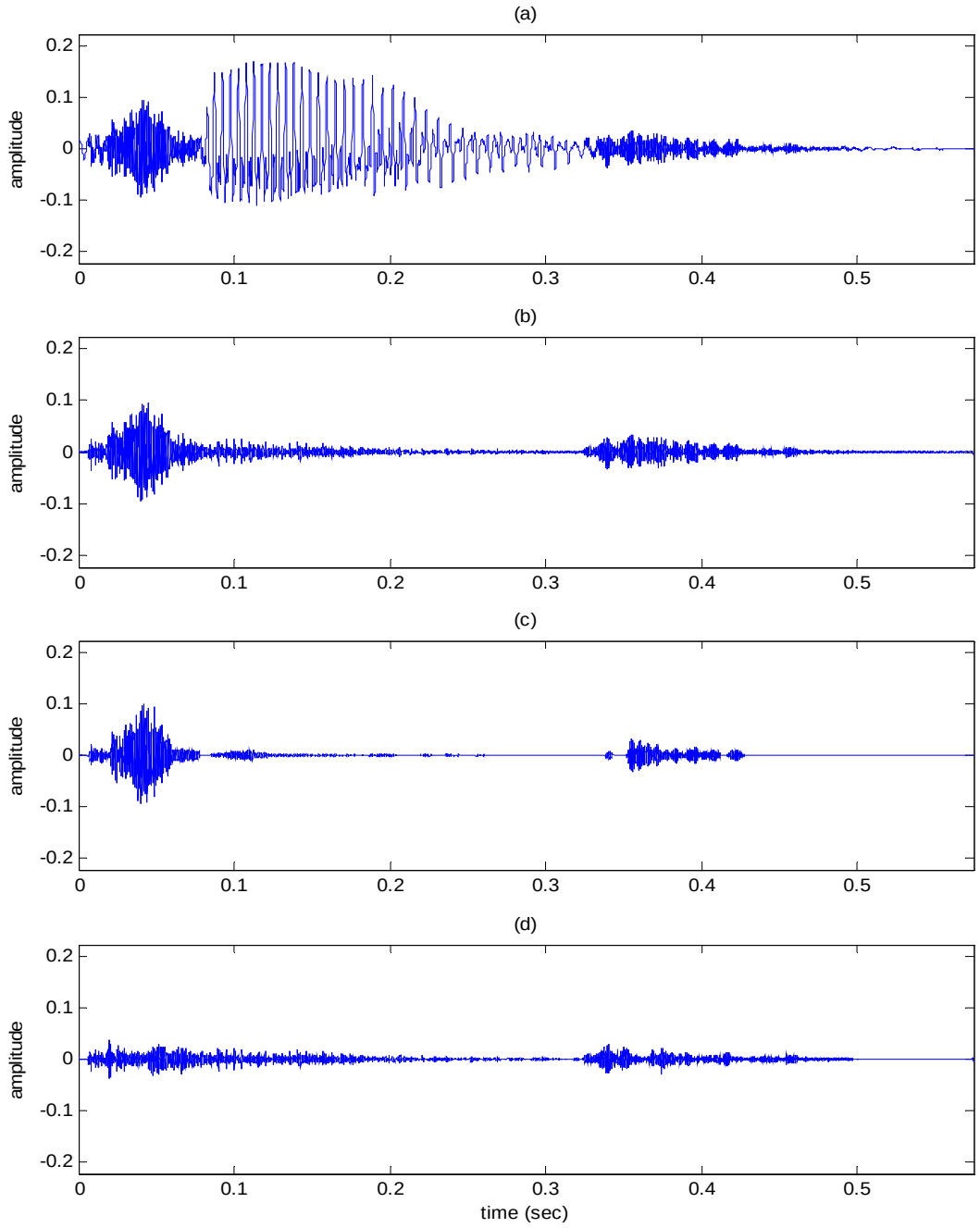

Figure 25: Waveforms of decomposed real speech signal “Juice” spoken by a female speaker: (a) original, (b) highpass filtered, (c) tonal, and (d) non-tonal components

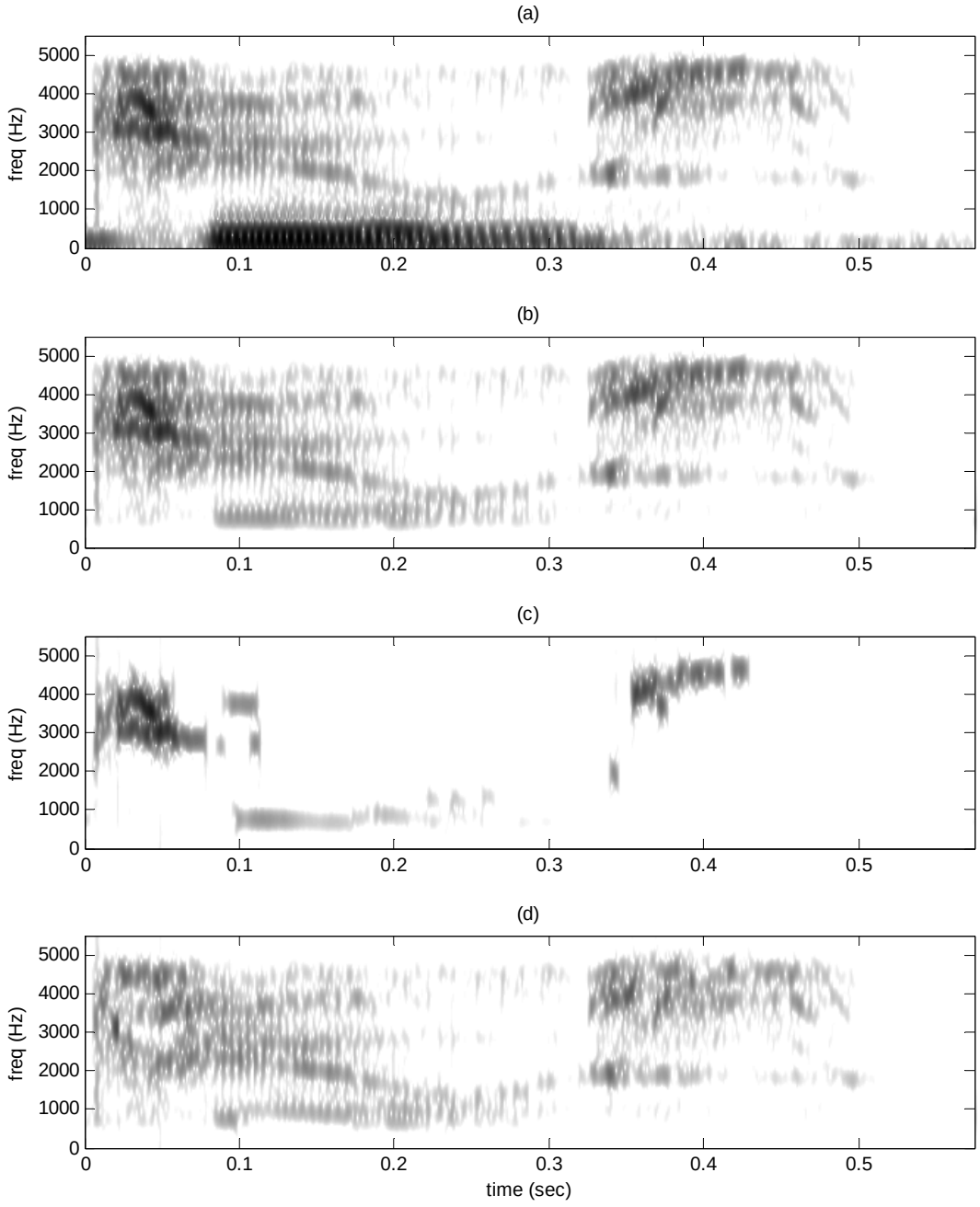

Figure 26: Spectrograms of decomposed real speech signal “Juice” spoken by a female speaker: (a) highpass filtered, (b) tonal, (c) non-tonal components

The sound of the tonal component was very garbled and not identifiable as the word "Juice". On the contrary, the non-tonal component was perceptually almost identical to the highpass filtered speech, despite having much less energy.

SNRs and time-varying bandwidths of each bandpass filter are shown in Figure 27, (a1-a3) and (b1-b3). The bandwidths open to pass the vowel sounds, /*u:*/ at around 0.01-0.1 seconds and closed to stop other activities. These characteristics are manifest in Figure 27 (c), which shows the upper and lower edges of bandwidths of each time-varying bandpass filter. The solid, dashed, and dotted lines are associated with the first, second, and third time-varying bandpass filters, respectively. The bandwidths are only opened enough to pass vowel sounds and closed to block other sounds from the tonal component. Figure 28 shows the output of each time-varying bandpass filter.

Another example of decomposition for a real speech signal spoken by a female speaker is illustrated in Figures 29-30. A mono-syllable word ("Pike", represented phonetically as /*paIk*/) was decomposed into tonal and non-tonal components. The original, highpass filtered, tonal, and non-tonal components decomposed by time-varying bandpass filters are shown in Figure 29. The energy in the highpass filtered speech is 16% of the energy in the original speech and the energy in the tonal component is 87% of the energy in the highpass filtered speech (14% of the original speech energy). The tonal component is dominated by the vowel /aI/, from approximately 0.07 to 0.17sec. The remaining 13% of the highpass filtered energy is in the non-tonal component (2% of the original speech energy) which includes energy associated with the noise burst accompanying the articulatory release of /p/ from approximately 0.01 to 0.07 sec., and the articulatory release of /k/ at around 0.38 sec. The sound of the tonal component was very garbled

and difficult to identify as the word “pike”. On the contrary, the non-tonal component was perceptually similar to the highpass filtered speech, despite having much less energy.

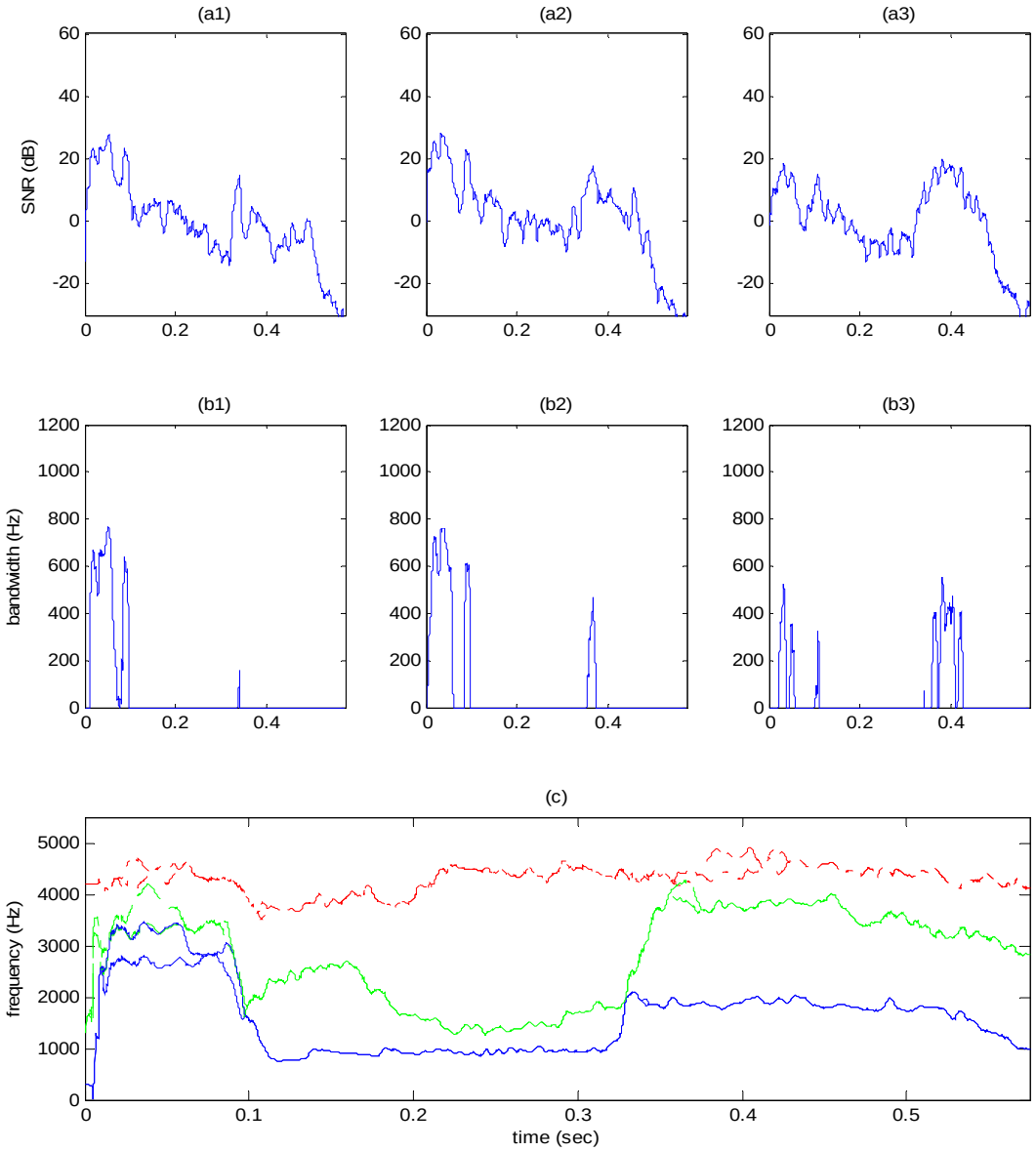

Figure 27: SNRs and time-varying bandwidths of each time-varying bandpass filter for a real speech signal “Juice”: (a) SNRs, (b) time-varying bandwidths, and (c) upper and lower edges of time-varying bandwidths. The solid, dashed, and dotted lines are associated with the 1st, 2nd, and 3rd time-varying bandpass filters, respectively.

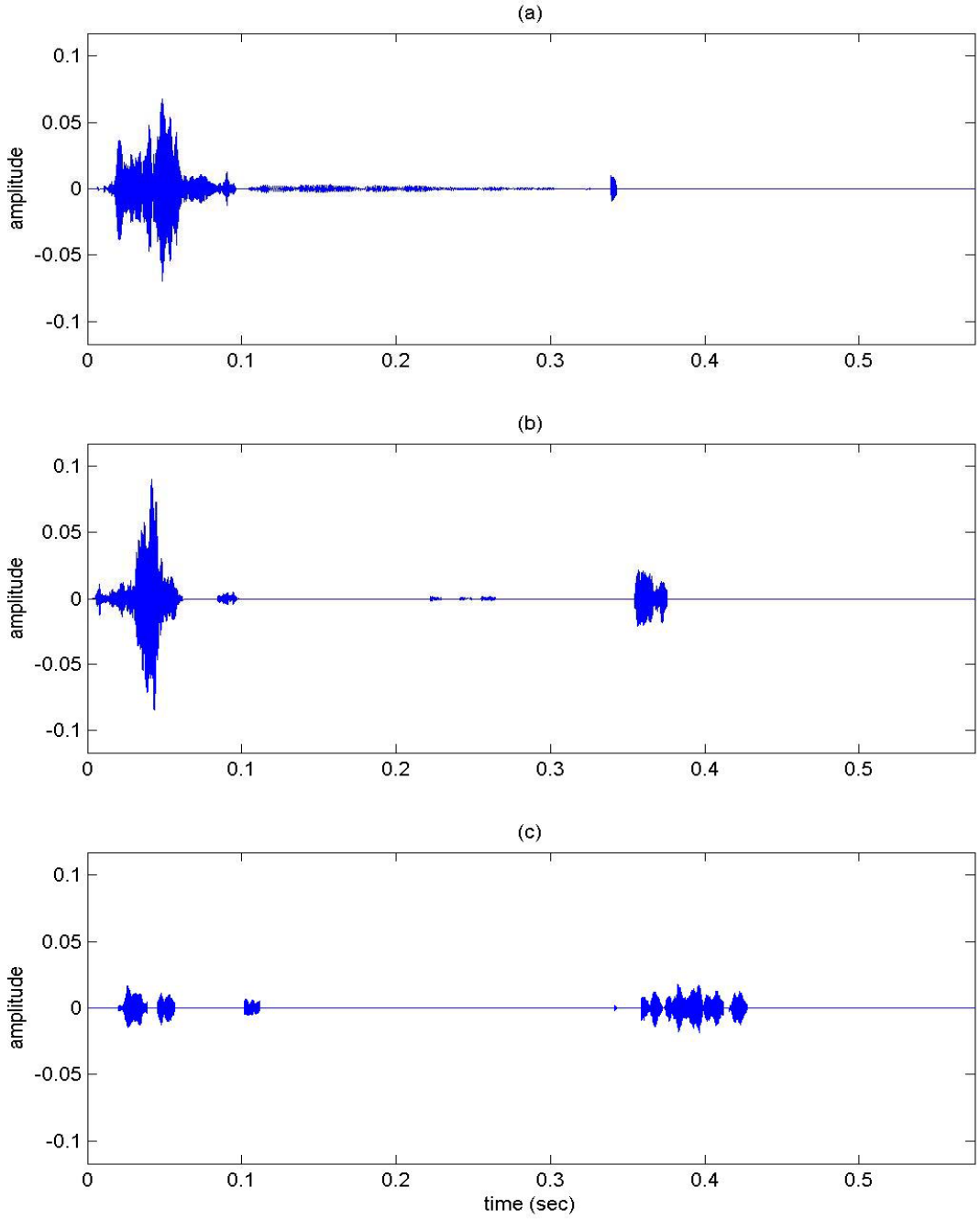

Figure 28: Individual output of each time-varying bandpass filter for a real speech signal "Juice" spoken by a female speaker: (a) 1st bandpass filter, (b) 2nd bandpass filter, and (c) 3rd bandpass filter

The spectrograms of highpass filtered, tonal, and non-tonal components were calculated to describe time-varying characteristics of decomposed components. As shown in Figure 30, most of the sustained vowel energy is included in the tonal component, and the non-tonal component emphasizes energy at the beginning and end of the tonal component. In particular, the non-tonal component includes spectral characteristics of both the /p/ and /k/ releases, as well as formant transitions from the /p/ release into the vowel /aI/. The location of the spectral energy in these transients contributes to the perception of place of articulation for both the consonants and the vowel.

Results obtained from the forty three mono-syllable words (twenty-three by a female speaker and twenty by a male speaker) and twelve two-syllable words (spoken by a female and by a male speaker) were similar to the results shown in this section. Figures 31 and 32 show the energy in the tonal and non-tonal components as a percent of the energy in the highpass filtered speech signal and the relative intelligibility of the tonal and non-tonal components, as evaluated subjectively by the author. The numbers 1 and 2 represent the female and male speaker respectively.

For the mono-syllable words (Figure 31), the relative energy in the tonal component ranged from 50% to 94% for the male speaker and from 31% to 89% for the female speaker. Overall, approximately 71% of the energy was in the tonal component, although that component was essentially unintelligible (1 by intelligibility assessment). The relative energy in the non-tonal component for the mono-syllable words ranged from 6% to 50% for the male speaker and from 11% to 69% for the female speaker. Overall, approximately 29% of the energy was in the non- tonal component, which was about as intelligible as but slightly less loud than the highpass filtered speech (4 or 5 by intelligibility assessment).

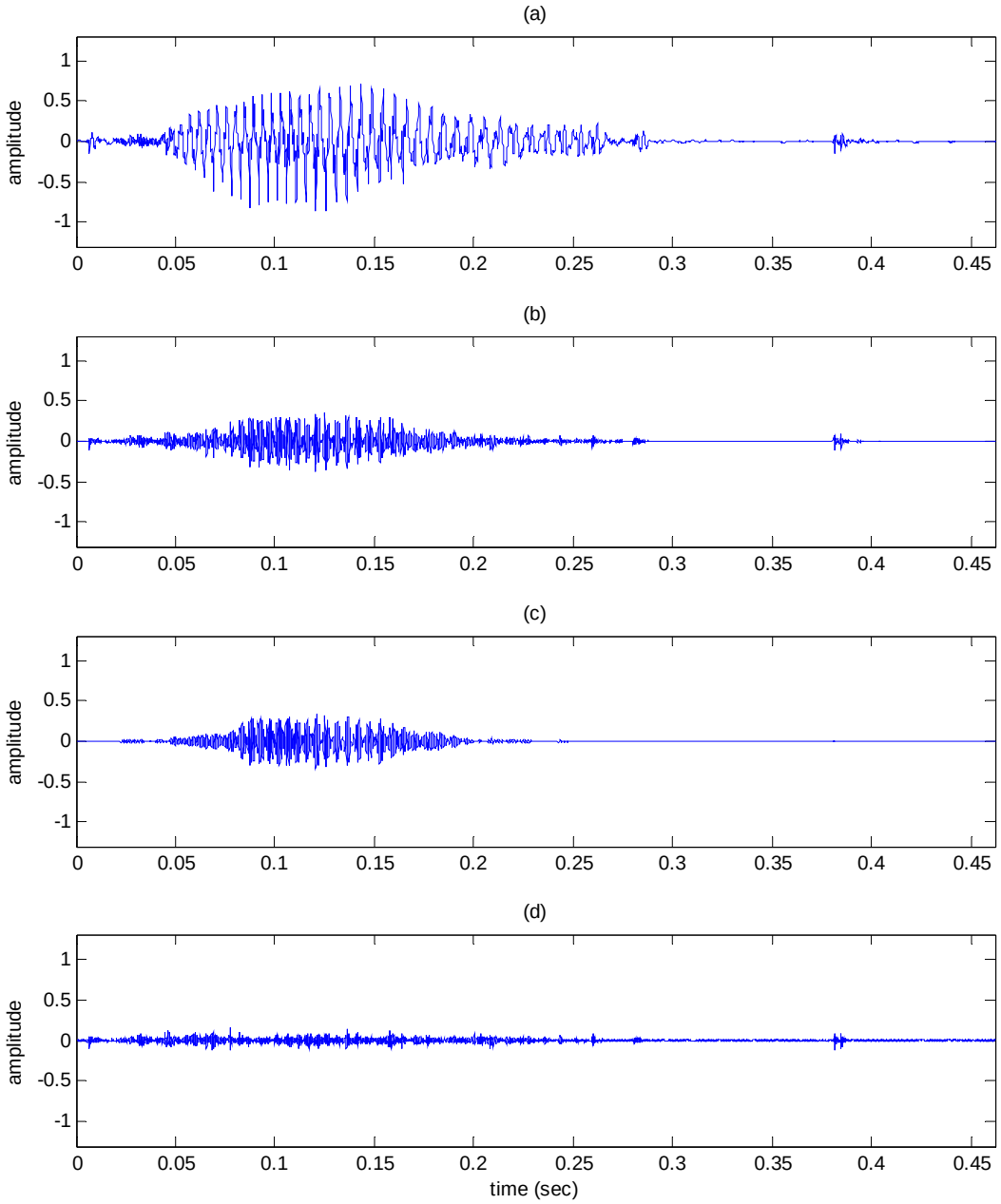

Figure 29: Waveforms of decomposed real speech signal “Pike” spoken by a female speaker: (a) original, (b) highpass filtered, (c) tonal, and (d) non-tonal components

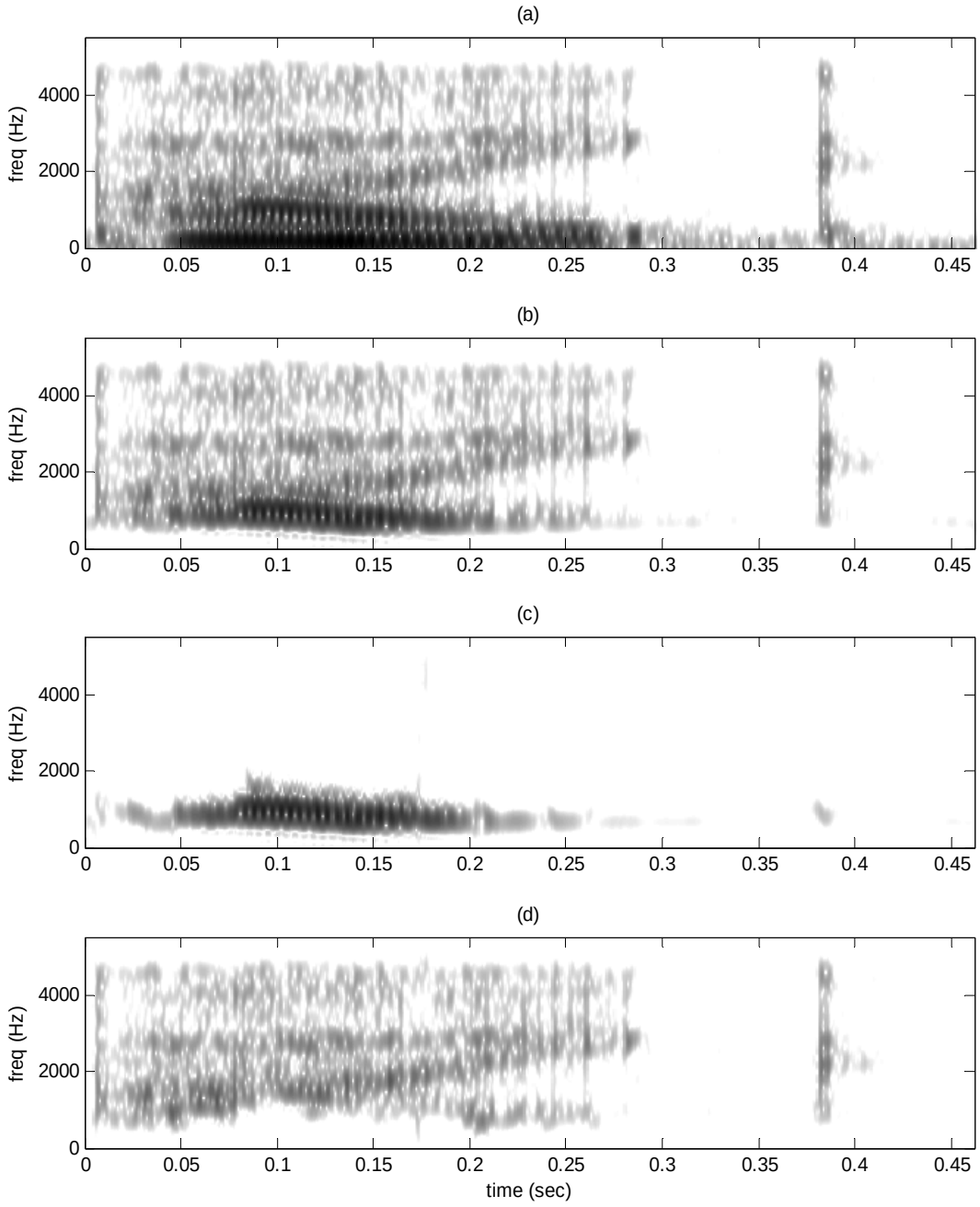

Figure 30: Spectrograms of decomposed real speech signal “Pike” spoken by a female speaker : (a) highpass filtered, (b) tonal, (c) non-tonal components

Similar results were found in the two-syllable words. As shown in Figure 32, the relative energy in the tonal component ranged from 76% to 91% for the male speaker and from 41% to 83% for the female speaker. Overall, approximately 74% of the energy was in the tonal component but the intelligibility of that component was very low (1 by intelligibility assessment). The relative energy in the non-tonal component for the two- syllable words ranged from 9% to 24% for the male speaker and from 17% to 59% for the female speaker. Overall, approximately 26% of the energy was in the non-tonal component but that component retained almost all of the speech signal's intelligibility (4 or 5 by intelligibility assessment).

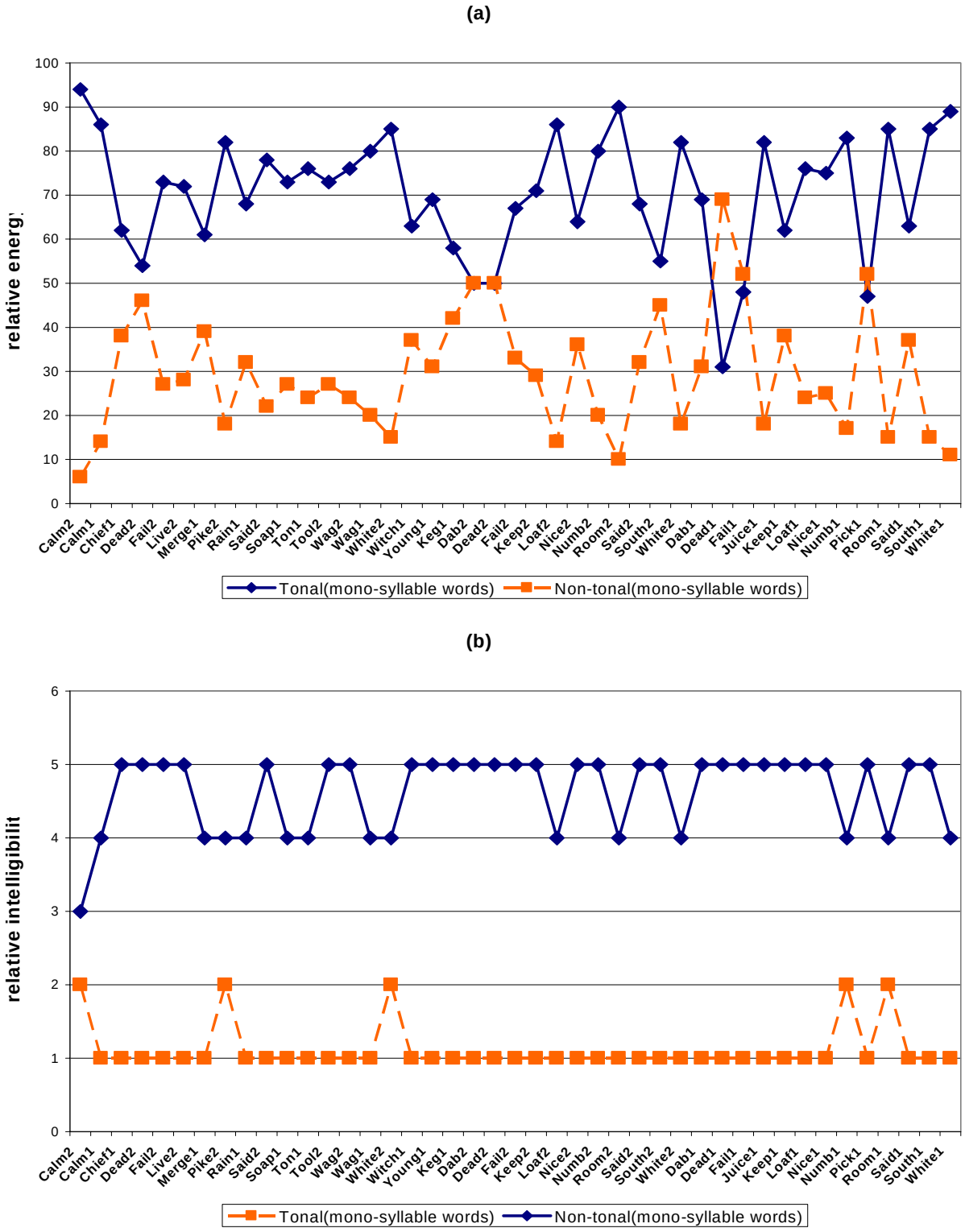

Figure 31: (a) relative energies in the tonal and non-tonal components and (b) relative intelligibility of the tonal and non-tonal components for mono-syllable words

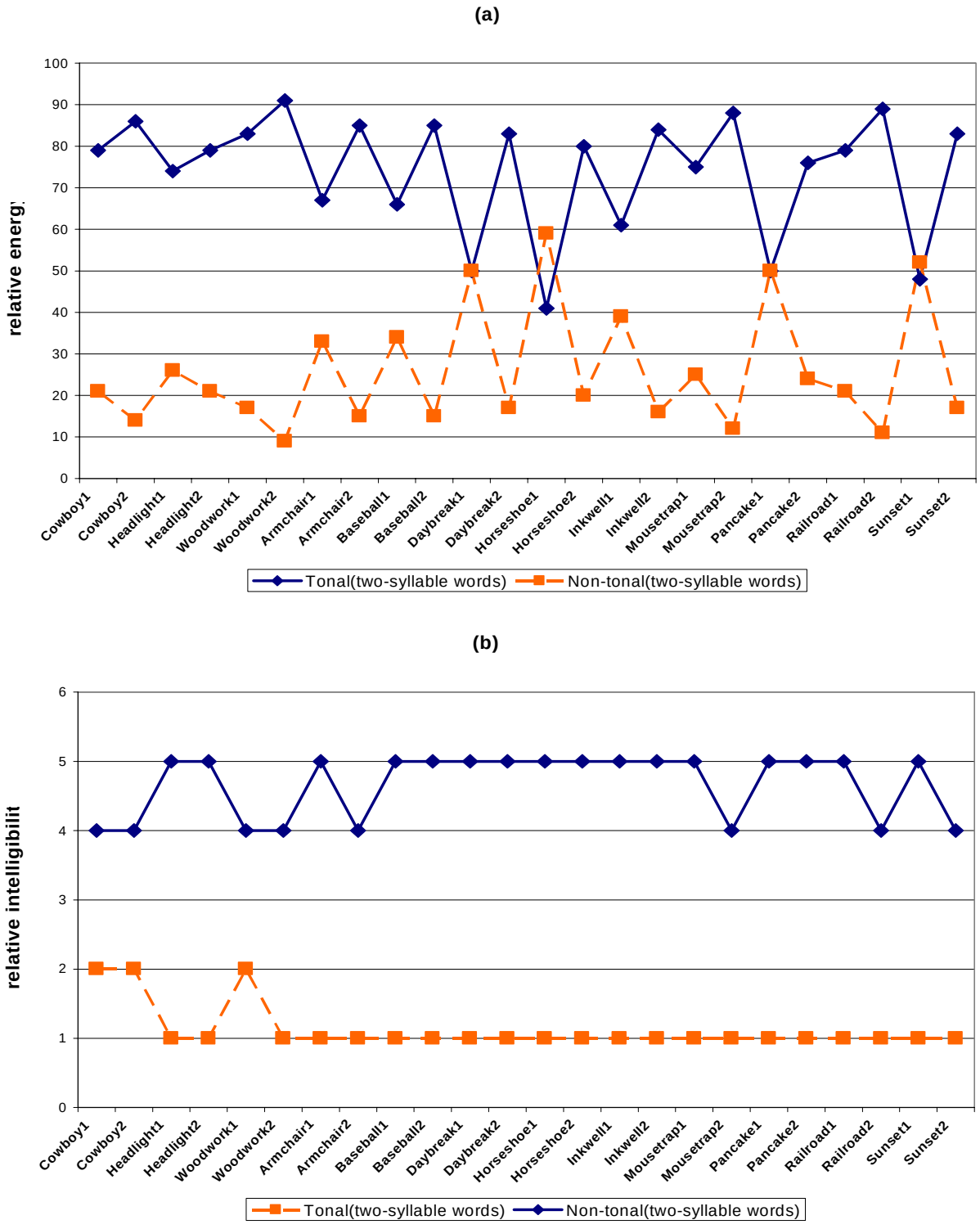

Figure 32: (a) relative energies in the tonal and non-tonal components and (b) relative intelligibility of the tonal and non-tonal components for two-syllable words

4.3 ALGORITHM PARAMETER SELECTIONS

As mentioned in the previous chapter, each time-varying bandpass filter requires two parameters: maximum bandwidth and bandwidth threshold (speech-to-noise ratio) at which the filter is activated. The maximum bandwidth should be large enough to capture most of the energy in the spectral band being tracked but small enough to be restricted to a single band. The bandwidth threshold is based on the ratio of speech to noise energy in a spectral band. It should be low enough to assure that the filter is active during a sustained sound and high enough to not be active during speech transitions or noise. The selections of maximum bandwidth and bandwidth threshold are based on the criteria of removing as much of the quasi-steady-state energy from the original speech as possible, while maintaining reasonable intelligibility in the non-tonal component. We believe this criterion reflects the optimal separation of tonal and non-tonal components.

The effects of variations of these parameters on energy and intelligibility of the tonal and non-tonal components were investigated by applying several maximum bandwidths and bandwidth thresholds to five two-syllable and three mono-syllable words. After preliminary evaluations, three different bandwidth thresholds (12 dB, 15 dB, and 18 dB) with 900 Hz of maximum bandwidth and three different maximum bandwidths (700 Hz, 900 Hz, and 1100 Hz) with 15 dB of bandwidth threshold were investigated.

The energy of the tonal and non-tonal components as a percent of the energy in the highpass filtered speech was computed. Results are summarized in Figures 33 and 34. The energy of the tonal component decreased by an average 10.5% as bandwidth threshold

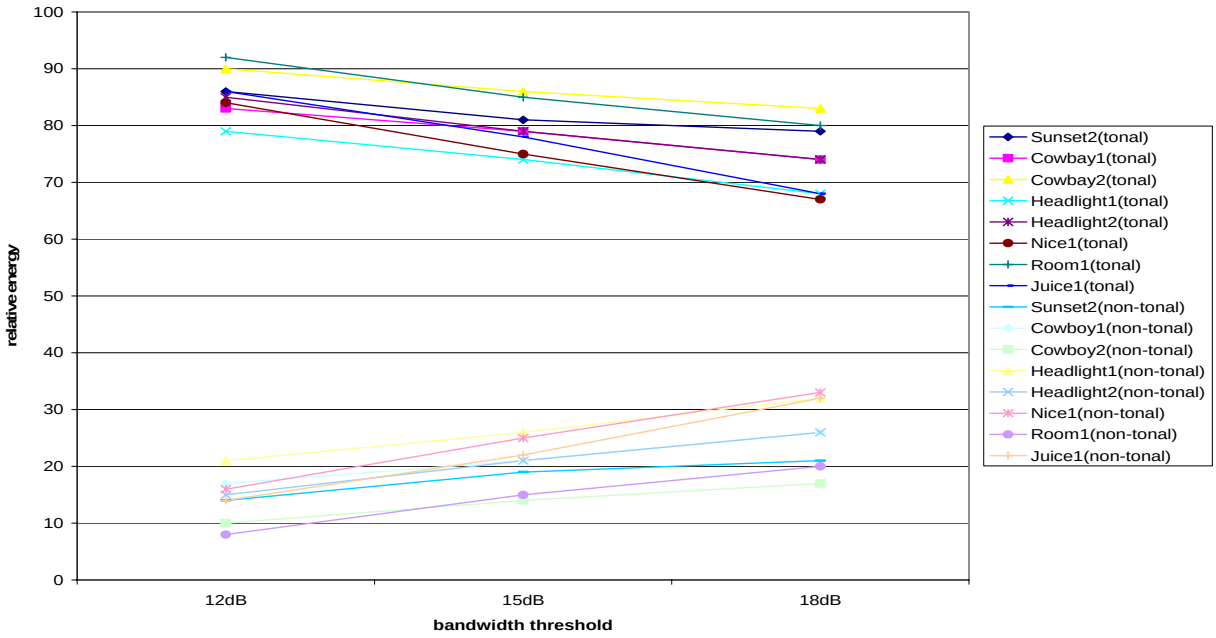

Figure 33: Relative energies in the tonal and non-tonal components with different bandwidth thresholds

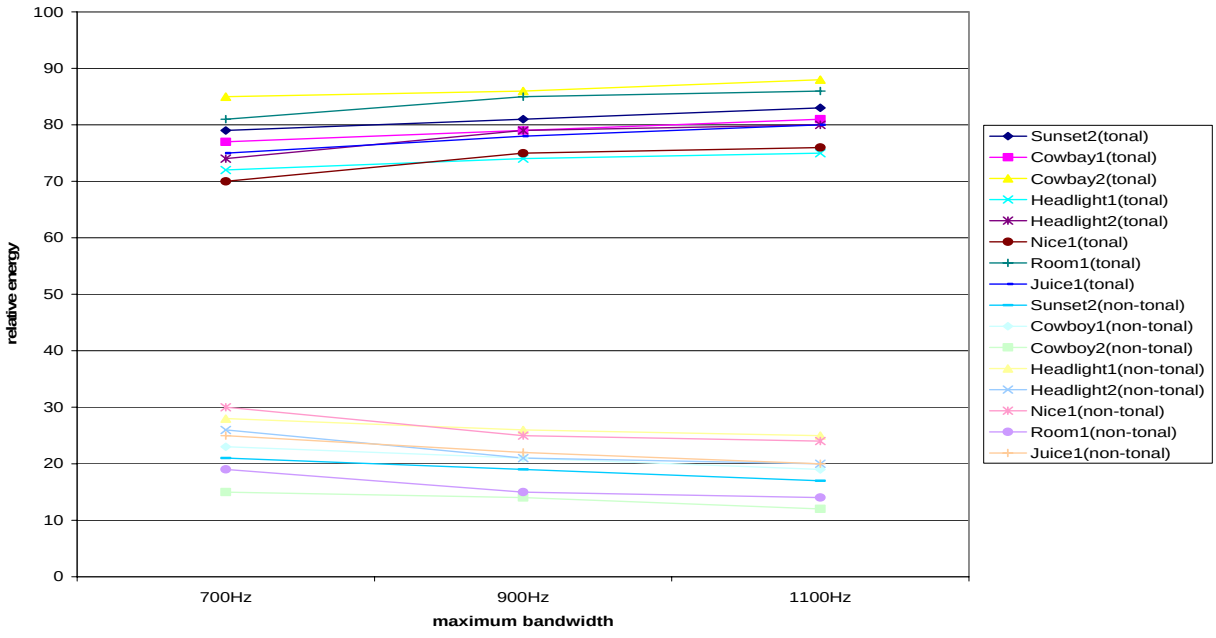

Figure 34: Relative energies in the tonal and non-tonal components with different maximum bandwidths

increased from 12 dB to 18 dB, and it increased by an average of 4.5% as maximum bandwidth increased from 700 Hz to 1100 Hz. The energy changes in the tonal and non-tonal components for different parameter values were relatively small, as shown in the Figure 33 and 34.

The intelligibility of the tonal and non-tonal components was assessed as described above and is shown in Tables 4 and 5. The average intelligibilities for the tonal components with 12 dB, 15 dB, and 18 dB bandwidth thresholds were 1.5, 1.3, and 1.1, respectively and for the non-tonal components with 12 dB, 15 dB, and 18 dB bandwidth thresholds were 3.9, 4.9, and 5.0, respectively. The tonal components were essentially unintelligible in all the sets of bandwidth thresholds. The non-tonal components are highly intelligible in the 15 dB and 18 dB bandwidth thresholds but began to lose intelligibility in the 12 dB bandwidth threshold. The non-tonal components with 12 dB bandwidth threshold were slightly less loud than the non-tonal components with 15 dB and 18 dB bandwidth thresholds.

The average intelligibility for the tonal components with 700 Hz, 900 Hz, and 1100 Hz maximum bandwidths were 1.1, 1.3, and 1.3, respectively and the non-tonal components with 700 Hz, 900 Hz, and 1100 Hz maximum bandwidths were 4.9, 4.9, and 5.0, respectively. The decomposition performance was not critically dependent on the maximum bandwidths. The tonal components were essentially unintelligible in all the sets of maximum bandwidths. The non-tonal components were highly intelligible for all maximum bandwidths tested, but the non-tonal components with maximum bandwidth of 1100 Hz were a little less loud than the non-tonal components with 700 Hz and 900 Hz of maximum bandwidths.

Based on these results, the 15 dB for bandwidth threshold and 900 Hz for maximum bandwidth were selected for the speech decompositions because they provided minimum energy in the non-tonal components with least loss of intelligibility.

Table 4: Relative intelligibility in the tonal and non-tonal components with different bandwidth thresholds

Words \ Bandwidth Threshold	12 dB		15 dB		18 dB	
	Tonal	Non-tonal	Tonal	Non-tonal	Tonal	Non-tonal
Sunset2	1	5	1	5	1	5
Cowboy1	2	4	1	5	1	5
Cowboy2	2	4	2	5	2	5
Headlight1	1	4	1	5	1	5
Headlight2	1	4	1	5	1	5
Nice1	2	4	1	5	1	5
Room1	2	3	2	4	1	5
Juice1	1	4	1	5	1	5

Table 5: Relative intelligibility in the tonal and non-tonal components with different maximum bandwidths

Words \ Maximum bandwidth	700 Hz		900 Hz		1100 Hz	
	Tonal	Non-tonal	Tonal	Non-tonal	Tonal	Non-tonal
Sunset2	1	5	1	5	1	5
Cowboy1	1	5	1	5	1	5
Cowboy2	1	5	2	5	2	5
Headlight1	1	5	1	5	1	5
Headlight2	1	5	1	5	1	5
Nice1	1	5	1	5	1	5
Room1	2	4	2	4	2	5
Juice1	1	5	1	5	1	5

5.0 PSYCHOACOUSTIC EVALUATIONS

The goal of this study was to investigate the roles of steady-state speech sounds and transitions between these sounds on the intelligibility of speech and the possibility of speech enhancement, based on the transitions, in background noise. The intelligibility of the different speech components and of enhanced speech is evaluated by psychoacoustic tests and the results are presented in this chapter. Intelligibility growth functions of speech were generated from the results of psychoacoustic intelligibility tests as described in section 2.3.1, and parameters extracted from the growth functions were analyzed statistically. A method of speech enhancement is described, and word identification scores and subjects' response times for original and enhanced speech in background noise were determined using the modified rhyme protocol to determine whether the enhanced speech provides improvement in word identification.

An experiment was also conducted to determine whether a fixed filter can produce enhanced speech as effectively as the time-varying filter. A filter function generating the long-term averaged spectrum of enhanced speech from the long-term averaged spectrum of original speech was calculated, and original speech was filtered by this filter function. Psychoacoustic evaluations with enhanced speech and the filtered original speech are also presented in this chapter.

5.1 TESTS ON SPEECH COMPONENTS

To evaluate the relative intelligibility of original, highpass filtered, tonal, and non-tonal speech components, psychometric functions to show growth of intelligibility for each component as signal amplitude increased were determined.

5.1.1 Methods

Three hundred consonant-vowel-consonant (CVC) words from the NU-6 word lists were decomposed to provide highpass filtered, tonal, and non-tonal components for each word [36]. Test words were presented in a quiet background. Five volunteer subjects with negative otologic histories and hearing sensitivity of 15 dB HL or better by conventional audiometry (250 – 8 kHz) were tested. Subjects sat in a sound-attenuated booth, and test words were delivered monaurally though headphones. Subjects were asked to repeat the words presented, and the number of errors in word identification was recorded by skilled examiners under supervision of a certified clinical audiologist. For each component, stimuli were presented at five intensity levels from 0% recognition until recognition reached 100% or did not increase.

Recognition results for each subject were fit to an error function, using the nonlinear least-squares fit routine 'lsqcurvefit' (MATLAB, The Mathworks, Inc.). The function minimum was set to zero, and estimates of the maximum (PB_{max} or maximum word recognition rate), midpoint (50% recognition point used to define threshold in classical signal detection studies but not related to the threshold measure in hearing sensitivity function), and slope (measured by the standard deviation parameter of the error function) were obtained. The mean squared difference between the fitted function and the original data divided by the total mean square of the data (R^2) was calculated to assess the adequacy of the fit, with $R^2 > 0.8$ being taken to indicate a

satisfactory fit. An example of growth function fit is illustrated in Figure 35, where the diamond symbols represent actual recognition scores and the solid line represents the error function fit to the data.

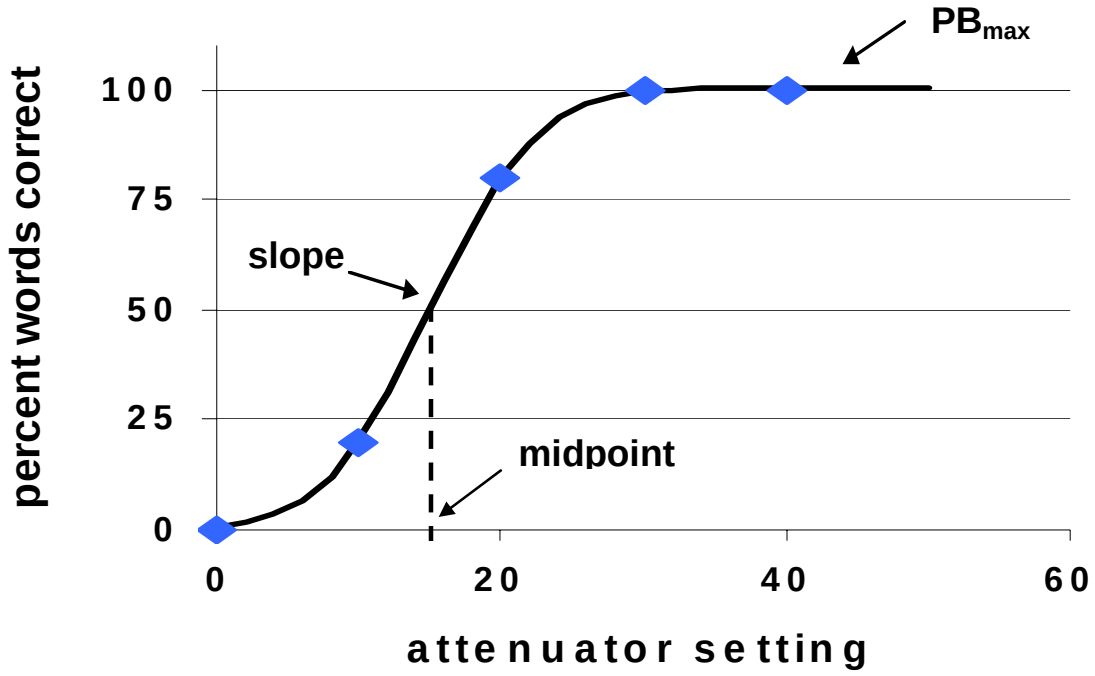

Figure 35: Example of growth function fit ($R^2 = 0.99$).

The growth function parameters obtained for the original, highpass filtered, tonal and non-tonal versions of the words were tested for significant differences across versions. Because of the potential for the data to be skewed (most of the data points were between 50% and 100%, with 100% as the maximum), a Friedman's test was used as a non-parametric analysis of variance, followed by Wilcoxon paired comparison tests of significant Friedman's results. P-values less than 0.05 were taken to indicate significant differences. Although this approach has

less power than parametric tests, sufficient data were not available to verify that the data were not significantly skewed.

5.1.2 Results

Table 6 shows the energy in the tonal and non-tonal components, averaged over the 300 CVC words, as a fraction of the energy in the original speech and highpass filtered speech. The non-tonal components averaged 2% of the original speech energy (18% of the highpass filtered speech energy), and the tonal component averaged 18% of the original speech energy (82% of the highpass filtered speech energy). The tonal component had loudness approximately equal to the highpass filtered speech, but the non-tonal component sounded less loud, as would be expected due to the lower energy.

Table 6: Mean of energy in the tonal and non-tonal components of mono-syllable words relative to energy in the highpass filtered speech and in the original speech. Standard deviation in parenthesis.

	Tonal component	Non-tonal component
% of HPF speech	82% (6.7)	18% (6.7)
% of original speech	12% (5.5)	2% (0.9)

Word recognition rates for each subject were successfully fit to error functions. Of the 20 sets of data (4 different word versions for 5 subjects), 18 were fit with $R^2 > 0.9$ and 2 with R^2 between 0.8 and 0.9. The upper graph in Figure 36 shows the growth of word recognition, based on error function parameters averaged across subjects for each version of the test words, as a

function of unadjusted speech level (the components were tested with amplitudes obtained directly from the algorithm). Highpass filtered speech, despite having much less energy than the original speech, was recognized at similar speech levels, while tonal and non-tonal components were recognized at similar, but higher levels. The lower graph shows sound level adjusted to compensate for the different component energies. That is, 0 dB represents the same energy level (original speech at 50% recognition) in each component. Highpass filtered speech had about the same maximum intelligibility as original speech, and the non-tonal component had only slightly lower maximum intelligibility. The tonal component had lower maximum intelligibility.

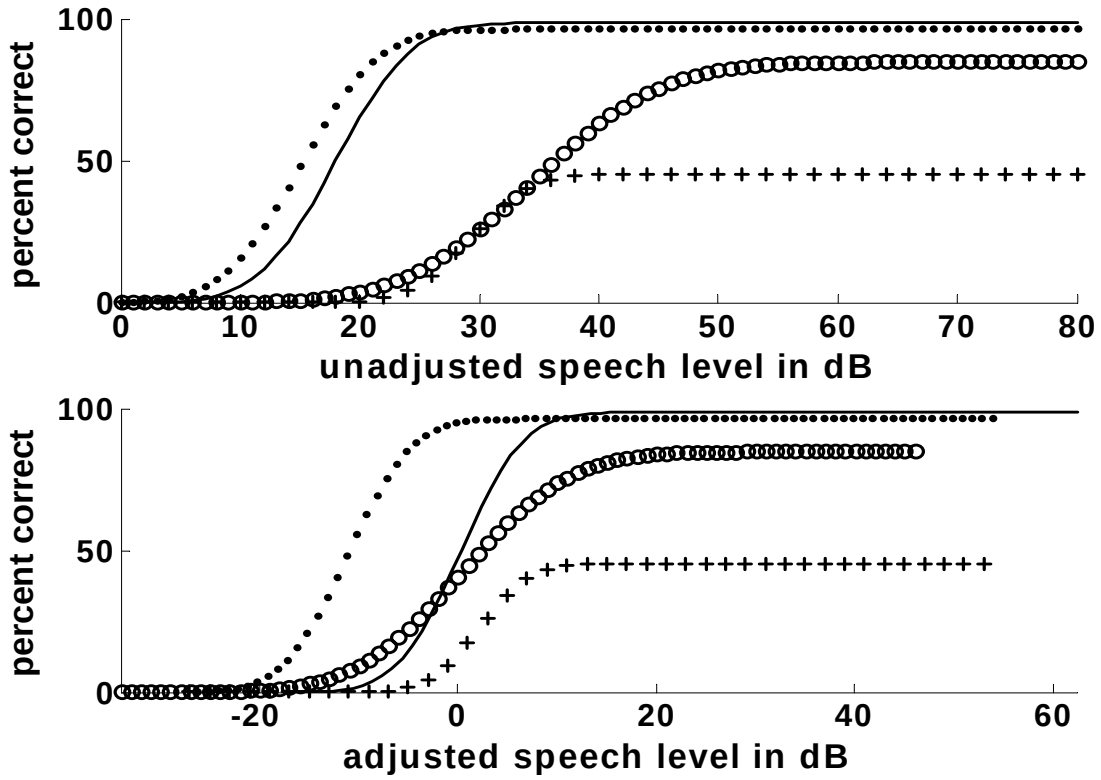

Figure 36: Growth of word recognition based on error function parameters: solid: original speech; dotted: highpass filtered speech; +-+: tonal component; o-o: non-tonal component.

Means and standard deviations of the parameters are summarized in Table 7. For the adjusted midpoint, sound levels were adjusted to compensate for the different component energies (the bottom graph in figure 36). The adjusted midpoint ($p = 0.027$) and PB_{max} ($p = 0.016$) were significantly different. Wilcoxon paired comparison results for these two parameters are summarized in Table 8. For $PB_{max,}$ the tonal component was significantly different from the other three versions, and for adjusted midpoint, the highpass filtered version was significantly different from the other three versions. The tonal component, despite having most of the energy of highpass filtered speech, had significantly lower PB_{max} than the other components. The adjusted growth function midpoint of the highpass filtered speech was significantly smaller than for original speech, suggesting that this component was picked out of noise at lower stimulus levels. The standard deviations of growth functions were the same for all versions, showing that the slopes of the growth functions showed no significant differences.

Table 7: Growth function parameters. Standard deviation in parenthesis.

	PB_{max}	Midpoint	Adjusted midpoint	Slope
Original speech	98.7 (3.0)	17.9 (2.7)	0.3	7.1 (3.2)
Highpass filtered speech	96.5 (2.1)	15.0 (3.8)	-11.2*	7.2 (2.5)
Tonal component	45.1 (19.3)*	29.2 (11.3)	2.2	5.6 (8.5)
Non-tonal component	84.9 (14.4)	34.4 (4.6)	0.5	12.1 (6.3)

* $p < 0.05$ for pair-wise comparisons with other components.

Table 8: Results of the Wilcoxon paired comparison tests

	PB_{max}	Adjusted midpoint
Original – Highpass filtered	0.144	0.043*
Original – Tonal	0.043*	0.893
Original – Non-tonal	0.109	0.893
Highpass filtered – Tonal	0.043*	0.043*
Highpass filtered – Non-tonal	0.225	0.043*
Tonal – Non-tonal	0.043*	0.715

* $p < 0.05$ for pair-wise comparisons with other components.

These results showed that highpass filtered speech was about as intelligible as original speech and the non-tonal component was only slightly less intelligible than the original and highpass filtered speech. The tonal component, however, was much less intelligible. These results support our hypothesis that transitional information in speech may be important to speech perception.

5.2 TESTS ON ENHANCED SPEECH

The motivation for using the non-tonal component for speech enhancement is that the non-tonal component, which emphasizes transitional information in speech, may be critical to the speech perception. As shown in section 5.1.2, the non-tonal component, despite having much less energy than the original and highpass filtered speech, had only slightly lower maximum intelligibility. We suggest that the transition information in the non-tonal component may be critical to speech perception but that it is particularly susceptible to noise. Selectively amplifying this component may improve the recognition of speech in noise.

To generate enhanced speech, speech sounds were decomposed, and the non-tonal component was amplified and then recombined with the original speech. The energy of enhanced speech was adjusted to be equal to the energy of the original speech, and the intelligibility of these two speech versions was evaluated using the modified rhyme protocol described in section 2.3.2 [31], [32].

5.2.1 Methods

Three hundred mono-syllable words (50 sets of rhyming words) were recorded by a male speaker, and these words were decomposed into tonal and non-tonal components, as described previously. These 300 words are presented in Appendix D. Bandwidth thresholds and maximum bandwidths for time-varying filters were 15 dB and 900 Hz respectively, based on the results in section 4.3.

To generate enhanced speech, the non-tonal component was multiplied by amplification factor *k* and then recombined with the base speech as

$$X_{enh}(t) = m * (X_{base}(t) + k * X_{nont}(t)) \qquad (5.1)$$

where $X_{enh}(t)$, $X_{base}(t)$, and $X_{nont}(t)$ represent the enhanced speech, base speech, and non-tonal component respectively and *m* represents energy adjustment constant (the energy of enhanced speech was adjusted to be equal to the energy of the base speech). Enhancements by three different amplification factors (4, 8, and 12) and two different base speech types (recombining with the original and highpass filtered speech) were preliminarily evaluated by the author. Amplification factors greater than 12 were also tested, but the enhancement effect was less than with lower factors. Based on these evaluations, an amplification factor 12 and original speech base were selected for speech enhancement.

Each stimulus set consisted of 6 mono-syllable words. An additional seventy two mono-syllable words (12 sets of rhyming words) were also processed for training purposes. These seventy two words were recorded by the same male speaker and did not include any of the first set of words. Test administration was computerized. Test words were presented with six different SNR levels (-25 dB, -20 dB, -15 dB, -10 dB, -5 dB, and 0 dB) of speech-weighted background noise. The sound pressure spectrum level of the noise was constant from 100 Hz to 1000 Hz and decreased at a rate of 12 dB/octave from 1000 Hz to 5513 Hz [37]. Because speech-weighted noise approximates the long-term sound pressure spectrum level of speech, it effectively interferes with the recognitions of speech sounds (increases auditory thresholds). The lowpass spectral shape of speech-weighted noise creates evenly distributed noise energy across critical bands in the auditory system because the bandwidths of critical bands increase with increasing frequency.

Each word was normalized to unit root-mean-square amplitude. The background noise was presented for 1.83 sec. and windowed by a Tukey window for a smooth onset and offset. The window rise and fall times were 0.25 sec. The amplitude of the background noise was adjusted to one of the six SNR levels for the word, and the word was presented with this background noise. Each SNR was defined by the amplitude ratio of the word and noise over the same time-interval. The interval between stimuli was 0.25 sec. The structure of the stimuli is illustrated in Figure 37. The order of presentations and noise levels were randomized. Subject responses were recorded by the computer, and the test results, including number of correct responses and response times, were saved.

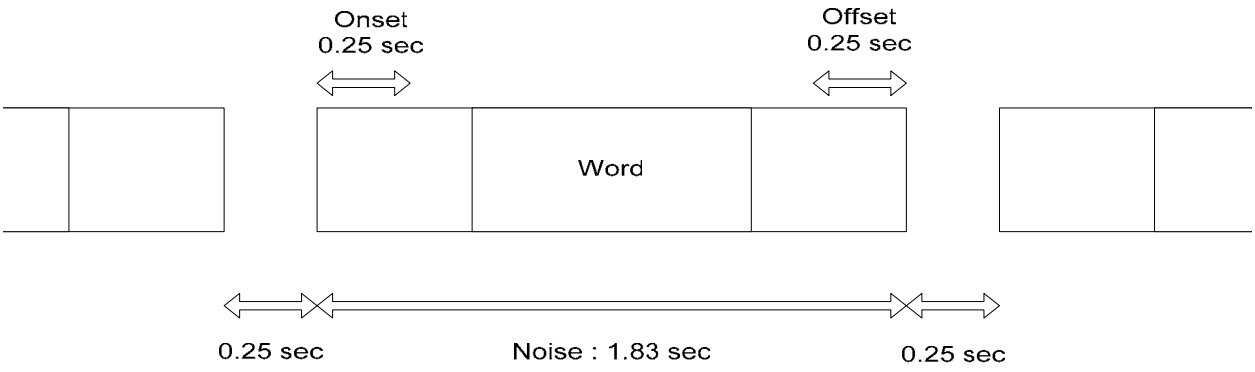

Figure 37: Structure of the stimuli

Eleven volunteer subjects with negative otologic histories and hearing sensitivity of 15 dB HL or better by conventional audiometry (250 – 8 kHz) were tested. Subjects sat in a sound-attenuated booth, and test words with background noise were delivered monaurally though headphones. At the beginning of each trial, a target word appeared on the computer monitor and remained until all six alternative words were presented. The first word among six alternative words was presented one second after the target word appeared on the computer monitor. The subjects were asked to push a mouse button as soon as they heard the target word. The subjects did not have a second chance to hear the test words. The response time was measured from the end of word presentation to the moment when the subject pushed the button.

Each subject had a training trial consisting of 12 sets of rhyming words (total 72 words). Among these 12 sets, 6 sets of rhyming words were presented as original speech and the remaining 6 sets were presented as enhanced speech. Among these two 6 sets of original and

enhanced speech, 3 sets were presented in quiet and the remaining 3 sets were presented with speech-weighted noise at different SNR levels.

In the main trial, 50 sets of rhyming words were repeated 6 times for each subject (total 300 sets of rhyming words per subject). One hundred fifty of the 300 sets were presented as original speech and the remaining 150 sets were presented as enhanced speech. The target words were randomly selected from the 300 mono-syllable words, and the selected target word was excluded in the future selections (the same target word did not appear more than once.). The sets were presented at 6 different SNR levels of speech-weighted noise (25 sets for each noise level and speech type). The order of presentations and noise levels were randomized. Subject responses were recorded by the computer, and the test results, including number of correct responses and response times, were saved. The test procedures were monitored by skilled examiners under supervision of a certified clinical audiologist.

Means, standard deviations, and 95% confidence intervals of recognition scores and response times for each subject and each noise condition were computed using MATLAB (The Mathworks, Inc.). The data distributions of recognition scores and response times for original and enhanced speech from 11 subjects were examined and confirmed to be normally distributed. Because the data appeared to be normally distributed, parametric statistical procedures (paired t-tests) were used to test for significant differences. A preliminary test with two additional subjects showed that, at SNR = -20 dB, recognition rates were reasonably high but that enhanced speech was substantially more recognizable. We hypothesized that enhanced speech would show a significantly better recognition and tested both recognition rates and response times at this SNR using a paired t-test procedure. Ninety-five percent confidence intervals were used to describe the results at all SNR levels.

5.2.2 Results

The average energy in the tonal and non-tonal components, averaged over the 50 sets of rhyming words for the main trial, as a fraction of the energy in the original speech and highpass filtered speech, is shown in Table 9. The non-tonal components averaged 5.8% of the original speech energy (36.6% of the highpass filtered speech energy), and the tonal component averaged 11.2% of the original speech energy (63.4% of the highpass filtered speech energy).

Table 9: Mean of energy in the tonal and non-tonal components of 50 sets of rhyming words used in main trials relative to energy in the highpass filtered speech and in the original speech. Standard deviation in parenthesis.

	Tonal component	Non-tonal component
% of HPF speech	63.4% (18.8)	36.6% (18.8)
% of original speech	11.2% (4.9)	5.8% (5.0)

The recognition rates and response times were averaged across subjects for summary graphs. Means and 95% confidence intervals of the word recognition scores for original and enhanced speech are shown in Figure 38, where the dashed line represents intelligibility of enhanced speech and the solid line represents intelligibility of original speech. For both speech versions, percent correct scores increased as the SNRs increased from -25 to 0 dB. The enhanced speech showed higher recognition scores at most SNR levels.

At lower SNRs (-25, -20, and -15 dB), the 95% confidence intervals for recognition of original and enhanced speech did not overlap, showing that the subjects could identify the

enhanced speech better than the original speech under more severe noise conditions. For example, at -25 dB SNR, the mean recognition scores from original and enhanced versions were 22% and 54% respectively. The differences were not significant at higher SNRs (-10, -5, and 0 dB).

The null hypothesis that there is no difference in speech recognition rates between original and enhanced speech at SNR = -20 dB was tested using a paired t-test. The mean difference across subjects of 25.5%, with standard deviation 7.4%, is significantly different from zero at $p<0.05$, and the null hypothesis was rejected. Paired test results are summarized, using 95% confidence interval of the differences averaged across subjects, for all SNRs in Table 10.

Means and 95% confidence intervals of the response times for original and enhanced speech are shown in Figure 39, where the dashed line represents response times for enhanced speech and the solid line represents response times for original speech. Paired test results are summarized, using 95% confidence interval of the differences averaged across subjects, for all SNRs in Table 11. Note that Figure 39 represents averages across subjects for each condition separately. The paired t-test results in Tables 10 and 11, however, compare differences between conditions within a given subject and are then averaged across subjects.

At -20 and -15 dB SNRs, the differences between original and enhanced speech versions were significant, showing that the subjects respond the enhanced speech faster than the original speech under severe noise conditions. However, the difference between original and enhanced speech for -25 dB SNR was not significant. At 0 dB SNR, the differences between original and enhanced speech versions were significant, showing that the subjects respond to the original speech faster than the enhanced speech at this SNR level. The differences between original and enhanced speech for -10 and -5 dB SNR were not significant.

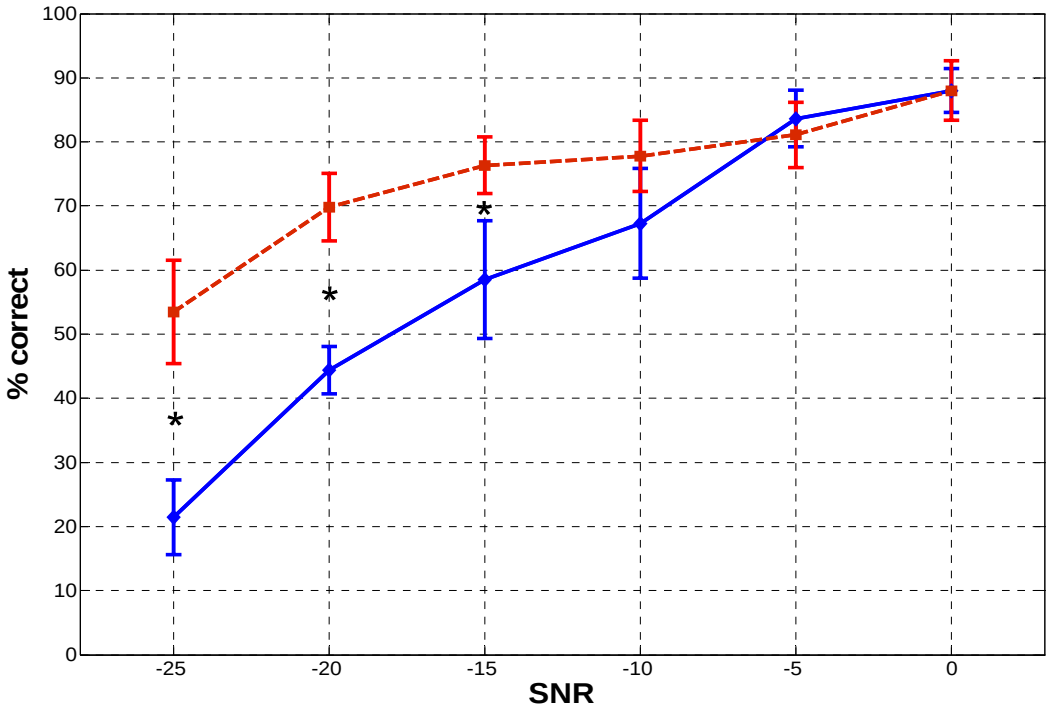

Figure 38: Means and 95% confidence intervals of word recognitions (%) for original (solid) and enhanced (dashed) speech. (* : paired differences not equal zero)

Table 10: Differences (enhanced speech – original speech) of means, standard deviations (SDs), and 95% confidence intervals (CIs) of word recognition scores.

SNR	Mean difference	SD difference	95% CI difference
-25 dB	32.0	12.1	23.85 ~ 40.15
-20 dB	25.5	7.4	20.46 ~ 30.45
-15 dB	17.8	12.2	9.64 ~ 26.00
-10 dB	10.5	18.6	-1.96 ~ 23.05
-5 dB	-2.5	6.3	-6.76 ~ 1.66
0 dB	0	9.3	-6.24 ~ 6.24

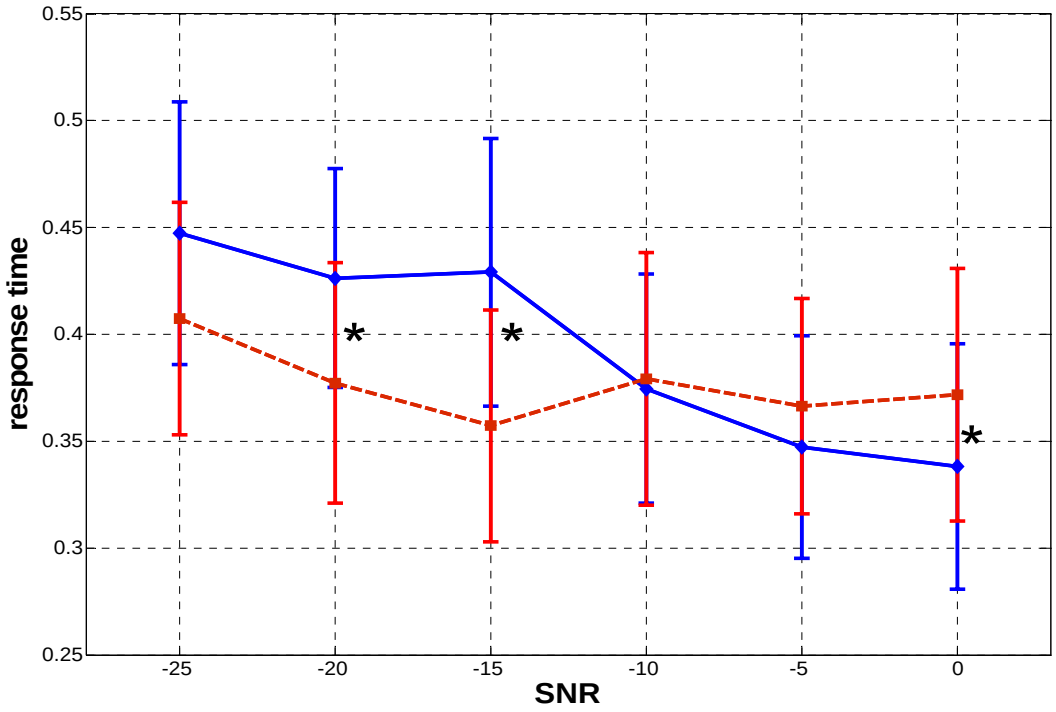

Figure 39: Means of response times (sec) for original (solid) and enhanced (dashed) speech. (* : paired difference not equal zero)

Table 11: Differences (enhanced speech – original speech) of means, standard deviations (SDs), and 95% confidence intervals (CIs) of response times.

SNR	Mean difference	SD difference	95% CI difference
-25 dB	-0.04	0.08	-0.10 ~ 0.02
-20 dB	-0.05	0.03	-0.07 ~ -0.03
-15 dB	-0.07	0.04	-0.10 ~ -0.04
-10 dB	0.005	0.05	-0.03 ~ 0.04
-5 dB	0.02	0.03	-0.001 ~ 0.04
0 dB	0.03	0.04	0.01 ~ 0.06

These results showed that at lower SNRs (-25, -20, and -15 dB), speech could be enhanced by selectively amplifying the non-tonal component, suggesting that transitional information in speech may be an important cue to the speech discrimination in severe noise. At higher SNRs, the recognition score of enhanced speech was similar to the recognition score of the original speech. The response times for enhanced speech were shorter than the response times for original speech at lower SNRs but longer at the highest SNR.

5.3 TESTS ON ENHANCED AND PSEUDO-ENHANCED SPEECH

The objective of this evaluation is to examine whether speech can be enhanced by fixed frequency filtering as effectively as by the time-varying filter described in the previous section. The fixed filter function should mimic the speech enhancement process of the time-varying filter algorithm by producing an output signal that is as close to the enhanced speech generated by the time-varying filtering as possible. To achieve this, the long-term averaged spectra for original and enhanced speech were calculated for the rhyming words, and a filter function to provide output speech with the same long-term averaged spectrum as the enhanced speech was calculated. Each original word was filtered by this filter function to generate what is referred to as pseudo-enhanced speech. The relative intelligibility of the enhanced and pseudo-enhanced speech versions were compared, using the modified rhyme protocol described in section 2.3.2.

5.3.1 Methods

Three hundred mono-syllable words (50 sets of rhyming words) described in section 5.2.1 were used to calculate the long-term averaged spectra of original and enhanced speech. A

periodogram was calculated for each word, and the averaged square-roots of the long-term spectra (ASRLSs) for original and enhanced speech were calculated by averaging the square-roots of these periodograms for the original and enhanced words. A pseudo-enhanced filter function to generate the ASRLS of enhanced speech from the ASRLS of original speech was calculated as

$$F(w) = \frac{X_{Enha}(w)}{X_{Orig}(w)} \qquad (5.2)$$

where $F(w)$ is the pseudo-enhanced filter function, $X_{Orig}(w)$ is the ASRLS of original speech, and $X_{Enha}(w)$ is the ASRLS of enhanced speech. Pseudo-enhanced speech was obtained by applying this zero-phase filter to original speech as shown in Figure 40.

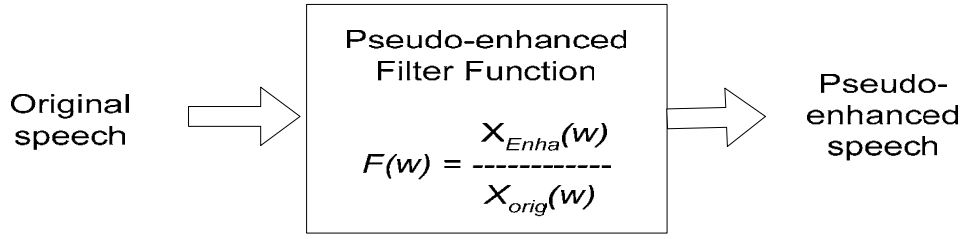

Figure 40: A block diagram of pseudo-enhanced filter function

The long-term averaged spectra of original and enhanced speech are shown in Figure 41 (a) and (b). The long-term averaged spectrum of original speech shows that most of the spectral energy is located in the low frequency region. The energy of the long-term averaged spectrum of the enhanced speech is more evenly distributed across frequencies. The information from 2500 Hz to 3300 Hz is relatively emphasized in the enhanced speech. The pseudo-enhanced filter

function, calculated as the ratio of these ASRLSs, is shown in Figure 41 (c). The filter function shows energy attenuation in the lower frequency regions and energy amplification in the middle to high frequency regions.

Each test word was passed through the pseudo-enhanced filter to create a pseudo-enhanced word. The energy of the pseudo-enhanced word was adjusted to be equal to the energy of the enhanced word. The long-term averaged spectra of enhanced and pseudo-enhanced speech are shown in Figure 42, where the solid line represents enhanced speech and the dashed line represents pseudo-enhanced speech. These two spectra show similar energy distributions at most frequencies, although the magnitude of pseudo-enhanced speech is a few dB higher than the magnitude of enhanced speech from 800 Hz to 1500 Hz.

The original, enhanced, and pseudo-enhanced speech are compared to investigate whether the speech structures of enhanced and pseudo-enhanced speech are different and whether the transient information is more effectively amplified in the enhanced speech than in the pseudo-enhanced speech. Waveforms and spectrograms for original, enhanced, and pseudo-enhanced speech (mono-syllable word “Meat” - represented phonetically as /*mit*/) are compared in Figure 43 and 44. The energies of the enhanced and pseudo-enhanced speech were adjusted to be equal to the energy of the original speech. The enhanced speech is dominated by the onset of */i/* at approximately 0.12 seconds (marked by an arrow on figure), and it also emphasizes the stop /*t*/ at 0.42 seconds. The pseudo-enhanced speech shows characteristics similar to the enhanced speech, but the onset of */i/* and stop /*t*/ are less emphasized in the pseudo-enhanced speech. In addition, the original harmonic structure of the vowel sound */i/* from 0.12 to 0.26 seconds is clearly retained in the pseudo-enhanced speech. Although the enhanced speech contains these harmonic energies (because a base speech in enhancement was the original speech), they are not

as regular as in the pseudo-enhanced speech. These differences between enhanced and pseudo-enhanced speech structures were typical through the 300 rhyming words.

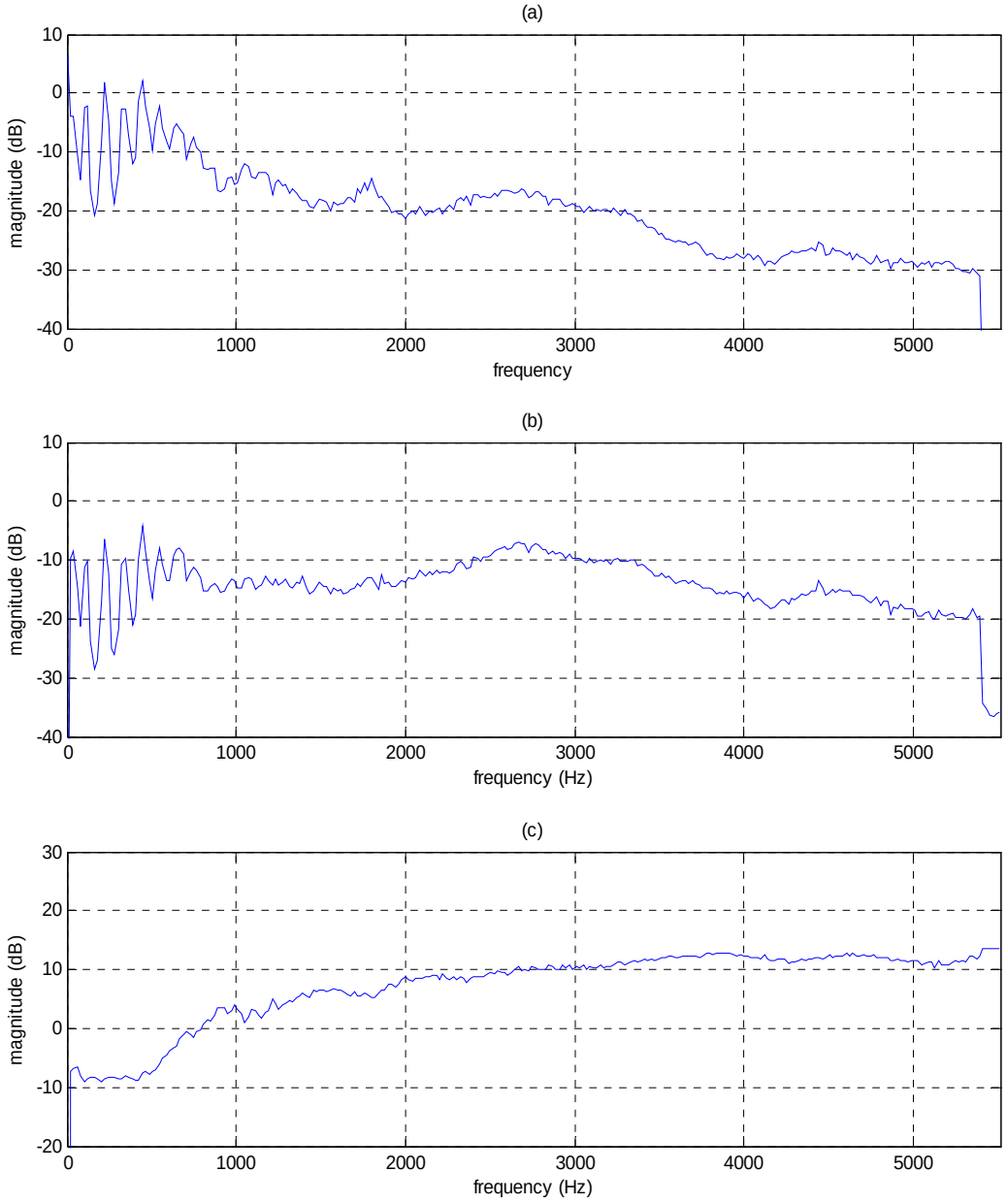

Figure 41: The long-term averaged spectra of (a) original and (b) enhanced speech and (c) the pseudo-enhanced filter function.

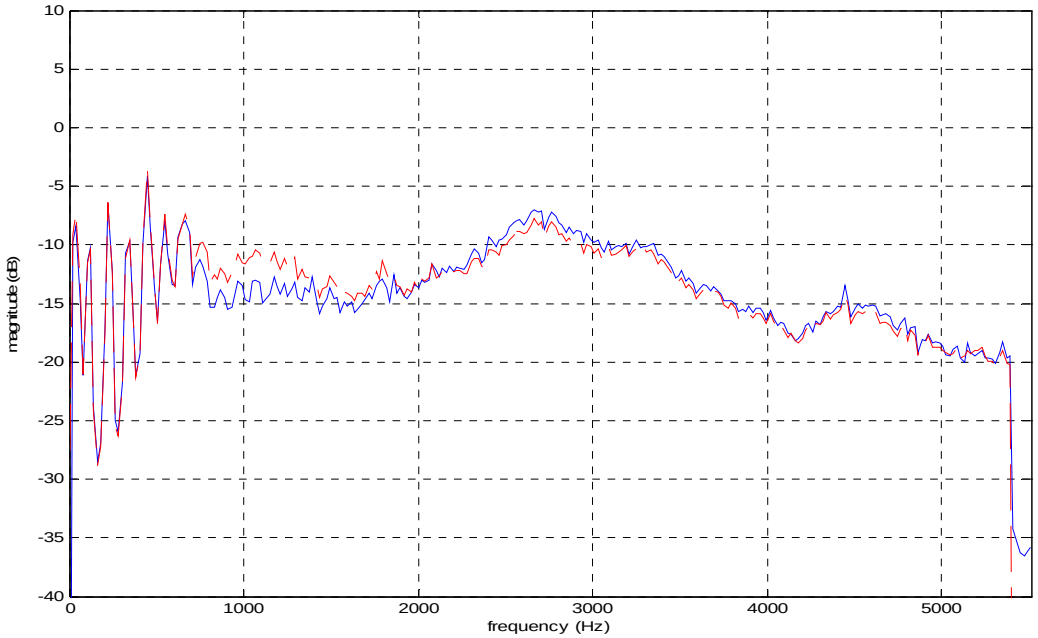

Figure 42: The long-term averaged spectra of enhanced (solid) and pseudo-enhanced (dashed) speech.

The relative intelligibility of the enhanced and pseudo-enhanced speech versions were compared by the modified rhyme protocol described in sections 2.3.2 and 5.2.1. Most test procedures were the same as described in section 5.2.1. Test words were only presented at -20 dB SNR level of speech-weighted background noise rather than 6 different SNR levels as described in section 5.2.1. The goal of this test was to examine whether speech can be enhanced by fixed frequency filtering as effectively as by time-varying filtering, so that testing with only one SNR level that well discriminates the recognitions of original and enhanced speech and provides reasonable intelligibility of the enhanced speech was adequate. Also, this test setting provided more statistical power in the test (more words presented at one SNR level for a hypothesis test.).

Five volunteer subjects with negative otologic histories and hearing sensitivity of 15 dB HL or better by conventional audiometry (250 – 8 kHz) were tested. These subjects were different subjects than those used in the previous experiment (section 5.2.1). Each subject had a training trial consisting with 12 sets of rhyming words (total 72 words) as described in section 5.2.1. In the main trial, 50 sets of rhyming words were repeated 4 times for each subject (total 200 sets of rhyming words per subject). One hundred sets were presented as enhanced speech and 100 sets were presented as pseudo-enhanced speech. The order of presentations was randomized. The 200 target words were randomly selected from the 300 mono-syllable words and the selected target word was excluded in the future selections (the same target word did not appear more than once.). Subject responses were recorded by the computer, and the test results, including number of correct responses and response times, were saved.

The test procedures were monitored by skilled examiners under supervision of a certified clinical audiologist. Means and standard deviations of recognition scores and response times for each subject and each speech type were computed using MATLAB (The Mathworks, Inc.). The data distributions of recognition scores and response times for enhanced and pseudo-enhanced speech were examined and the data appeared to be normally distributed. The recognition scores and response times obtained for the enhanced and pseudo-enhanced versions of the words were compared by the paired t-test.

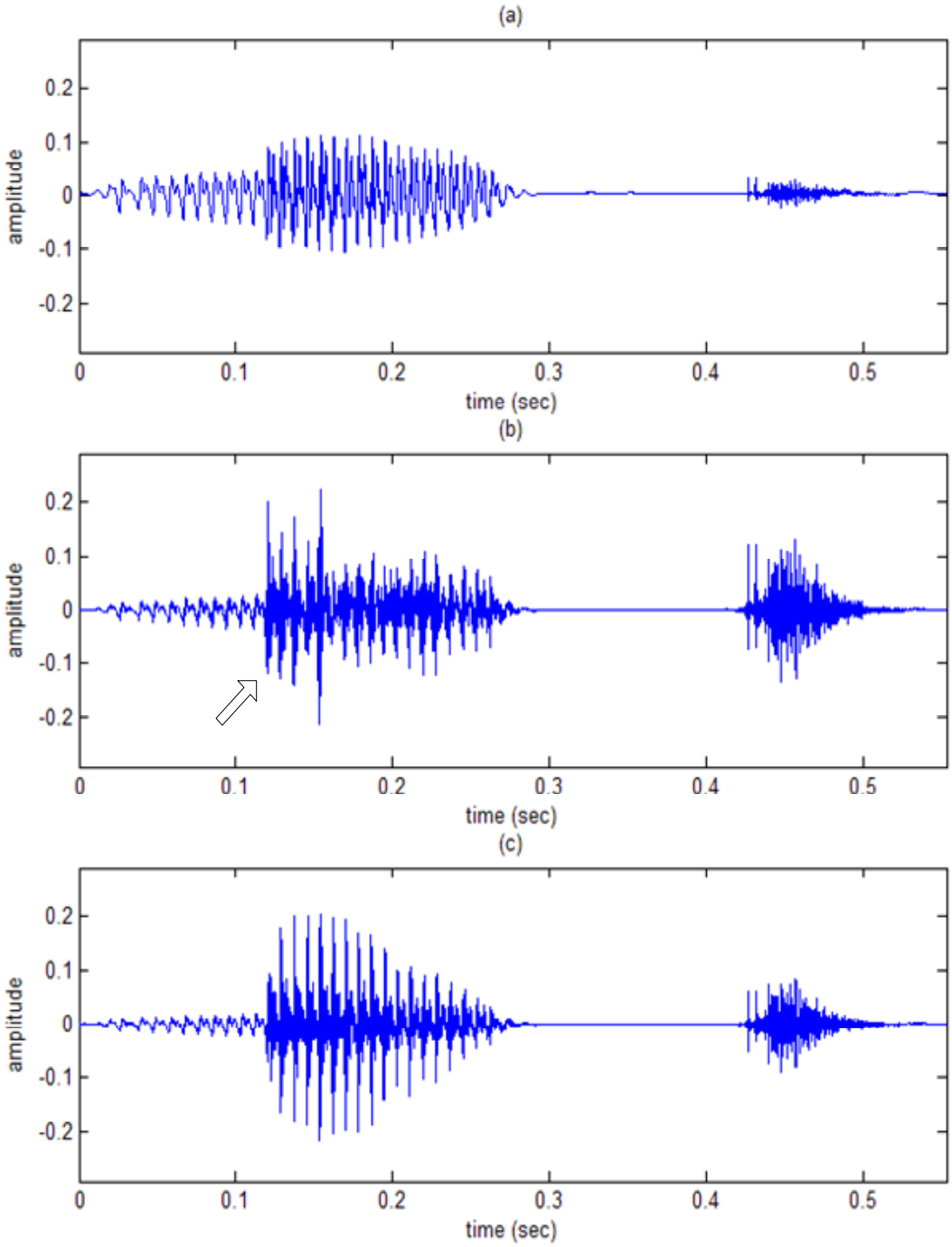

Figure 43: Waveforms of a mono-syllable word “Meat” spoken by a male speaker: (a) original, (b) enhanced, (c) pseudo-enhanced speech

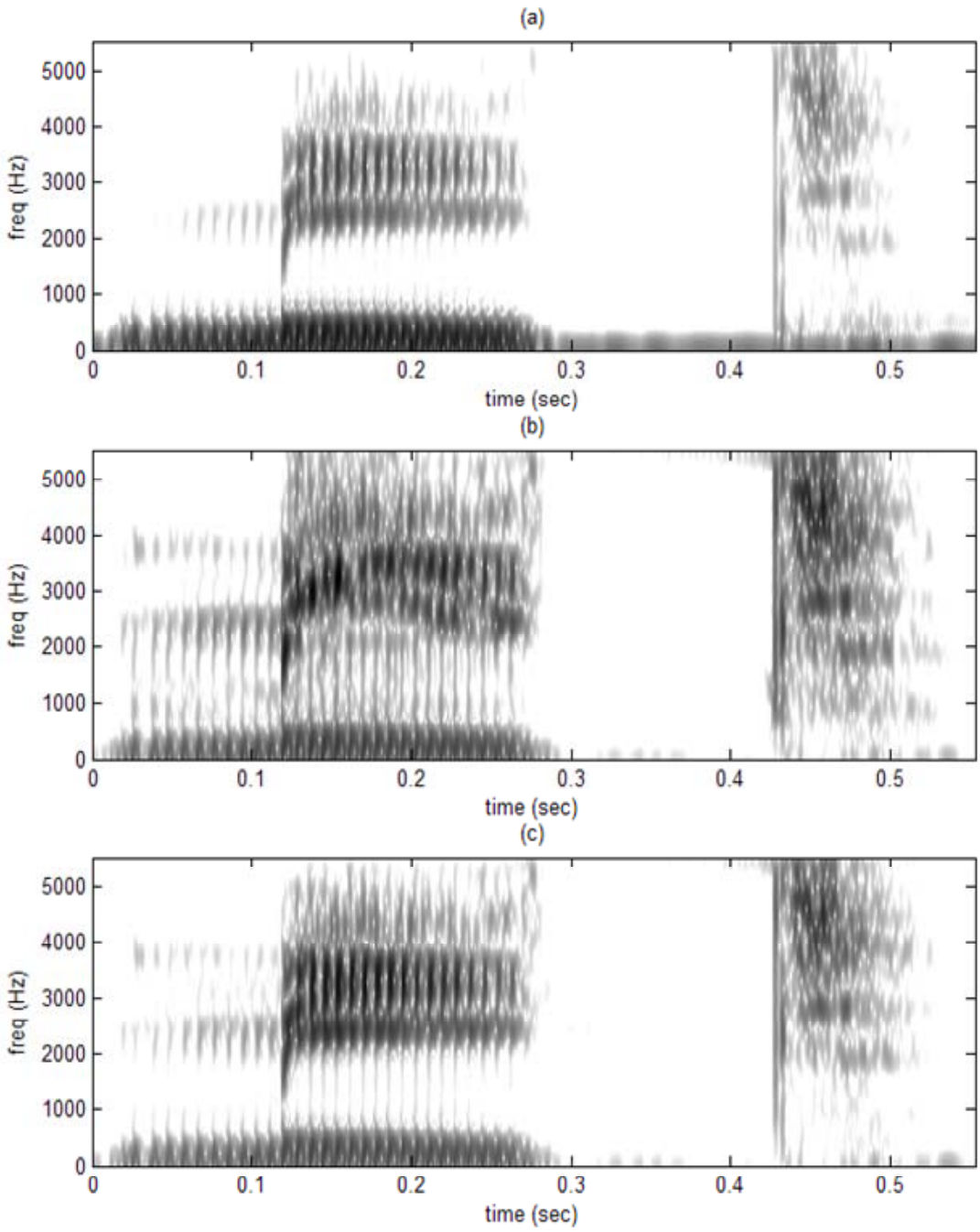

Figure 44: Spectrograms of a mono-syllable word “Meat” spoken by a male speaker: (a) original, (b) enhanced, (c) pseudo-enhanced speech

5.3.2 Results

The recognition rate and response time of enhanced speech were about the same as in the previous test (section 5.2.2). Differences of means, standard deviations, and 95% confidence intervals of the word recognition scores and response times for enhanced and pseudo-enhanced speech are shown in Table 12. The differences of recognition scores and response times between enhanced and pseudo-enhanced speech versions were not significant.

Table 12: Differences (pseudo-enhanced speech – enhanced speech) of means, standard deviations (SDs), and 95% confidence intervals (CIs) of word recognition scores (WRSs) and response times (RTs).

	Mean difference	SD difference	95% CI difference
WRS (%)	5.2	5.4	-1.57 – 11.97
RT (sec)	-0.001	0.02	-0.03 – 0.03

These results showed that at -20 dB SNR, speech can be enhanced by fixed frequency filtering as effectively as by the time-varying filter. Although, qualitatively, enhanced speech appears to emphasize transitions in speech more effectively than pseudo-enhanced, the recognition scores and response times were not significantly different.

6.0 DISCUSSION AND FUTURE RESEARCH

6.1 DISCUSSION

We have introduced a new dynamic method to extract transient information from speech. Time-varying bandpass filters whose center frequencies and bandwidths are controlled to pass most of the energy in the three largest formant components in highpass filtered speech were designed to extract tonal energy. The tonal component was composed of the sum of the filter outputs, and we referred to the signal with the tonal component removed as the non-tonal component of speech. Rao and Kumareasan's study focused on the slowly-varying tonal component, but we are focusing on what they eliminated – non-tonal component [18]. We suggest that the non-tonal component primarily represents transitions between and within vowels and hubs of consonants.

Speech sounds can be classified as vowels and consonants. Vowels contain most of the speech energy and are dominated by lower frequencies. Consonants contain less energy and the energy is distributed in higher frequency regions. Traditional methods of studying the auditory system and speech recognition have generally emphasized the steady-state vowel sounds rather than consonant sounds. In this study, most of the steady-state vowel energy was removed from speech signals, and the remaining sounds which contained most of the consonant energy were analyzed.

The basic idea of this study is that the auditory system may be particularly sensitive to time-varying frequency edges, which probably reflect the transition components in speech.

Although these transitions represent a small proportion of the total speech energy, they may be critical to speech perception. In order to investigate the role of non-tonal speech components on speech intelligibility, mono-syllable words were decomposed into tonal and non-tonal components, and the energy and intelligibility of each component was tested psychoacoustically. The non-tonal components have less energy than the original speech, but psychometric measures of maximum word intelligibility showed almost equal intelligibility to original speech. The tonal components had much greater energy but were significantly less intelligible. We suggest that the tonal component corresponds to speech energy that characterizes sustained vowel sounds and consonant hubs.

These results suggest that non-tonal components are important in speech perception. If the auditory system is sensitive to transient information, emphasis of the non-tonal components may provide a method to enhance intelligibility, especially in noisy conditions. The transients are expected to be distributed across time and frequency and may require time-frequency techniques to identify them. The decomposition algorithm described in this study provides one method of extracting a signal that emphasizes transient speech components.

Most traditional studies of speech enhancement have focused on noise reduction. In this study, the speech signal itself was enhanced by the time-varying filters. The non-tonal component was amplified and recombined with the original speech, and the intelligibility of the enhanced speech was compared to the original speech in background noise. The psychometric measures of word intelligibility demonstrate that the enhanced speech can provide significant improvement in speech intelligibility at low SNR levels. At higher SNR, the differences between original and enhanced speech were not significant because the noise was relatively soft so that the transient information for both original and enhanced speech was not greatly affected.

At lower SNR levels (-20 and -15 dB), the response times for enhanced speech were shorter than the response times for the original speech. At these SNR levels, the enhanced speech showed higher word recognition scores than the original speech. These results – higher word recognition scores with shorter response times and lower word recognition scores with longer response times - support the relations between word recognition scores and response times observed by Mackersie *et al.* [32]. At the highest SNR level (0 dB), the response times for original speech were shorter than the response times for the enhanced speech. The word recognition scores between original and enhanced speech were not significantly different at this SNR level. These results may indicate that, compared to original speech, subjects increase listening efforts to identify the enhanced speech at high SNR level [38].

These results suggest that amplification of transient information can enhance speech in noise. This enhancement method can be applied to any speech communication system where clean speech can be accessed but outside noises interrupt the communications, such as cellular phone communications in a loud restaurant, communications between control office and firefighters, battle field communications between command center and soldiers *etc*.

Another way to implement the enhancement method was evaluated. A fixed frequency filter function designed to generate the long-term averaged spectrum of enhanced speech from the long-term averaged spectrum of original speech was calculated, and pseudo-enhanced speech was generated by filtering the original speech by this filter function. The relative intelligibility of the enhanced and pseudo-enhanced speech versions were compared by the modified rhyme protocol.

The difference between enhanced and pseudo-enhanced speech versions was not significant. In general, the transient information is distributed in higher frequency regions, and

the transient information may be emphasized in the pseudo-enhanced speech because of the frequency characteristics of the filter function. These results suggest that for a specific speech material and speaker, speech enhancement similar to that obtained with a time-varying filtering can be archived by a fixed frequency filter [5], [6], [7]. However, this experiment was constrained to a certain condition (i.e. single speaker, single speech material, word identification rather than conversational task, *etc*.) and the fixed filter may not be robust to speaker, speech material, or environment. Preliminary studies suggest that different filter functions are obtained under different conditions (Appendix E). Since the pseudo-enhanced speech was calculated based on the time-frequency techniques, time-frequency techniques may be required to define the filter functions for various conditions.

Waveforms and spectrograms of original, enhanced, and pseudo-enhanced speech sounds were compared. The pseudo-enhanced speech shows characteristics similar to the enhanced speech, but the harmonic structures of vowel sounds were more clearly retained in the pseudo-enhanced speech. Although pseudo-enhanced speech may contain some of the same amplified transient information, the enhanced speech appears to emphasize it more than the pseudo-enhanced speech does. These differences may be caused by the time-varying filtering, and the time-frequency techniques may be more effective to emphasize transition information in speech.

6.2 FUTURE RESEARCH

- The first task for future work would be to establish a better understanding of the tonal and non-tonal components. Quantitative measures of the component differences, including definitions of component characteristics in terms of time and frequency and role in speech perception, could be a significant contribution to the field of speech enhancement.

- The decomposition algorithm is based on the computation of time-varying filters. Other approaches for speech decomposition are interesting for future studies. For examples, wavelets and cosine transform can be applied for new decompositions. Comparisons between the time-varying filter approach and other candidates may refine the speech decomposition as well as provide a solution to speed up the processing.
- Recent findings in auditory research suggest that a nonlinear active process in outer hair cells (OHCs) may play a role in the processing of noisy speech, and this role may be related to the processing of transition information. Measuring otoacoustic emissions, sounds produced by outer hair cells as a byproduct of signal transduction in the cochlea, is a non-invasive method to measure the response of OHCs to different types of acoustic stimuli. Comparisons between otoacoustic emissions derived from tonal and non-tonal components will be an interesting subject for future research. These comparisons will characterize the responses of OHCs to transition and quasi-steady-state stimuli and provide better understanding of OHC functions. These studies will facilitate the design of algorithms to identify and process transition components in speech.
- Psychoacoustic evaluations to examine whether speech can be enhanced by fixed frequency filtering rather than time-varying filtering were performed. The intelligibility difference between the two versions was not significant for a specific speaker, environment, and speech material. Psychoacoustic evaluations for the enhanced and pseudo-enhanced speech with different speakers, speech materials, and environments will be interesting for future study. These evaluations will characterize the differences between enhanced and pseudo-enhanced speech that were not revealed at the previous

evaluation and possibly identify the optimal filters across speakers, speech materials, and languages.

- Waveforms of original, enhanced, and pseudo-enhanced speech were illustrated in section 5.3.1. Quantitative measures of the waveform differences and effects of these differences to the results of psychoacoustic tests will be interesting for future research.

APPENDIX A

ANALYTICAL TESTS OF SPEECH INTELLIGIBILITY

This appendix describes preliminary evaluations of some analytical techniques to quantify the effectiveness of speech enhancement procedures. The articulation index (AI) and speech intelligibility index (SII) were calculated from statistical models, developed from articulation theory. Two different automatic speech recognition systems (BBN Byblos and Dragon systems) were used to estimate recognition scores for original speech, highpass filtered speech, tonal component, and non-tonal component. The purpose of this investigation was to determine whether these models and systems could indicate the same differences in speech intelligibility that were observed in psychoacoustic tests (section 5.1) and hence provide a preliminary indication of speech enhancement effectiveness to guide psychoacoustic testing.

Articulation Index And Speech Intelligibility Index

The intelligibility of speech usually improves as speech energy increases from barely audible levels to higher levels [24]. The intelligibility of speech, however, may not improve if the energy of speech is increased to excessive levels.

The relations between frequency and speech intelligibility were studied by French and Steinberg [5]. They investigated the intelligibility of highpass and lowpass filtered speech with

varying cut-off frequencies and found that increasing amounts of either highpass or lowpass filtering resulted in reduced intelligibility. Increasing the cut-off frequency of a highpass filter above 1.5 kHz or lowering the cut-off frequency of a lowpass filter below 3 kHz caused intelligibility to drop below 80% of the unfiltered speech, respectively.

In general, speech intelligibility tests with human subjects require extensive time and effort. To avoid these complex tests, statistical models based on articulation theory have been proposed [5], [6], [7], [8]. Articulation theory has been developed from the results of speech intelligibility tests and used to calculate articulation indices (AIs) based on the statistics of the tests [5], [6], [7], [8]. The AI is basically designed to predict what the intelligibility of speech would be when transmitted over a particular communication system. Thus, the AI is a physical measure of the communication system. The speech intelligibility index (SII) was developed from the AI to provide a reliable and easily applicable method to predict speech intelligibility [9]. Recently, this method was introduced as an ANSI standard [9].

Implementations of AI and SII

In this study, the AI model by Fletcher and Galt was implemented because their model has been suggested to predict speech intelligibility more accurately than other models [5], [6], [7], [8], [39]. The major difference between Fletcher and Galt's AI model and other models is perceptional considerations. In Fletcher and Galt's model, various parameters (e.g. loudness, critical bands, masking by noise and speech itself, and speech detection thresholds) are used to transform physical measurements on speech into the perceptional domain, and the AI was calculated from not only physical measurements of speech but also perceptional correlations of these parameters. In the other AI models, the AIs are directly calculated from the physical

measurements of speech with only minor consideration of perceptional correlations. Rankovic compared Fletcher and Galt's AI model with ANSI S3.5 (1969) [39]. He compared these AIs with recognition scores obtained from speech intelligibility tests by human subjects and concluded that Fletcher and Galt's calculation was more accurate than ANSI S3.5 (1969).

The AIs by Fletcher and Galt were calculated from the system response (transfer function) of the transmission channel, estimated by the ratio of the output to input spectrum. The system response was divided into 20 fixed narrow frequency bands, and the spectral energy of the system response in each band was estimated [7]. The AI was calculated from these spectral energies as a function of the channel gain.

Speech intelligibility index (SII) was a measure that was correlated with the intelligibility of speech under various listening conditions. The SII was calculated from the input variables, including speech spectrum level, noise spectrum level, and hearing threshold level. Different frequencies contributed different amounts to speech intelligibility and a higher SNR contributed to intelligibility within a certain range. For SII calculations, the speech spectrum was divided into 18 fixed narrow frequency bands defined in the standard [9]. For each band, a spectral energy was calculated, and the SII was determined from the spectral energy as a function of the channel gain.

Results of AI and SII

Five mono-syllable words spoken by a female speaker were randomly selected for examination (from the audio CDROM that accompanies Contemporary Perspectives in Hearing Assessment, by Frank E. Musiek and William F. Rintelmann, Allyn and Bacon, 1999). These speech samples were "South", "Dab", "Juice", "Nice", and "Pick". For each word, the AIs and

SIIs for the original signal, highpass filtered signal, tonal component, and non-tonal component were calculated. The spectra of the original signals were considered as input spectra to the transmission channels in the AI calculations. Average spectral energy, used in the AI and the SII calculations, for the five speech samples is shown in Figure A1 and A2. The solid, dashed, dotted, and dot-dashed lines are associated with the original signal, highpass filtered signal, tonal component, and non-tonal component, respectively. The differences in spectral energies between the original speech and the other components are more pronounced in the low frequency bands, below 750 Hz, than in the high frequency bands. These differences in the low frequency bands should not affect the intelligibility of these components [35].

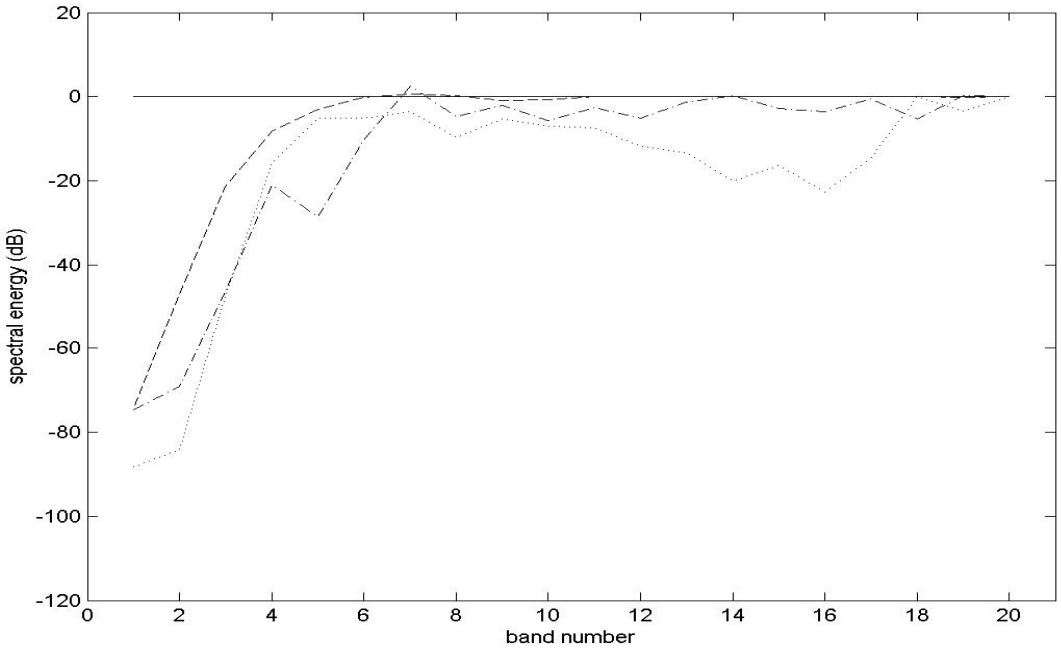

Figure A1: Ensemble spectral energies in each band for five speech samples, original speech signal (solid), highpass filtered speech signal (dashed), tonal component (dotted), and non-tonal component (dot-dashed). These spectral energies were used in the AI calculations.

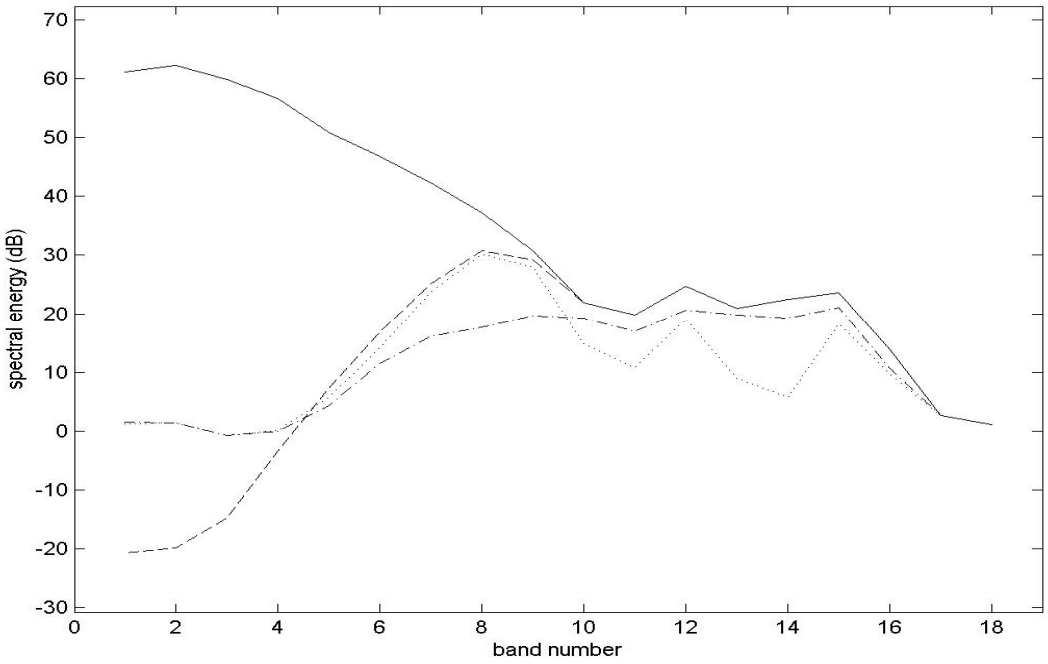

Figure A2: Ensemble spectral energies - original speech signal (solid), highpass filtered speech signal (dashed), tonal component (dotted), and non-tonal component (dot-dashed) – for the five speech samples. These spectral energies were used in the SII calculations.

Ensemble AIs and SIIs for each component across five speech samples are shown in Figure A3 and A4. The AI values at 0 dB of channel gain for the original speech signal, highpass filtered speech signal, tonal component, and non-tonal component were 0.95, 0.80, 0.66, and 0.72, respectively. The AI curves for the highpass filtered speech signal, tonal component, and non-tonal component show shifts in channel gains and smaller maximum scores with respect to the original signal. The SII values at 0 dB of channel gain for the original speech signal, highpass filtered speech signal, tonal component, and non-tonal component were 0.99, 0.97, 0.96, and 0.98, respectively. The shapes of all four SII curves are similar, except shifts in channel gains with respect to the SII curve of original signal. When the speech level exceeds a

critical value, further increases in speech level do not result in increased intelligibility. Rather, intelligibility decreases – the so called rollover effect. The AI and SII models predict the rollover effect when the channel gain exceeds 68 dB above the gain which is required to detect speech in quiet [7], [9].

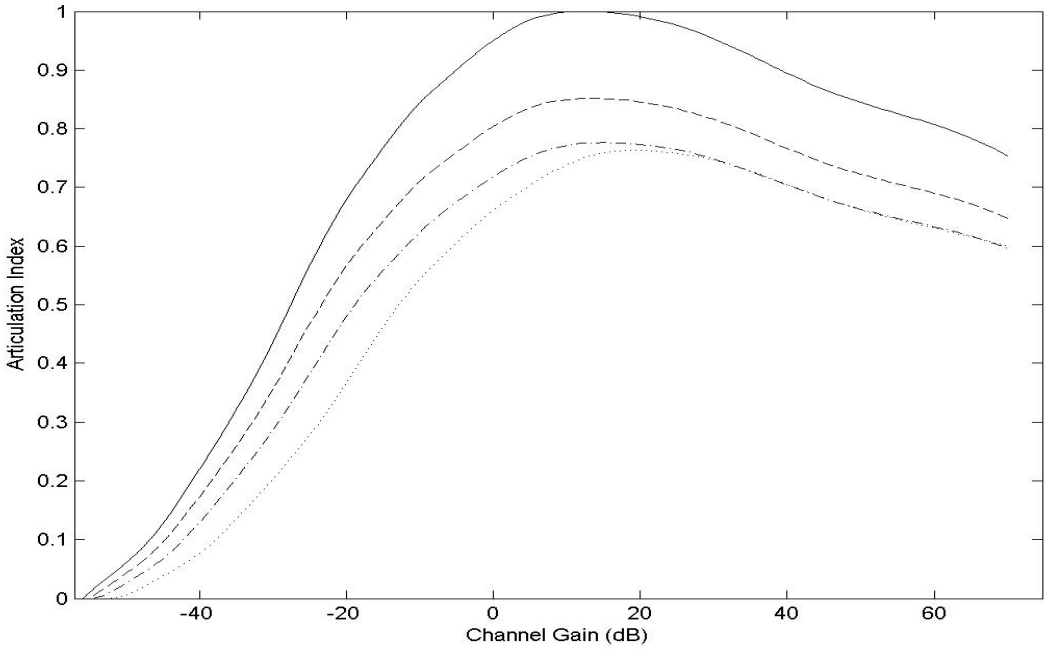

Figure A3: Ensemble AIs for original speech signal (solid), highpass filtered speech signal (dashed), tonal component (dotted), and non-tonal component (dot-dashed) across five speech samples

A system with an AI of less than 0.3 is generally considered unsatisfactory for everyday speech communications [40]. A system with an AI between 0.3 and 0.5 is generally considered barely acceptable and a system with an AI of 0.5 or greater is generally considered as satisfactory. A system with an SII of less than 0.45 is a poor communication system, and a system with an SII of 0.75 or greater is considered to be a good communication system [9]. All

AIs calculated here are greater than 0.5 and all SIIs are greater than 0.75 at 0 dB of channel gain, implying that all four channels (components) are satisfactory in terms of the AI and SII scores.

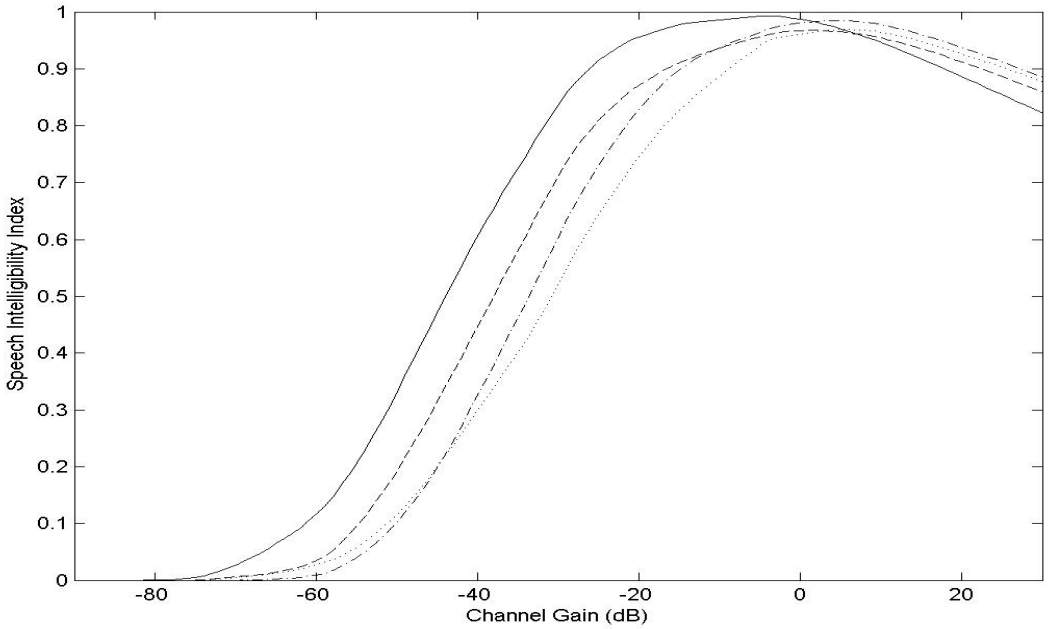

Figure A4: Ensemble SIIs for original speech signal (solid), highpass filtered speech signal (dashed), tonal component (dotted), and non-tonal component (dot-dashed) across five speech samples

AI and SII curves of tonal and non-tonal components are essentially identical. That is, the AI and SII methods do not indicate intelligibility differences between tonal and non- tonal components, even though there are clear differences subjectively. Both AI and SII approaches only consider fixed frequency bands, and the spectral computations in mid-to-high frequency bands, which have greatest weights in the computations, are similar for both tonal and non-tonal components. Using fixed frequency bands may not be appropriate to explain the intelligibility changes caused by the time-varying characteristics of speech, which are captured in the time-

varying filters. Hence, it was concluded that the AI and SII methods will probably not be effective to evaluate the intelligibility differences observed in the preliminary assessments.

Automatic Speech Recognition Test

Speech recognition systems have been developed with various applications from simple keyword recognition to complex sentence dictations [41], [42], [43], and [44]. Although the performance of speech recognition systems has improved over the past two decades, many fundamental questions are still unanswered. The speech signals are time-varying signals, and even the same speaker can produce different speech sounds with different times for the same word. These variations in speech signals make analyzing and designing automatic speech recognition systems difficult.

Two different automatic speech recognition systems (BBN Byblos and Dragon systems) were investigated to test the intelligibility of highpass filtered speech, tonal and non-tonal components with respect to the original speech, and the test results are compared to the results obtained from psychoacoustic tests with human subjects. The basic structures and concepts of automatic speech recognition are introduced below, following references [41], [43], [44], [45], and [46].

In modern automatic speech recognition, the speech signal is treated as a stochastic pattern, and the recognition systems apply statistical pattern recognition approaches to the input speech. Generally, automatic speech recognition systems identify the input speech pattern using pre-defined acoustic models of speech sounds. The acoustic model is generated from the information in a set of speech data, in a process referred to as training. The performances of the recognition systems are largely dependent on the selection of the training speech data. Optimal

speech data sets are not currently available, and the selection of the sets has depended on applications.

A widely accepted and efficient model used in automatic speech recognition is hidden Markov models (HMMs). The HMM is a statistical model having a finite number of states and state transitions to model the frequency and time-varying nature of the speech signal. Each state of the model has an observation density function that specifies the probability of state represented as a combination of Gaussians. The parameters of HMMs are usually estimated from training data using the maximum likelihood method.

The input speech is first sampled and then digitized. The start and end of the digitized input speech are detected, and acoustic feature vectors of the input speech are extracted. The acoustic feature vectors, consisting of parameters that contain recognition information about the sounds in the utterance, are selected to have good discrimination for distinguishing speech sounds and statistical properties that are relatively invariant across speakers and speaking environments.

The extracted acoustic feature vectors are compared to the trained acoustic model. This process is referred to as recognition. The recognition task can be described as finding the most likely sequence of words through the network such that the likelihood of the observed acoustic features is maximized. Two methods - modular and integrated approaches - are typically used to find the maximum likelihood. In the modular approach, each module is considered as one knowledge source and the speech feature of each frame is matched to the acoustic model in a sequential manner. Each module is tested and designed separately, so that the specific module can be easily modified and developed without affecting the other modules. The drawback of this approach is that in each module, decisions are made without knowledge of the other modules,

and decision errors are likely to be propagated from one module to the next and accumulate to cause search errors.

The integrated approach uses all knowledge sources to make a decision. Therefore, all the knowledge sources need to be characterized and integrated for the network to achieve high performance. This approach is generally used in modern automatic speech recognition systems. The drawback of this approach is that, compared to the modular approach, much more computation is required because all of the knowledge sources are utilized simultaneously to reach recognition decisions.

In general, automatic speech recognition systems are evaluated by calculating the rates of incorrect recognitions – word error rates. The word error rate represents the fraction of input speech samples that are not correctly recognized by the recognition system. The improvement of system performance can be measured by a reduction in word error rate. In this study, the word error rate was used as a measure of the intelligibility of input speech samples.

Automatic Speech Recognition Test - BBN Byblos System

The Byblos system (BBN Systems and Technologies) was one of two systems which was used in this project for testing the relative intelligibility of highpass filtered speech, tonal and non-tonal components with respect to the original speech. In Byblos, speech is modeled statistically, and the system attempts to determine the parameters of the model. Then, the system compares the acoustic features of input speech to the pre-defined parameters of the model to determine the most likely words.

The Byblos system considers speech as the output of a HMM. The acoustic model of the system is produced by acoustic training based on the HMM. A 5-state HMM is used to model

each phoneme. The HMM is controlled by the transition probabilities of the Markov chains and the output densities associated with the states. These parameters are updated in each speech frame to increase the likelihood of the observed speech.

The Byblos system extracts speech features from raw audio files by examining only one frame at a time. The system computes the energy and its cepstral coefficients for every 10 msec frame, and these features are grouped as a vector. The entire group of features is used in the recognition process.

The Byblos system uses knowledge of language to increase recognition performance. The language model is generated by analyzing the probabilities of word sequences in a series of documents. A grammar in the language model is built from the statistical patterns of word sequences and forces fewer and more accurate choices in the recognition.

The recognition step in the Byblos system is performed by comparing the extracted speech features to the acoustic and language models. The system first uses less detailed models to reduce the number of word candidates and then uses more detailed models to finally choose the best word.

The speech data set used in the training and recognition steps is referred as a corpus, and the selection of the data set significantly affects the performance of the Byblos system. In general, the selections of the corpus have depended on the type of application. The corpus includes information on speaker genders and timing information of beginning and ending of sentences.

The Byblos system was tested to determine whether the system can identify the intelligibility differences of decomposed components observed in preliminary intelligibility tests

and psychoacoustic tests. The system was trained by the original training corpus, and word error rates for decomposed testing corpora were calculated.

The Byblos system was installed on parallel connected computers. The testing speech data were decomposed into highpass filtered speech, tonal component, and non-tonal component by Matlab software (The MathWorks, Inc., USA). A 5-state hidden Markov model was used in the training process. Speech is analyzed with a 25 msec frame duration at a rate of 100 frames/sec. Forty LPC coefficients were used for LPC smoothing and 256 FFT points were used for processing each frame. The Byblos system was trained by the original training corpus, and code books were generated through this training process, by another research group in Electrical Engineering at the University of Pittsburgh [47]. The code books contained training information, such as acoustic and language model parameters. Based on the code books, the original speech, highpass filtered speech, tonal component, and non-tonal component in the testing corpora were transcribed (decoded) to compare to the already known correct transcription. The word error rates (in %) for original speech, highpass filtered speech, tonal component, and non-tonal components were calculated in the decoding process.

Conversational telephone speech data were used in training experiments [47]. For the training corpus, the Swbd40hrs, a gender-balanced 40 hours subset of the Switchboard training corpus, was used. The corpus was composed of 364 female and 386 male speakers. The female speech, 19 hours 56 minutes long, contained 20,993 utterances, and the male speech, 20 hours 7 minutes long, contained 18,478 utterances. The speech data were composed of telephone conversations between two speakers. The audio files were recorded with two channels, and each channel hosted one side of the telephone conversation. The sampling frequency was 8 kHz and the format of the audio files was NIST_1A format.

For decoding experiments, the Hub5.English.Dev01 corpus was used. This corpus was composed of conversational telephone speech data and generated from the 2001 Hub-5 Evaluation [47]. The corpus composed of 23 female and 25 male speakers. The female speech was 1 hour 13 minutes long and contained 1123 utterances. The male speech was 1 hour 18 minutes long and contained 1135 utterances. This corpus included three different conditions – original Switchboard, Switchboard-2 Phase-3, and Switchboard-2 Phase-4 (cellular phone conversation). The Hub5.English.Dev01 corpus contained Switchboard-2 Phase-4 conditions which were not observed in the training corpus, Swbd40hrs.

The decoding results for each of the decomposed components are summarized in Table A.1. The original speech corpus has a word error rate of 68.6%, while the highpass filtered speech corpus displays a 93.4% word error rate. The tonal component corpus shows a word error rate of 93.1% and the non-tonal component corpus has a word error rate of 93.3%. The word error rate increases by 24.8% in highpass filtered speech corpus compared to the original speech corpus. The word error rate of 93.1% in the tonal component corpus is 24.5% greater than the word error rate of the original speech corpus. The non-tonal component corpus is 24.7% greater than the word error rate of the original speech corpus. In essence, the system was not able to recognize the highpass filtered speech, tonal component, or non-tonal component corpora.

Intelligibility characteristics observed in the psychoacoustic tests (high intelligibility in original, highpass filtered, and non-tonal components and low intelligibility in tonal component) are not shown in these automatic speech recognition tests. Because the time that would be required to decompose all of the training data was prohibitive, no attempt was made to train the system on highpass filtered speech or the non-tonal component.

Table A1: Decoding results (word error rates) for each decomposed component

Testing Corpora	Word Error Rates (%)
Original Speech	68.6%
Highpass Filtered Speech	93.4%
Tonal Component	93.1%
Non-tonal Component	93.3%

Automatic Speech Recognition Test - Dragon System

A commercial system for automatic speech recognition was tested to determine whether the system can identify the intelligibility differences of decomposed components observed in the psychoacoustic tests. The system was trained by the original and highpass filtered speech sounds, and word error rates for each component were calculated.

The Dragon system was installed on a PC computer. The training speech data consisted of multiple paragraphs, including 5 or 6 sentences for each paragraph. The testing speech data consisted of forty-five mono-syllable words. Both training and testing speech data were recorded from a female speaker. The training of the system was performed as follows. The pre-defined training data were decomposed into highpass filtered speech, tonal component, and non-tonal component by Matlab software (The MathWorks, Inc., USA). The system has a built-in acoustic model, and only detailed parameters of the acoustic model are adjusted during the training process. Thus, training speech data, at least, have to be recognized by the built-in acoustic model. The system could be trained by the original and highpass filtered speech sounds but could not be

trained by the tonal and non-tonal components. That is, the built-in acoustic model could not recognize the tonal and non-tonal components at all. The testing speech data were decomposed and decoded by the systems trained with original and highpass filtered speech data. The word error rates (in %) for original speech, highpass filtered speech, tonal component, and non-tonal components were calculated in the decoding process.

The testing results from the systems trained by the original and highpass filtered speech data are presented in Table A.2 and A.3, respectively. When the system was trained by the original speech, the original testing data had a word error rate of 4%, while highpass filtered testing data displayed a 18% word error rate. The tonal component testing data showed a word error rate of 96%, and the non-tonal component testing data had a word error rate of 87%. These results show that most tonal and non-tonal components are incorrectly recognized.

When the system was trained by the highpass filtered speech, the original testing data had a word error rate of 24% while highpass filtered testing data had a 27% word error rate. The tonal component testing data had a word error rate of 93%, and the non-tonal component testing data had a word error rate of 80%. These results demonstrate that most of the tonal component is incorrectly recognized, but the word error rate decreased by 7% in non-tonal component compared to the system trained by the original speech data. Although the Dragon system correctly recognizes original and highpass filtered speech, most tonal and non-tonal components are recognized incorrectly. The differences in intelligibility measures on the non-tonal component between psychoacoustic tests (section 5.1) and Dragon system may imply differences between the human hearing system and the automatic speech recognition systems.

Table A2: Decoding results (word error rates) for each decomposed component (trained by original speech data)

Testing data	Word Error Rates (%)
Original Speech	4%
Highpass Filtered Speech	18%
Tonal Component	96%
Non-tonal Component	87%

Table A3: Decoding Results (word error rates) for each decomposed component (trained by highpass filtered data)

Testing data	Word Error Rates (%)
Original Speech	24%
Highpass Filtered Speech	27%
Tonal Component	93%
Non-tonal Component	80%

Neither automatic speech recognition system was unable to demonstrate recognition of highpass filtered speech and non-tonal component similar to the psychoacoustic results. The recognition using automatic speech recognition systems do not extend effectively beyond the type of speech that they were trained on. Since human listeners do perform recognition over a range of speech types, as demonstrated in the psychoacoustic tests, human listeners probably attend to aspects of speech that are not considered by the automatic speech recognition systems.

Hence, these automatic speech recognition systems will probably have limited effectiveness in evaluating speech enhancement techniques.

APPENDIX B

RELATIVE POSITIVE (NEGATIVE) CHIRP ENERGIES OF TONAL COMPONENTS

Table B1: Relative positive (negative) chirp energies of tonal components for constant frequency change in chirp and constant chirp duration. Key : E_o : Chirp energy of original synthetic signal, E_t : Chirp energy of tonal component

Chirp rate (Hz/msec)	Chirp duration (msec) for constant frequency change (1460 Hz) in chirp	Relative chirp energy in tonal component for constant frequency change (1460 Hz) in chirp 100 x (E_t / E_o) (%)	Frequency change in chirp (Hz) for constant chirp duration (20 msec)	Relative chirp energy in tonal component for constant chirp duration (20 msec) 100 x (E_t / E_o) (%)
133	11	11 (11)	2660	11 (9)
122	12	12 (11)	2440	15 (12)
112	13	15 (14)	2240	17 (14)
104	14	18 (18)	2080	20 (19)
97	15	26 (23)	1940	28 (26)
86	17	36 (34)	1720	37 (34)
81	18	40 (38)	1620	41 (39)
73	20	47 (44)	1460	47 (44)
63	23	63 (61)	1260	61 (59)
58	25	68 (66)	1160	69 (69)
49	30	79 (77)	980	87 (86)
42	35	87 (87)	840	96 (96)
37	40	94 (95)	740	99 (100)
32	45	99 (100)	640	104 (105)
29	50	102 (103)	580	107 (106)
27	55	102 (103)	540	108 (106)
24	60	103 (104)	480	108 (107)

APPENDIX C

DECOMPOSITION RESULTS OF SYNTHETIC CHIRP SIGNAL (4 CHIRPS)

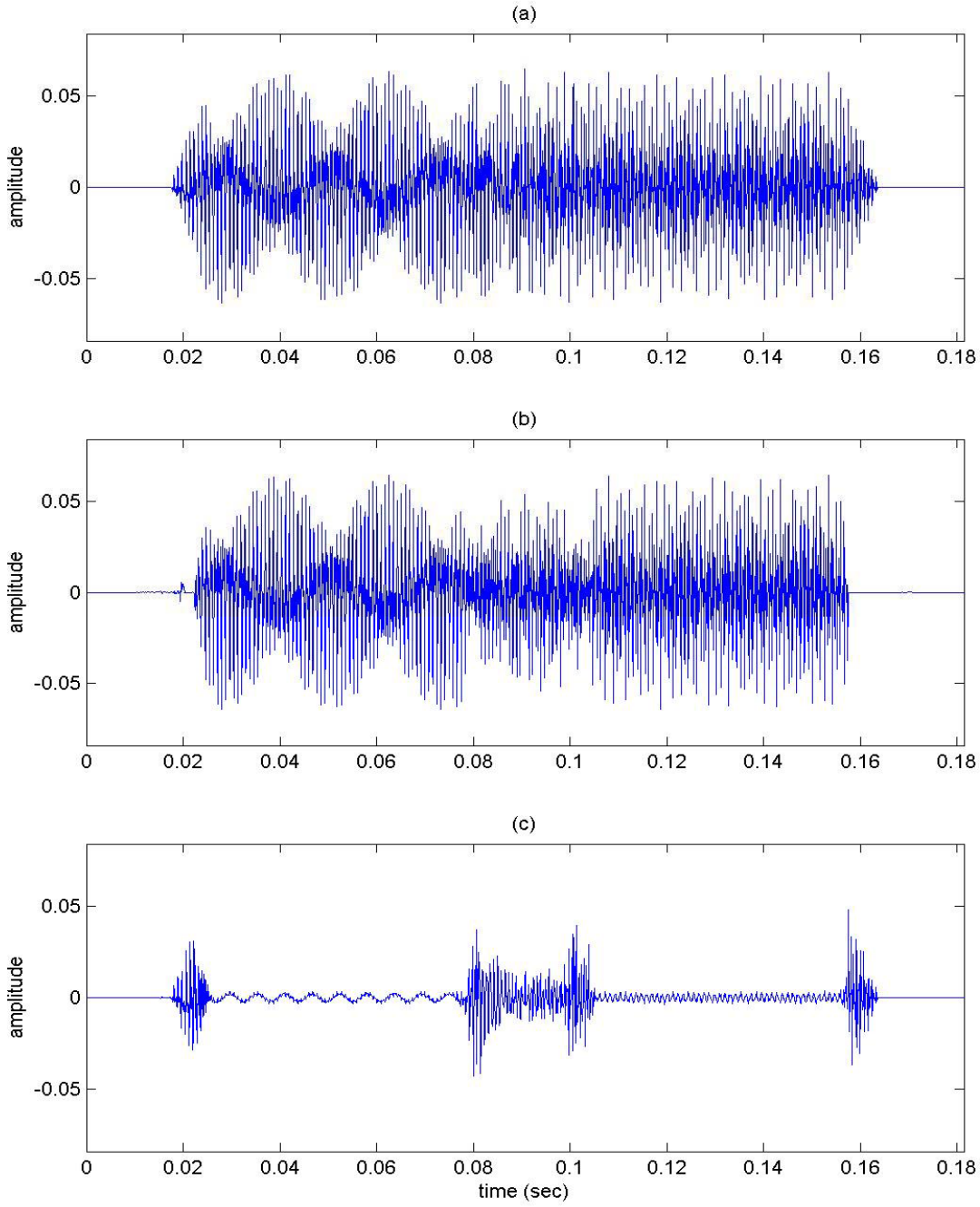

Figure C1: Waveforms of decomposed synthetic chirp signal (4th tone-chirp-tone in low frequency): (a) original, (b) tonal, and (c) non-tonal components. All four tones+chirps+tones had 38 Hz/msec of chirp rates and the chirp durations were fixed at 20 msec.

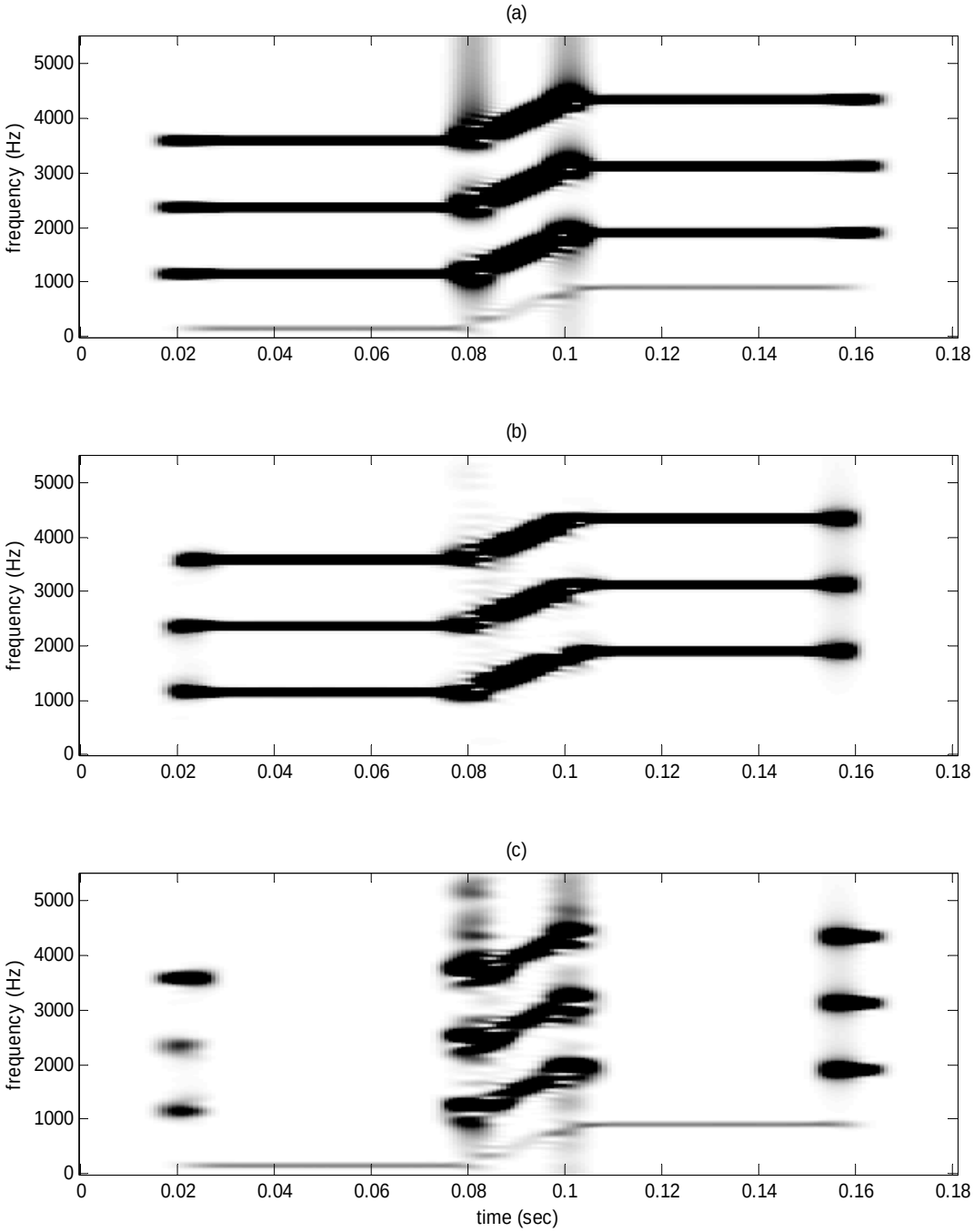

Figure C2: Spectrograms of decomposed synthetic chirp signal (4th tone-chirp-tone in low frequency): (a) original, (b) tonal, (c) non-tonal components

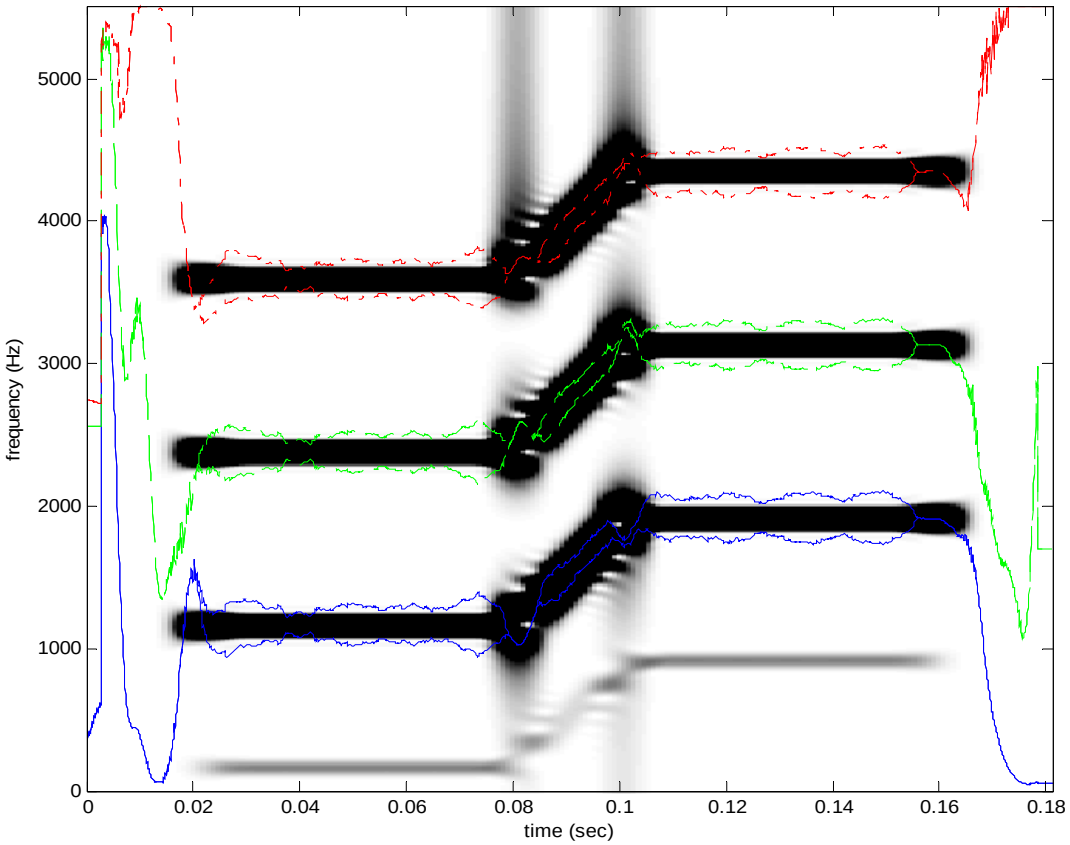

Figure C3: Upper and lower edges of time-varying bandwidths (plotted with spectrogram). 4th tone-chirp-tone in low frequency. The solid, dashed, and dotted lines are associated with the 1st, 2nd, and 3rd time-varying filters, respectively. No filter tracks 4th tone-chirp-tone component in low frequency and the tracking filter was not affected by the 4th tone+chirp+tone.

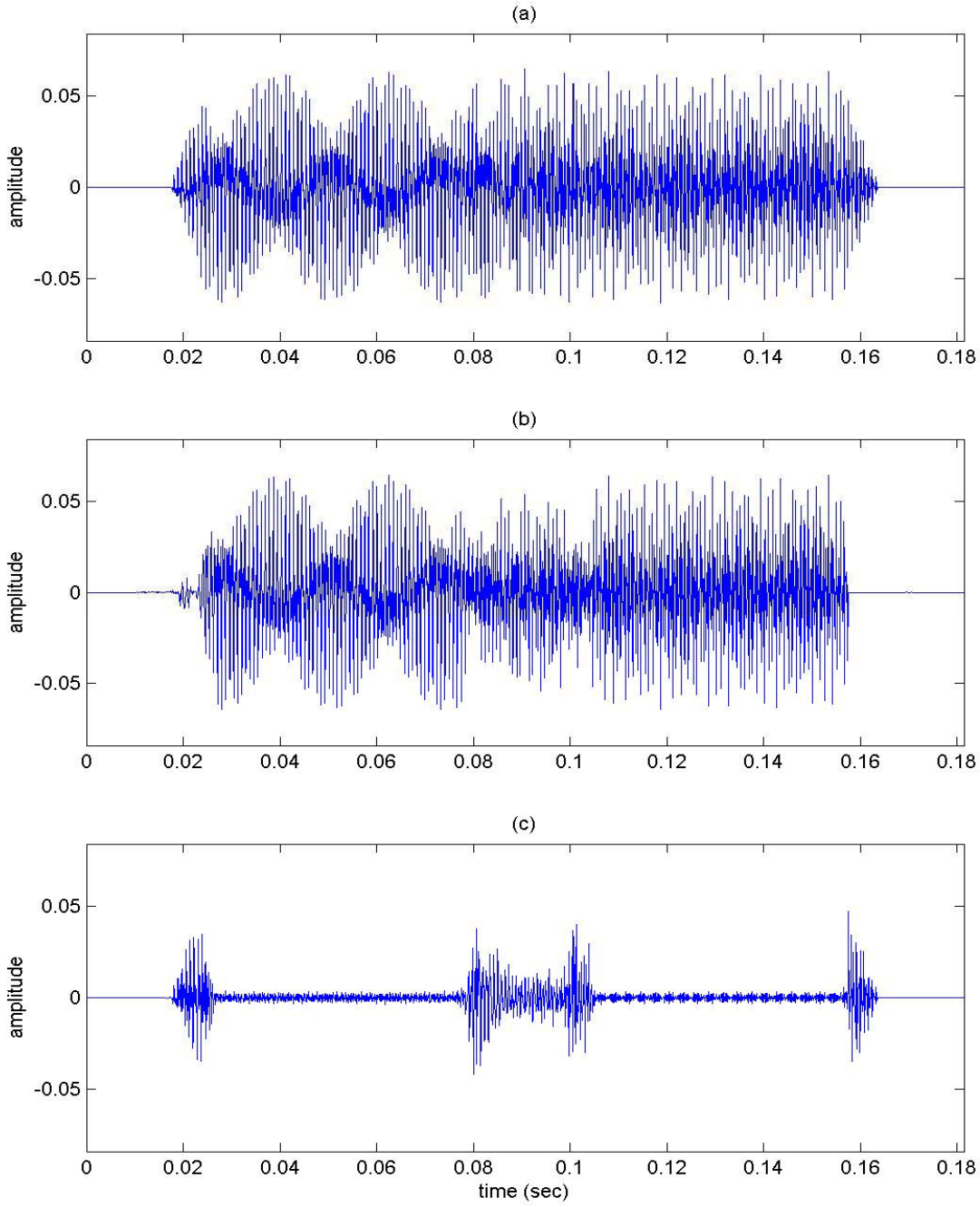

Figure C4: Waveforms of decomposed synthetic chirp signal (4th tone-chirp-tone in high frequency) : (a) original, (b) tonal, and (c) non-tonal components. All four tones+chirps+tones had 38 Hz/msec of chirp rates and the chirp durations were fixed at 20 msec.

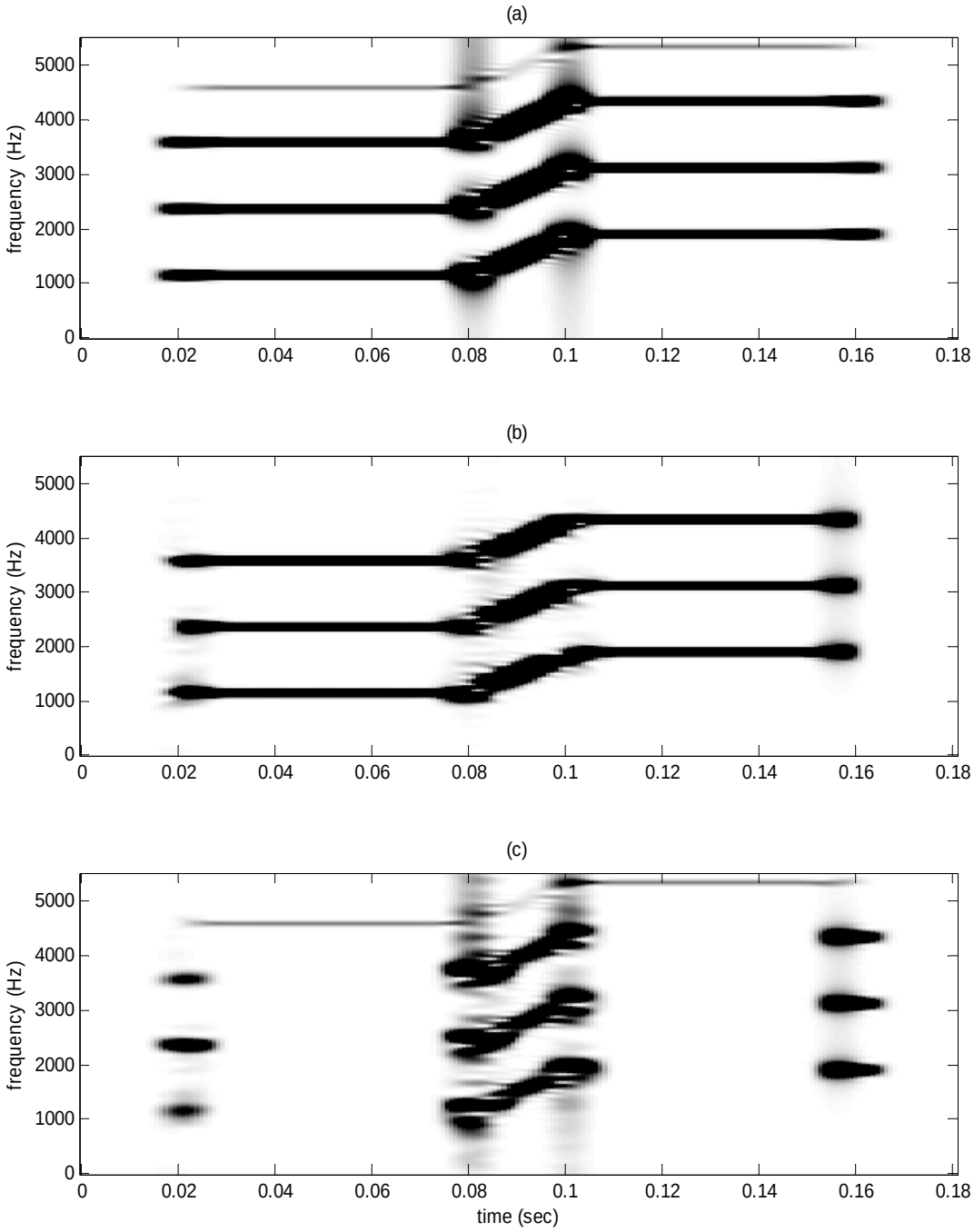

Figure C5: Spectrograms of decomposed synthetic chirp signal (4th tone-chirp-tone in high frequency): (a) original, (b) tonal, (c) non-tonal components

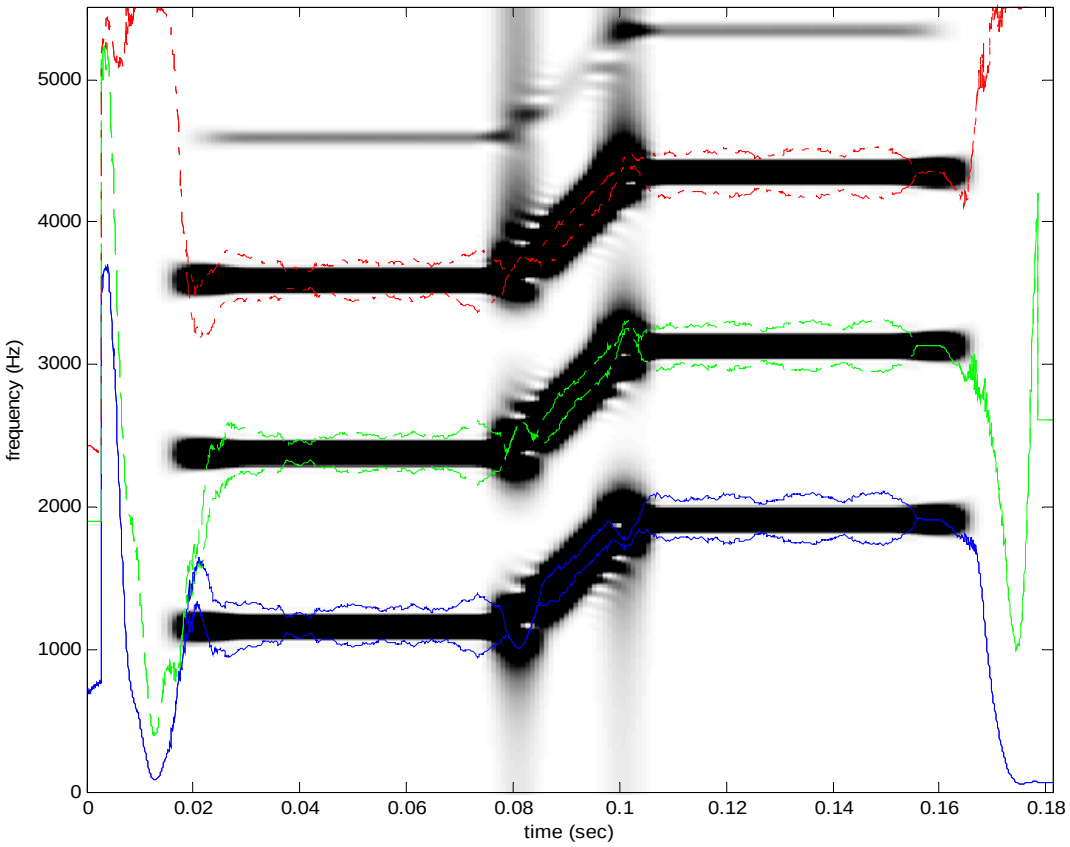

Figure C6: Upper and lower edges of time-varying bandwidths (plotted with spectrogram). 4th tone-chirp-tone in high frequency. The solid, dashed, and dotted lines are associated with the 1st, 2nd, and 3rd time-varying filters, respectively. No filter tracks 4th tone-chirp-tone component in low frequency and the tracking filter was not affected by the 4^{th} tone+chirp+tone.

APPENDIX D

THREE HUNDRED RHYMING WORDS

	1	2	3	4	5	6
1	lick	pick	tick	wick	sick	kick
2	seat	meat	beat	heat	neat	feat
3	pus	pup	pun	puff	puck	pub
4	look	hook	cook	book	took	shook
5	tip	lip	rip	dip	sip	hip
6	rate	rave	raze	race	ray	rake
7	bang	rang	sang	gang	hang	fang
8	hill	till	bill	fill	kill	will
9	mat	man	mad	mass	math	map
10	tale	pale	male	bale	gale	sale
11	sake	sale	save	same	safe	sane
12	peat	peak	peace	peas	peal	peach
13	king	kit	kill	kin	kid	kick
14	sad	sass	sag	sat	sap	sack
15	sip	sing	sick	sin	sill	sit
16	sold	told	hold	cold	gold	fold
17	buck	but	bun	bus	buff	bug
18	lake	lace	lame	lane	lay	late
19	gun	run	nun	fun	sun	bun
20	rust	dust	just	must	bust	gust
21	pan	path	pad	pass	pat	pack
22	dim	dig	dill	did	din	dip
23	wit	fit	kit	bit	sit	hit
24	din	tin	pin	sin	win	fin
25	teal	teach	team	tease	teak	tear
26	tent	bent	went	sent	rent	dent
27	sung	sup	sun	sud	sum	sub
28	red	wed	shed	bed	led	fed
29	hot	got	not	tot	lot	pot
30	dud	dub	dun	dug	dung	duck
31	pip	pit	pick	pig	pill	pin
32	seem	seethe	seep	seen	seed	seek
33	day	say	way	may	gay	pay
34	rest	best	test	nest	vest	west

35	pane	pay	pave	pale	pace	page
36	bat	bad	back	bath	ban	bass
37	cop	top	mop	pop	shop	hop
38	fig	pig	rig	dig	wig	big
39	tap	tack	tang	tab	tan	tam
40	cave	cane	came	cape	cake	case
41	game	tame	name	fame	same	came
42	oil	foil	toil	boil	soil	coil
43	fin	fit	fig	fizz	fill	fib
44	cut	cub	cuff	cuss	cud	cup
45	feel	eel	reel	heel	peel	keel
46	dark	lark	bark	park	mark	hark
47	heap	heat	heave	hear	heath	heal
48	men	then	hen	ten	pen	den
49	raw	paw	law	saw	thaw	jaw
50	bead	beat	bean	beach	beam	beak

APPENDIX E

SENSITIVITY OF THE FILTER FUNCTION FOR PSEUDO-ENHANCED SPEECH

The sensitivity of the pseudo-enhanced filter functions *F(w)* (described in section 5.3.1) to different speakers and speech materials were examined. Five filter functions were calculated. Among the five filter functions, three filter functions were calculated from the same speech material with different speakers, and two filter functions were calculated from different speakers and speech materials. Each filter function was designed to generate the long-term averaged spectrum of enhanced speech from the long-term averaged spectrum of original speech.

Fifty mono-syllable words spoken by 3 different speakers (1 female and 2 males) were used to generate the first three filter functions (same speech material with different speakers). These three filter functions are referred as $F_1(w)$ - female, $F_2(w)$ - male, and $F_3(w)$ - male. The male speaker for $F_3(w)$ was the same speaker who recorded 300 rhyming words in section 5.2.1. These 50 words were picked from the 300 rhyming words described in section 5.2.1. The other two filter functions (different speech materials and speakers – one female and one male) were calculated from forty three CVC words from the NU-6 word lists described in section 5.1.1 [36]. These two filter functions were referred as $F_4(w)$ – female and $F_5(w)$ - male.

The long-term averaged spectra of original and enhanced speech and filter function $F_1(w)$, $F_2(w)$, and $F_3(w)$ are shown in Figure E1, E2, and E3 respectively. The long-term averaged spectra of the original speech show that most spectrum energy is located at the lower

frequency region. The energies in the middle to high frequency regions were emphasized in the long-term averaged spectra of the enhanced speech. The filter functions show energy amplification in the middle to high frequency regions. The long-term averaged spectra of enhanced and pseudo-enhanced speech for the $F_1(w)$, $F_2(w)$, and $F_3(w)$ are shown in Figure E4, E5, and E6 respectively, where the solid line represents enhanced speech and the dashed line represents pseudo-enhanced speech. These two spectra show similar energy distributions at most frequencies, but for $F_2(w)$ and $F_3(w)$, magnitudes of the pseudo-enhanced speech are a few dB higher than the magnitude of the enhanced speech from 800 Hz to 1500 Hz.

The long-term averaged spectra of original and enhanced speech and filter functions for $F_4(w)$ and $F_5(w)$ are shown in Figure E7 and E8 respectively. The long-term averaged spectra of enhanced and pseudo-enhanced speech for the $F_4(w)$ and $F_5(w)$ are shown in Figure E9 and E10 respectively, where the solid line represents enhanced speech and the dashed line represents pseudo-enhanced speech.

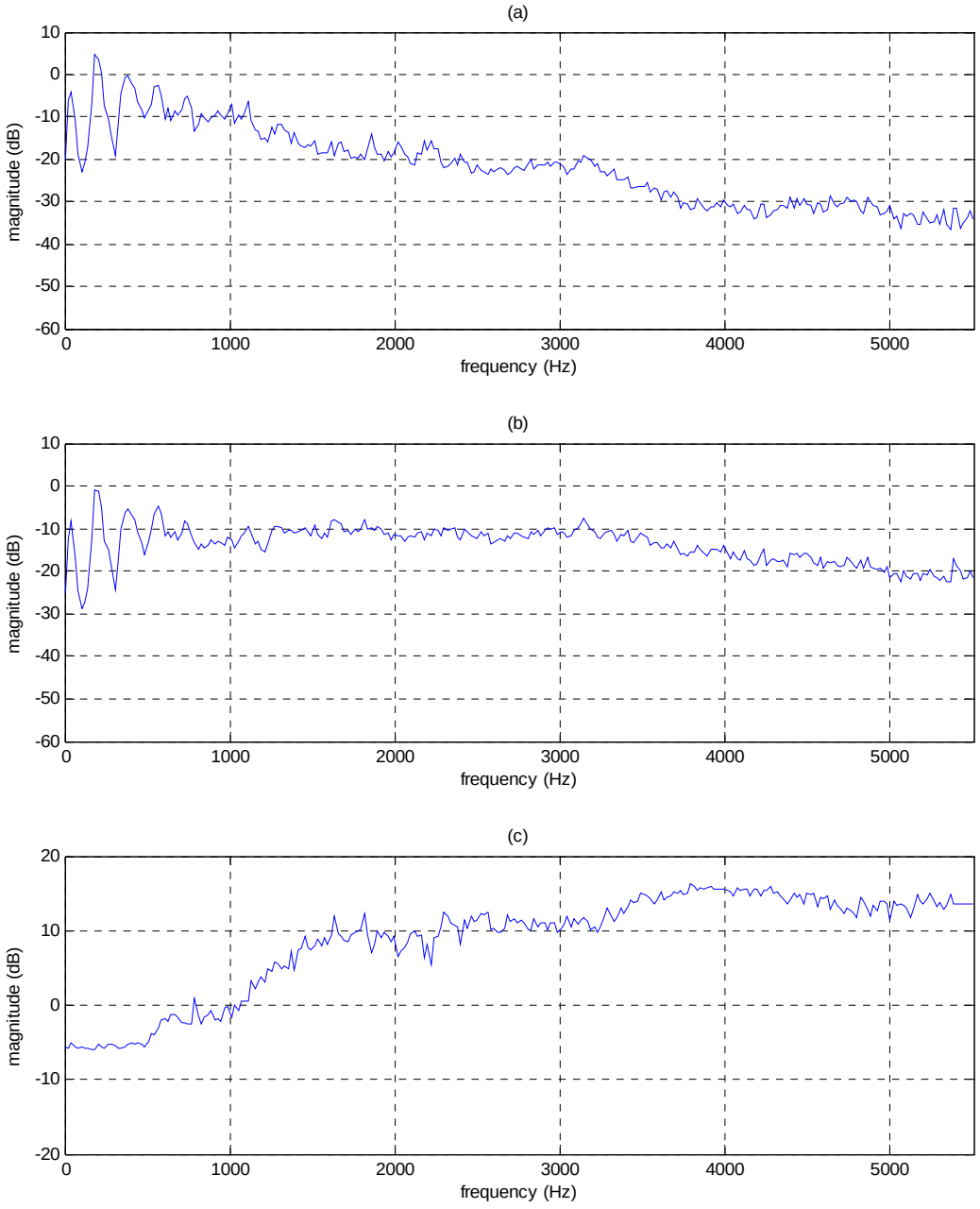

Figure E1: The long-term averaged spectra of (a) original and (b) enhanced speech for the $F_1(w)$ - female speaker and (c) the magnitude of filter function whose input and output were the long-term averaged spectra of original and enhanced speech respectively.

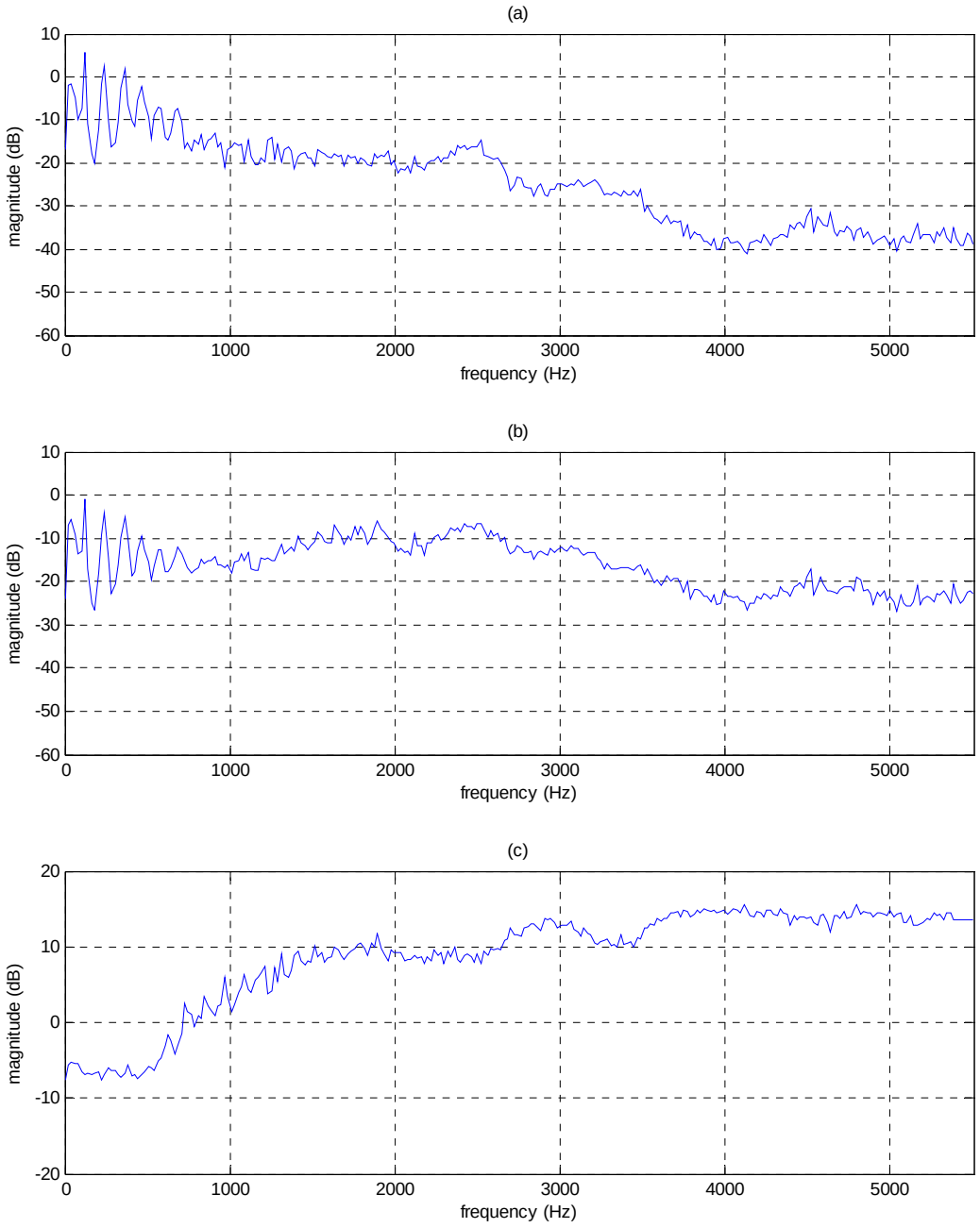

Figure E2: The long-term averaged spectra of (a) original and (b) enhanced speech for the $F_2(w)$ - male speaker and (c) the magnitude of filter function whose input and output were the long-term averaged spectra of original and enhanced speech respectively.

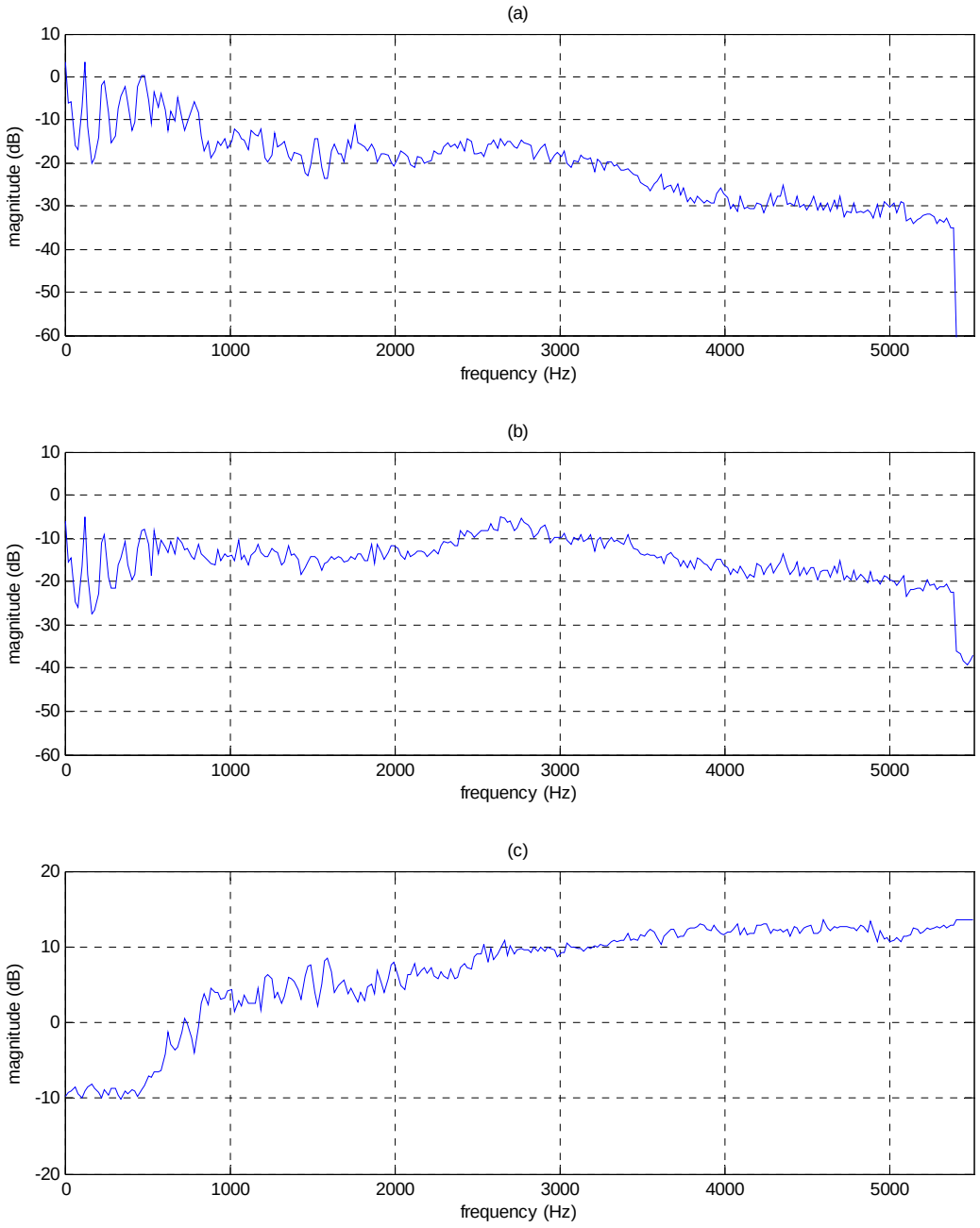

Figure E3: The long-term averaged spectra of (a) original and (b) enhanced speech for the $F_3(w)$ - male speaker and (c) the magnitude of filter function whose input and output were the long-term averaged spectra of original and enhanced speech respectively.

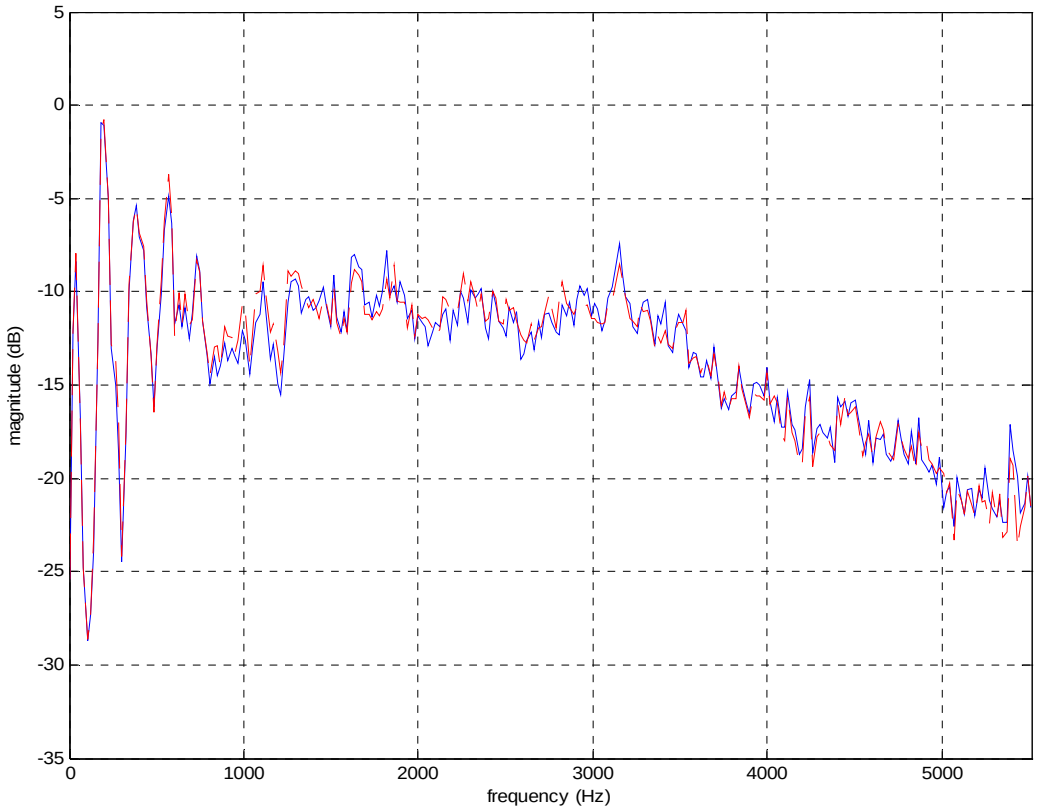

Figure E4: The long-term averaged spectra of enhanced (solid) and pseudo-enhanced (dashed) speech for the $F_1(w)$ - female speaker.

Figure E5: The long-term averaged spectra of enhanced (solid) and pseudo-enhanced (dashed) speech for the $F_2(w)$ - male speaker.

Figure E6: The long-term averaged spectra of enhanced (solid) and pseudo-enhanced (dashed) speech for the $F_3(w)$ - male speaker.

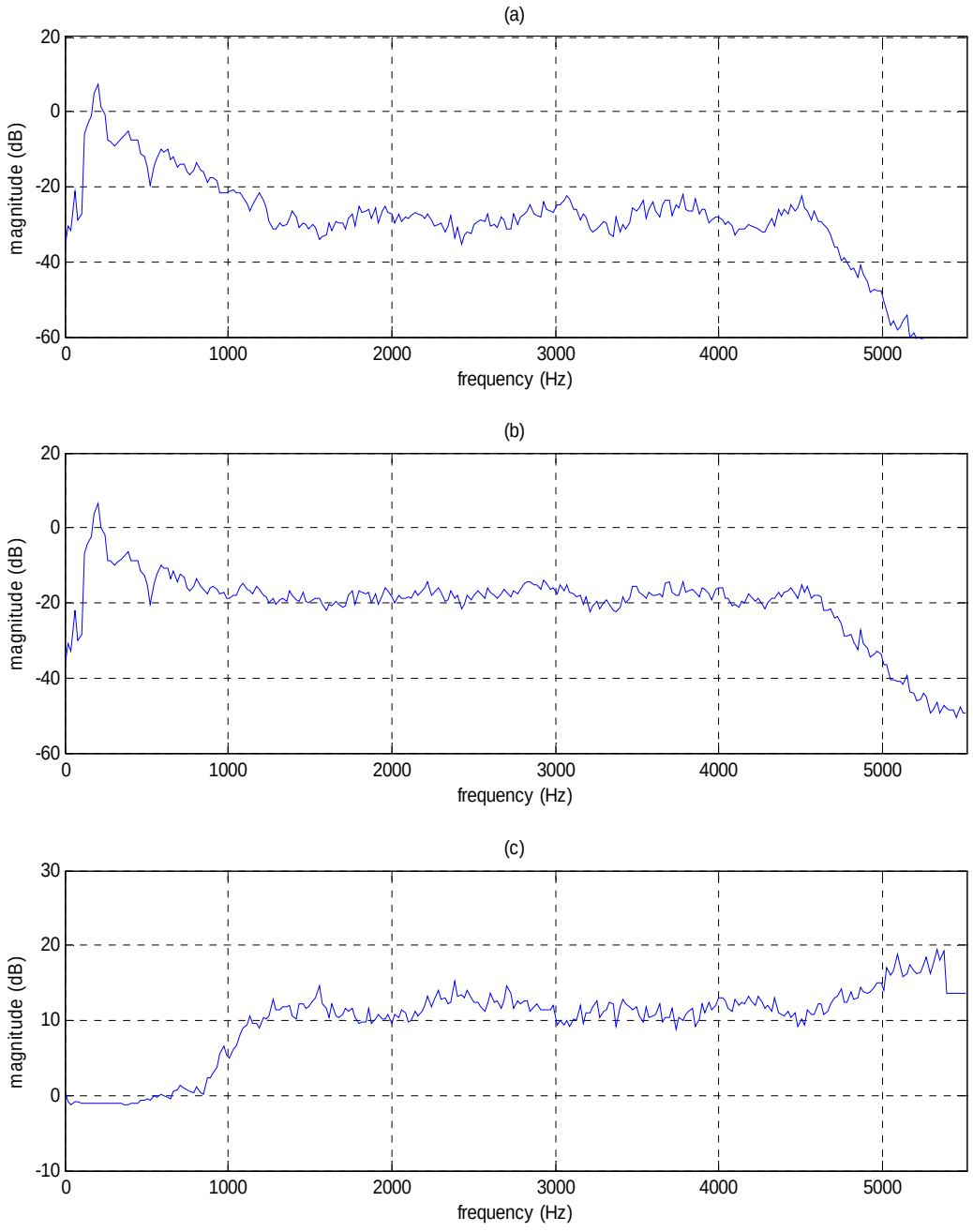

Figure E7: The long-term averaged spectra of (a) original and (b) enhanced speech for the $F_4(w)$ - female speaker and (c) the magnitude of filter function whose input and output were the long-term averaged spectra of original and enhanced speech respectively.

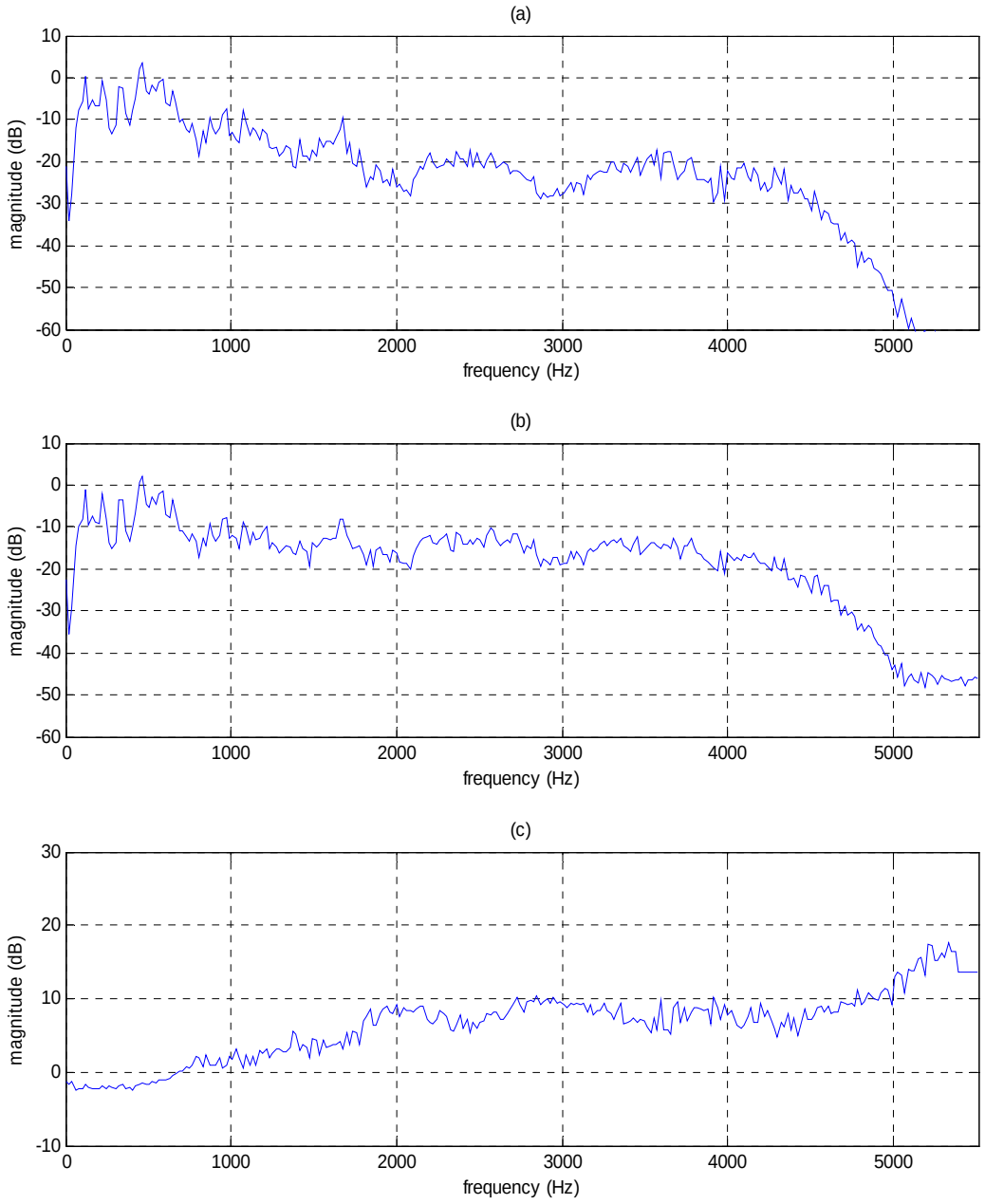

Figure E8: The long-term averaged spectra of (a) original and (b) enhanced speech for the $F_5(w)$ - male speaker and (c) the magnitude of filter function whose input and output were the long-term averaged spectra of original and enhanced speech respectively.

Figure E9: The long-term averaged spectra of enhanced (solid) and pseudo-enhanced (dashed) speech for the $F_4(w)$ - female speaker.

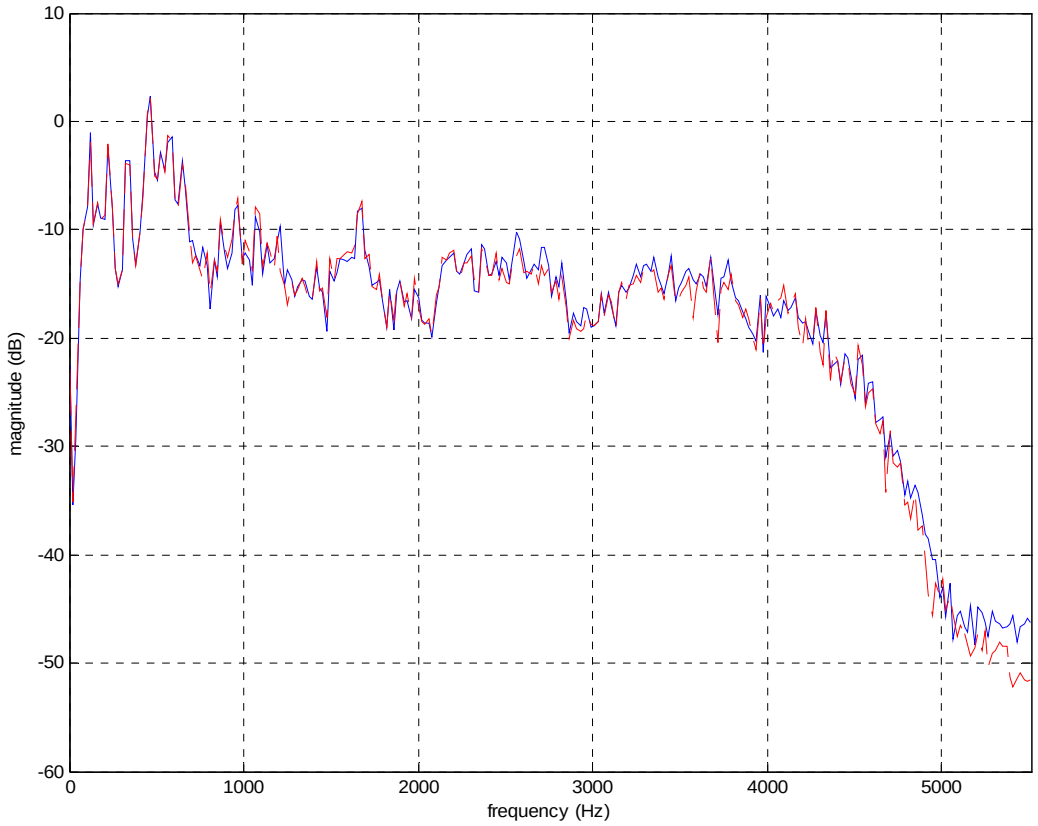

Figure E10: The long-term averaged spectra of enhanced (solid) and pseudo-enhanced (dashed) speech for the $F_5(w)$ - male speaker.

Three filter functions - $F_1(w)$, $F_2(w)$, and $F_3(w)$ – from the same speech material with different speakers are compared in Figure E11, where the dashed, solid, and dash-dotted lines represent filter functions for the $F_1(w)$ - female, $F_2(w)$ - male, and $F_3(w)$ - male respectively. The filter function generally shows energy amplification in the middle to high frequency regions. The magnitude differences between $F_1(w)$ and $F_2(w)$ are less than 5 dB across frequencies. Between $F_1(w)$ and $F_3(w)$ or $F_2(w)$ and $F_3(w)$, approximately, 10 dB differences are shown from 1500 to 2000 Hz and 5 *dB* differences are shown from 3300 to 4500 Hz.

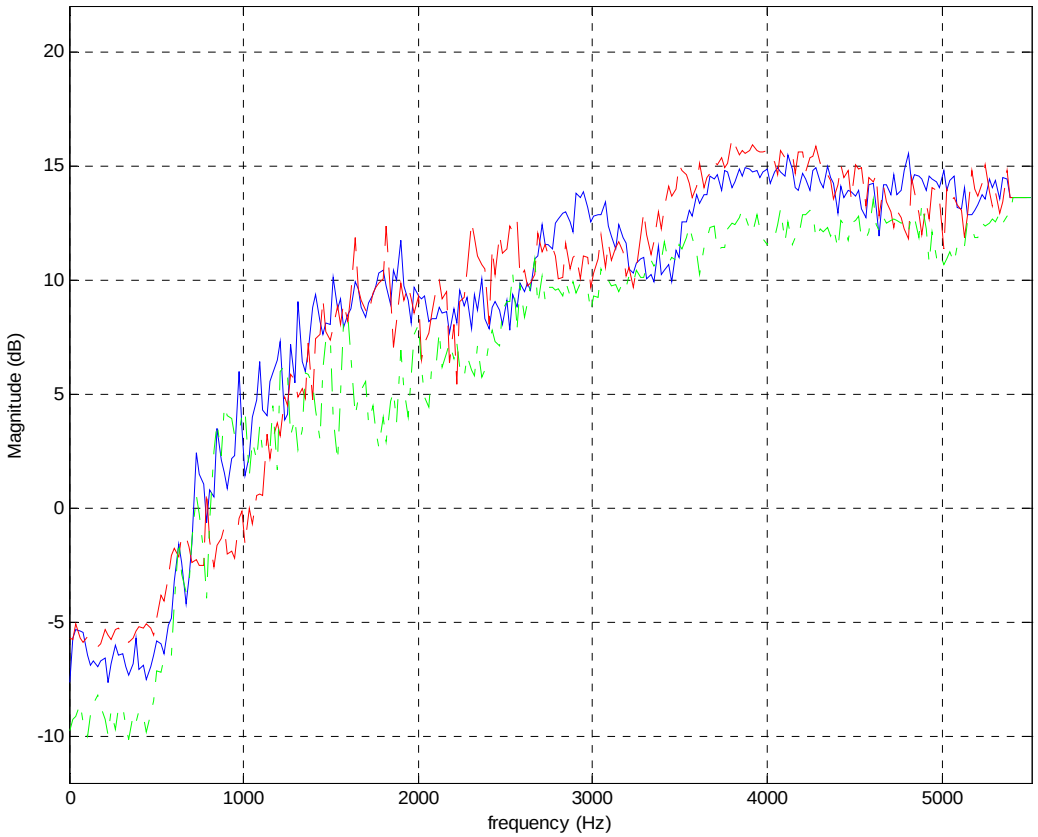

Figure E11: The magnitudes of three filter functions. Dashed, solid, and dash-dotted lines represent a filter function for the $F_1(w)$, $F_2(w)$, and $F_3(w)$ respectively.

The two filter functions ($F_4(w)$ – female and $F_5(w)$ – male) from different speech materials and different speakers are compared in Figure E12, where the solid and dashed lines represent filter functions for the $F_4(w)$ and $F_5(w)$ respectively. $F_4(w)$ and $F_5(w)$ show approximately 3-13 dB differences from 1000 to 5000 Hz.

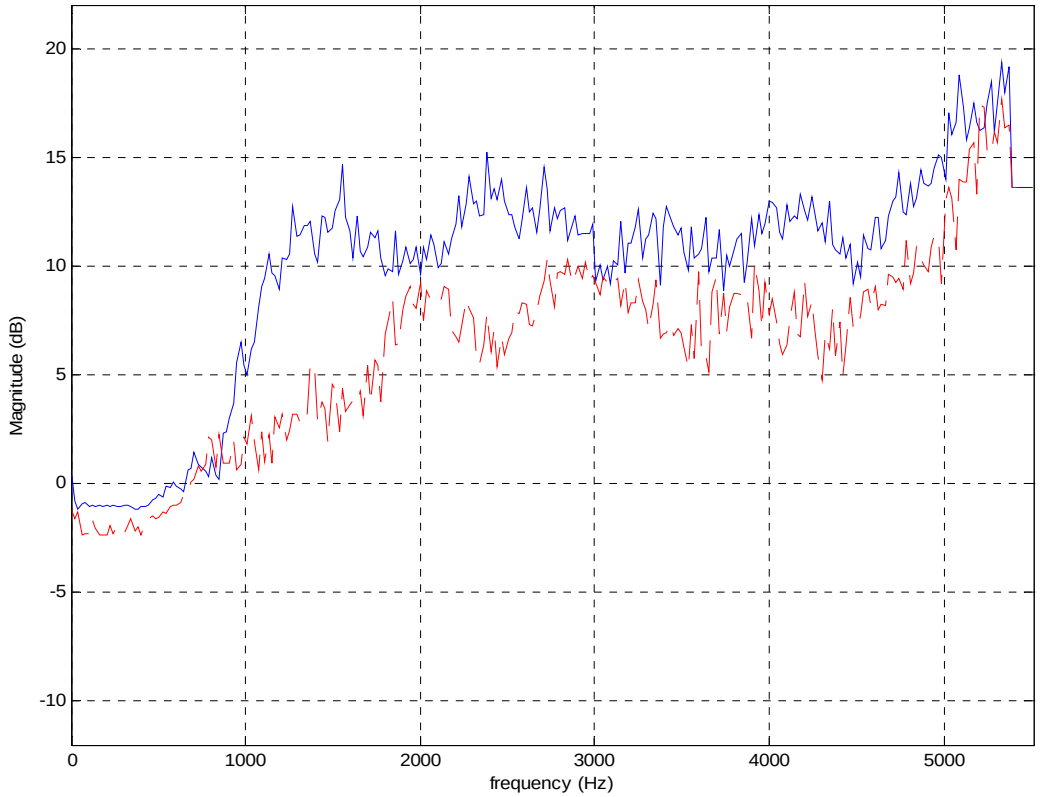

Figure E12: The magnitudes of the rest two filter functions. Solid and dashed lines represent a filter function for the $F_4(w)$ and $F_5(w)$ respectively.

All five filter functions were compared in Figure E13, where the dashed, solid (thick), dash-dotted, dotted, and solid (thin) lines represent filter functions for the $F_1(w)$, $F_2(w)$, $F_3(w)$, $F_4(w)$, and $F_5(w)$ respectively. Different filter functions show different magnitudes across frequencies. These results suggest that the filter function is sensitive to different speakers and speech materials.

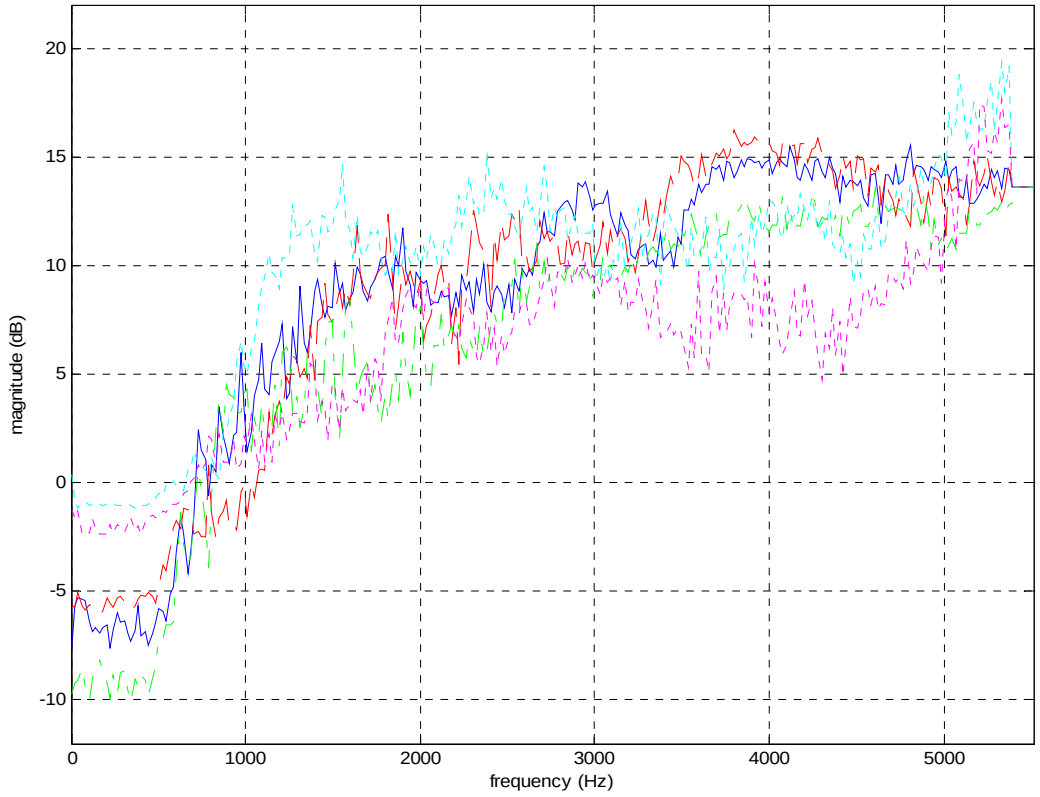

Figure E13: The magnitude of five filter functions. Dashed, solid (thick), dash-dotted, dotted, and solid (thin) lines represent a filter function for the $F_1(w)$, $F_2(w)$, $F_3(w)$, $F_4(w)$, and $F_5(w)$ respectively.

The variations in filter functions for different speakers and speech materials suggest that the fixed frequency filter calculated from the particular speech material and speaker is not robust to different speech materials and speakers. These results may imply two things. First, for the particular speaker and speech material, the fixed frequency filter may provide easier implementation and more computational efficiency than the time-varying filter does. Second, a time-frequency technique may be necessary to find an appropriate filter for a specific speaker and material because of the unreliable characteristics of the fixed frequency filter for the

different speakers and speech materials. Psychoacoustic evaluations for the enhanced and pseudo-enhanced speech across conditions will be needed to determine the validity of this conclusion.

BIBLIOGRAPHY

[1] K. Stevens, *Acoustic phonetics*, Cambridge: MIT Press, 1998.

[2] S. Gelfand, *Hearing*, New York: Marcel Dekker Inc., 1990.

[3] R. Kent and C. Read, *Acoustic analysis of speech,* Albany: Singular Thomson Learning, 2002.

[4] J. Durrant and J. Lovrinic, *Bases of hearing science,* Baltimore: Williams and Wilkins, 1995.

[5] N. French and J. Steinberg, "Factors governing the intelligibility of speech," *J. Acoust. Soc. Amer.*, vol. 19, pp. 90-114, 1947.

[6] K. Kryter, "Methods for the calculation and use of the articulation index," *J. Acoust. Soc, Am.*, vol. 34, pp. 1689-1697, 1962.

[7] H. Fletcher and R. Galt , "The perception of speech and its relation to telephony**,"** *J. Acoust. Soc, Am.*, vol. 22, pp. 89-151, 1950.

[8] ANSI S3.5-1969, "American National Standard methods for the calculation of the articulation index," *American National Standards Institute,* New York, 1969.

[9] ANSI S3.5-1997, "Methods for calculation of the speech intelligibility index," *American National Standards Institute,* New York, 1997.

[10] J. Cunningham, T. Nicol, C. King, S. Zecker, and N. Kraus, "Effects of noise and cue enhancement on neural responses to speech in auditory midbrain, thalamus, and cortex*,"* *Hearing Research*, vol. 169, pp. 97-111, 2002.

[11] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components*,"* *IEEE Trans. on Speech and Audio Processing*, vol. 6, pp. 1-11, 1998.

[12] I. Daubechies and S. Maes, *"A nonlinear squeezing of the continuous wavelet transform based on auditory nerve model,"* Wavelets in Medicine and Biology edited by A. Aldroubi and M. Unser, New York : CRC Press, pp.527-546, 1996.

[13] L. Daudet and B. Torresani, "Hybrid representations for audiophonic signal encoding," *Signal Processing*, vol. 82, pp. 1595-1617, 2002.

[14] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition," *IEEE International Conference on Acoust., Speech, and Signal Processing*, vol. 3, pp. 1783-1786, 2000.

[15] E. Yu and C. Chan, "Phase and transient modeling for harmonic+noise speech coding," *IEEE International Conference on Acoust., Speech, and Signal Processing,* vol. 3, pp. 1467 -1470, 2000.

[16] Q. Zhao, Q. Gao, and H. Chi, "Detection of spectral transition for speech perception based on time-frequency analysis," *ICICS 97*, pp. 522-525, 1997.

[17] H. Voelcker, "Toward a unified theory of modulation-part1:phase-envelope relationships," *Proc. IEEE*, vol. 54, pp. 340-354, 1966.

[18] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 240-254, 2000.

[19] T. Quatieri, T. Hanna, and G. O'Leary, "AM-FM separation using auditory-motivated filters," *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 465-480, 1997.

[20] B. Boashash and L. White, "Instantaneous frequency estimation and automatic time-varying filtering," *International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1221-1224, April 1990.

[21] A. Francos and M. Porat, "Non-stationary signal processing using time-frequency filter banks," *13th International Conference on Digital Signal Proceeding*, vol. 2, pp. 765-768, July 1997.

[22] K. Nie, G. Stickney and F. Zeng, "Encoding frequency modulation to improve cochlear implant performance in noise," *IEEE Trans. on Biomedical Engineering,* vol. 52, pp. 64-73, 2005.

[23] S. Yoo, J. R.Boston, J. D. Durrant, A. El-Jaroudi, C. C. Li. "Speech decomposition and intelligibility," *Proceedings of the World Congress on Medical Physics and Biomedical Engineering*, Sydney Australia, August 2003.

[24] J. Boston, S. Yoo, J. Durrant, K. Kovacyk, S. Karn, C. Li, and A. El-Jaroudi, "Relative intelligibility of dynamically extracted transient versus steady-state components of speech," *75th (147th) Meeting of The ASA*, May 2004.

[25] S. Yoo, J. Boston, J. Durrant, K. Kovacyk, S. Karn, S. Shaiman, A. El-Jaroudi, and C. Li, "Relative energy and intelligibility of transient speech components," *EUSIPCO*, pp. 1031-1034, Sep. 2004.

[26] S. Yoo, J. Boston, J. Durrant, K. Kovacyk, S. Karn, S. Shaiman, A. El-Jaroudi, and C. Li, "Relative energy and intelligibility of transient speech information," *ICASSP*, vol. 1, pp. 69-72, Mar. 2005.

[27] M. Li, H. McAllister, N. Black, T. De Perez, "Perceptual time-frequency subtraction algorithm for noise reduction in hearing aids," *IEEE Trans. on Biomedical Engineering,* vol. 48, pp. 979-988, 2001.

[28] B. Moore, *An introduction to the psychology of hearing,* New York: Academic Press, 2003.

[29] D. Robinson and C. Watson, *Psychophysical methods in modern psychoacoustics,* In J. Tobias (ed.), Foundation of modern auditory theory, New York: Academic Press, 1973.

[30] A. Syrdal, R. Bennett, and S. Greenspan, *Applied speech technology,* Massachusetts: CRC Press, 1995.

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

[32] 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(2), pp. 140-148, April 1999.

[33] G. Fairbanks, "Test of phonemic differentiation: The rhyme test," *J. Acoustic Society of America*, vol. 30, pp. 596-600, July 1958.

[34] R. Kumaresan and A. Rao, "Model based approach to envelope and positive instantaneous frequency estimation of signal with speech applications," *J. Acoustic Society of America*, vol. 105, pp. 1912-1924, March 1999.

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

[36] T. Tillman, R. Carhart, "An expanded test for speech discrimination utilizing CNC momosyllabic words," *Northwestern Univ. Auditory Test No 6*, Technical Report, 1966.

[37] ANSI S3.6-1996, "American National Standard specification for audiometers," *American National Standards Institute,* New York, 1996.

[38] D. Downs and M. Crum, "Processing demands during auditory learning under degraded conditions," *J. Speech and Hearing Research*, vol. 21, pp. 702-714, 1978.

[39] C. Rankovic, "Factors governing speech reception benefits of adaptive linear filtering for listeners with sensorineural hearing loss," *J. Acoust. Soc, Am.*, vol. 103, pp. 1043-1057, 1998.

[40] K. Kryter, "Validation of the articulation index," *J. Acoust. Soc, Am.*, vol. 34, pp. 1698-1702, 1962.

[41] E. Keller, *Fundamentals of speech synthesis and speech recognition : basic concepts, state of the art and future challenges*, New York: Wiley, 1994.

[42] A. Waibel and K. Lee, *Readings in speech recognition*, San Mateo: Morgan Kaufmann Publishers, 1990.

[43] L. Rabiner and B. Juang, *Fundamentals of speech recognition*, New Jersey: Prentice Hall, 1993.

[44] C. Lee, F. Soong, and K. Paliwal, *Automatic speech and speaker recognition*, Massachusetts: Kluwer Academic Publishers, 1996.

[45] F. Jelinek, *Statistical methods for speech recognition*, Massachusetts: MIT Press, 1997.

[46] *BBN Byblos version 2.0 summer 2001 delivery manual*, BBN Systems and Technologies, 2001.

[47] P. Dognin, "A bandpass transform for speaker normalization," Ph.D dissertation, University of Pittsburgh, 2003.