PROCESSING PERCEPTUALLY IMPORTANT TEMPORAL AND SPECTRAL CHARACTERISTICS OF SPEECH

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the

Graduate School of The Ohio State University

By

Eric W. Tarr, B.S., B.A., M.S.

Graduate Program in Electrical and Computer Engineering

The Ohio State University

2013

Dissertation Committee:

Dr. Ashok Krishnamurthy, Co-Adviser

Dr. Susan Nittrouer, Co-Adviser

Dr. Bradley Clymer

Dr. Hooshang Hemami

ABSTRACT

Speech signals are time-varying in frequency, with perceptually important characteristics spanning both temporal and spectral dimensions. Determining perceptual importance requires isolating and manipulating signal characteristics for inspection in listening experiments. Pitch synchronous signal processing methods are presented to represent temporal and spectral characteristics independently. Ways to potentially enhance the perception of speech by processing the signal's spectral characteristics are discussed. These methods were applied to listening experiments to examine the perceptual importance of a signal's temporal envelope and spectral envelope. Pitch synchronous processed speech signals were perceived as sounding natural, and also maintained a similar level of intelligibility to unprocessed speech signals. This finding was an improvement compared with similar signal processing methods. Additionally, the perceptual segregation of speech in the presence of noise and the perceptual integration of dissimilar signals were examined.

To my wife, family, and friends

ACKNOWLEDGMENTS

I would like to thank the my co-advisors, Dr. Ashok Krishnamurthy and Dr. Susan Nittrouer for their instruction throughout my graduate education. I would like to thank Dr. Bradley Clymer and Dr. Hooshang Hemami for their advise and encouragement. Also, thanks to the Ohio State Department of Otolaryngology and the National Institute of Health for their support of this research effort.

VITA

August 22, 1984	Born - Minneapolis, Minnesota, USA
2008	B.S. Electrical and Computer Engineering, The Ohio State University
2008	B.A. Mathematics, Capital University
2010	M.S. Electrical and Computer Engineering, The Ohio State University
2013	Ph.D. Electrical and Computer Engineering, The Ohio State University
2008-present	Graduate Research Associate, The Ohio State University.

PUBLICATIONS

Research Publications

E. Tarr, A. Krishnamurthy, S. Nittrouer, "A pitch synchronous framework for representing, recovering, removing, and replacing acoustic characteristics of speech signals". *Journal of the Acoustical Society of America, Submitted May 2013.*

E. Tarr, S. Nittrouer, "Explaining coherence in coherence masking protection for adults and children". *Journal of the Acoustical Society of America, 133, 2013.*

S. Nittrouer, J. Lowenstein, E. Tarr, "Amplitude rise time does not cue the /ba/-/wa/ contrast for adults or children" *Journal of Speech, Language, and Hearing Research, 2013.*

J. Lowenstein, S. Nittrouer, E. Tarr, "Children weight dynamic spectral structure more than adults: Evidence from equivalent signals" *Journal of the Acoustical Society of America, 132, EL443-EL449, 2012.*

S. Nittrouer, A. Caldwell, J. Lowenstein, E. Tarr, C. Holloman, "Emergent literacy skills in kindergartners with cochlear implants" *Ear & Hearing, 33, 683-697, 2012.*

E. Tarr, S. Nittrouer, "Coherence masking protection for mid-frequency formants". *Journal of the Acoustical Society of America, 130, EL290-EL296, 2011.*

S. Nittrouer, E. Tarr, "Coherence masking protection for speech signals in children and adults". *Attention, Perception, and Psychophysics, 73, 2606-2623, 2011.*

FIELDS OF STUDY

Studies in:

Speech Processing	Prof. Ashok Krishnamurthy
Speech Perception	Prof. Susan Nittrouer
Signal Processing	

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction to the Field of Speech Processing

1.1 Introduction

Human speech is an important signal. It is the primary method of human interaction and relationship. This signal exists in a system of speech communication between speakers and listeners, with sub-systems of human production and perception. The human speech production and auditory perception systems are highly sophisticated and specialized. In many ways, these systems are complementary and well matched to function in communication. An additional aspect of communication is the system of transmission. The acoustic environments in which speech occurs are diverse, and influence communication.

The entire system of speech communication is quite exquisite. However, it has limitations. The production of speech signals is complicated, and varies in many different ways depending on an individual speaker. Acoustic environments introduce interfering signals to the transmitted speech. After all this, the system of auditory perception has to analyze the complicated signal, which varies from one speaker to another, and which has been interfered by additional background signals. Additionally, auditory perception has its own limitations and can even function improperly for some individuals.

In order to broaden the access of speech communication, technology has been developed specifically for the human voice. Communication systems to transmit and receive speech include radio, telephones, and VoIP. Medical devices to amplify the sound of the human voice include hearing aids and cochlear implants. Audio and music technology has been developed to capture, store, and reproduce the human voice, as well as make it sound more pleasing. Essentially, this technology has been developed to make communication possible, improve communication outcomes, and increase communication satisfaction.

Given these purposes of speech technology, the problem becomes determining how to process the speech signal to achieve a specific purpose. There are a wide variety of possibilities ranging from speech-specific signal processing methods to general signal processing methods. One potential way of selecting an appropriate signal processing method is to understand which characteristics of the signal that are important to perception. Then, the problem can be focused on attempting to preserve these characteristics, and next, considering if these characteristics can be enhanced.

Speech signals are time-varying in frequency, with characteristics spanning both temporal and spectral dimensions. One aspect of speech research has been determining how the auditory system analyzes these various characteristics as acoustic information and translates them into phonetic information. The conventional approach to this research has been to isolate and manipulate a specific characteristic to determine its contribution to speech perception. In some cases, synthetic speech is used to allow for precise control of characteristics prior to the signal being produced. In other cases, natural speech is used, but requires additional processing the isolate and manipulate the specific characteristic.

1.2 Related Work

Before discussing the signal processing methods and perceptual experiments in the remaining chapters, it is appropriate to discuss related work completed previously.

1.2.1 Speech Processing Research of Temporal and Spectral Characteristics

Speech signal processing has been used to investigate the perceptual importance of certain acoustic characteristics in a signal. Examples of this are signal processing techniques used to represent characteristics of temporal information or spectral information.

Processing Temporal Information

A general method of uniquely defining amplitude in a signal over time is based on the analytic version of a signal. By the Fourier transform, any real-valued signal, $x(t)$ has energy in both positive and negative frequencies. An analytic signal, $x_a(t)$ is a complex-valued version of a real-valued signal that has the identical energy for positive frequencies, but no energy for negative frequencies. The Hilbert Transform of a real-valued signal, $\mathcal{H}[x(t)]$, is the signal that produces the analytic signal, $x_a(t) = x(t) + j\mathcal{H}[x(t)]$.

The analytic signal can be used to estimate the temporal information of a signal called the temporal envelope, $a(t) = |x(t) + j\mathcal{H}[x(t)]|$. The instantaneous phase of a signal can also be estimated from the analytic signal, $\phi(t) = \arctan(\frac{\mathcal{H}[x(t)]}{x(t)})$. The original signal can be reconstructed from the temporal envelope and the instantaneous phase signal, $\tilde{x}(t) = a(t)\cos(\phi(t))$.

The Hilbert Transform has been used in speech coding for telecommunication systems based on perceptual encoding in the human auditory system [Flanagan and Golden(1966), Flanagan(1980)]. Here, it was suggested that a speech signal could be represented by passing the signal through a parallel bank of contiguous band-pass filters. Each channel could then be decomposed into a temporal envelope and instantaneous frequency signal using the Hilbert Transform. This speech coding technique, called auditory chimera processing, was used in a listening experiment to examine the perceptual importance of the temporal envelope and the instantaneous phase signal to speech recognition [Smith, Delgutte, and Oxenham(2002)].

Processing Spectral Information

Along with representing the temporal information of speech, signal processing has been used to represent different aspects of spectral information in speech. Linear predictive coding (LPC) is one such processing technique. Similar to recovering the temporal envelope and instantaneous phase signal using the Hilbert Transform, LPC recovers the spectral envelope and residual signal from speech. The original signal can be reconstructed from the spectral envelope and residual signal [O'Shaughnessy(1987), Childers(1987)]. Additionally, several methods have been proposed to change the shape of the spectral envelope independent of the residual signal [Makhoul(1976), Parker and Hall(1979), Kuwabara(1984)].

1.2.2 Speech Perception Research of Temporal and Spectral Characteristics

Speech perception has been traditionally studied by asking how specific cues support the recovery of phonetic structure. Synthetic syllables are used to experimentally

control for the specific cue. In these syllables, all but one acoustic elements are held constant across a series of stimuli at settings providing ambiguous information about phonetic identity. That one signal component, the cue, is manipulated across the series, spanning a range from a setting that disambiguates labeling in favor of one phonetic category to a setting that disambiguates labeling in favor of another phonetic category. All stimuli are presented to listeners multiple times for phonetic labeling, and a labeling function is derived from listeners' responses to represent categories of perception. This method of investigation has been used to explain how variation in small characteristics of the signal, known as acoustic cues, specifies phonetic categories. However, research with this methodology has not been able to explain all aspects of human speech perception.

Other research methods have also been used to measure the perceptual importance of temporal and spectral acoustic characteristics. Rather than manipulating a specific cue, the acoustic characteristic of interest can be removed from a speech signal, leaving a residual signal to be analyzed. If recognition is decreased, then the acoustic characteristic is considered to contribute to perception. In another method, the acoustic characteristic of interest is recovered from a speech signal, independent of the other characteristics in the residual signal. A new signal is synthesized based only on the acoustic characteristic of interest, lacking the remaining characteristics of the residual signal. If recognition persists after removing all other characteristics, then the acoustic characteristic is considered to contribute to perception. Examples of acoustic characteristics that have been investigated are the temporal envelope and the spectral envelope of speech.

Perceiving Temporal Information

One aspect of early speech research investigated the perceptual importance of certain acoustic characteristics in a signal. The purpose of the research was to efficiently transmit the important characteristics of speech through telecommunication systems. In one study, Licklider and Pollack (1948) were interested in the importance of the signal's fluctuations in amplitude over time. Their method involved removing a signal's fluctuations in amplitude and measuring the impact on speech recognition. They found that infinite peak clipping amplitude distortion impaired intelligibility by a "surprisingly little" amount. It has been argued that this result indicates that temporal amplitude information is unimportant for speech.

This conclusion was challenged several years later in Shannon et al. (1995). Rather than attempting to remove the temporal amplitude information in a signal, this study measured recognition for speech represented primarily by temporal amplitude information, while removing much of the other information. The results of Shannon et al. (1995) indicated that near-perfect speech recognition could be obtained using temporal amplitude information, even if the spectral information is greatly reduced. It was then concluded that the temporal amplitude information may be important after all, especially in some applications. Shannon et al. (1995) specifically suggested one application of this result would be the development of alternative signal processing strategies for auditory prostheses.

One aspect of the research in Shannon et al. (1995) was to investigate whether slow-varying temporal amplitude information contributed to speech recognition differently than fast-varying temporal amplitude information. The initial finding was

that the slow-varying temporal amplitude information was sufficient for speech perception. However, the perceptual importance of fast-varying temporal amplitude information has been proven. It has been shown that this information is important to speech perception in the presence of noise [Hopkins and Moore(2009)]. Specifically, sensitivity to this information has been shown to be important for listening "in the dips" when the interfering noise is fluctuating [Moore(2008)]. Furthermore, it has be shown that listeners with impaired hearing are unable to make use of the fast-varying temporal amplitude information in a speech signal [Lorenzi, Gilbert, Carn, Garnier, and Moore(2006), Hopkins and Moore(2007)]. This has been used to explain why listeners with impaired hearing have difficulty understanding speech in the presence of background sounds, especially when the interference is fluctuating [Duquesnoy(1983)]. The relationship between slow- and fast-varying temporal amplitude information as an acoustic cue and various linguistic cues was presented in Rosen (1992).

Perceiving Spectral Information

Complementing the research on the perceptual importance of the temporal amplitude information, research has been conducted on the perceptual importance of the spectral information in speech signals. Just as Licklider and Pollack (1948) investigated speech recognition after temporal amplitude information was removed from a speech signal, research has been conducted to investigate speech recognition after spectral information has been removed from a speech signal. Sine-wave speech (SWS) has been proposed as a method to remove much of the spectral information of speech. SWS is synthesized by estimating frequencies of formants and creating time-varying tones at these frequencies to replace the original signal [Remez et al.(1981)]. Near-perfect speech recognition has been achieved with SWS given sufficient training

and sentence context [Barker and Cooke(1999),Nittrouer and Lowenstein(2010)], suggesting much of the spectral information of speech is unimportant. This conclusion contradicted a theory of perceptual organization that argued listeners perceptually process complex signals such as speech based on spectral details that had been removed in SWS.

Auditory Scene Analysis Auditory scene analysis has been proposed as a theory to explain the method in which listeners analyze the sensory information of complex and interfering signals [Bregman(1990)]. In this hypothesis, the organization of sensory information is based on segregating signals into perceptual streams. Signals that share temporal and spectral characteristics are integrated for further perceptual analysis. The Gestalt Principles of similarity, proximity, and continuity were adapted to auditory perception as temporal common fate and spectral harmonicity. Conversely, interfering signals can also be segregated into separate perceptual streams if they do not share temporal and spectral characteristics.

It is difficult for the auditory system to appropriately segregate all interfering signals. The consequence of this is perceptual masking, of which there are several types. Energetic masking refers to interference that occurs in the periphery of the auditory system by the presence of another physical signal. Informational masking refers to perceptual interference that occurs more central for reasons not associated with the presence of other physical signals [Ihlefeld and Shinn-Cunningham(2008)]. Although masking can occur for signals in spectral proximity, the auditory system can overcome masking for spectrally distant signals outside a perceptual critical band [Zwicker and Terhardt(1980)].

Several experiments have been conducted to examine ASA principles of perceptual integration and segregation for spectrally proximate and spectrally distant signals. In these experiments two spectrally distant components of synthetic speech were presented with masking noise in a labeling task. If these components were perceived separately, speech recognition was poorer than if these components were perceptually integrated [Gordon(1997)]. A separate series of experiments has replicated the results for adult listeners, and also examined child listeners [Nittrouer and Tarr(2011), Tarr and Nittrouer(2011), Tarr and Nittrouer(2013)]. Based on a sequence of experiments designed to examine ASA principles of perceptual organization, it was demonstrated that the principle of harmonicity is not necessary, and the principle of common fate is not sufficient to perceptually segregate speech in the presence of masking noise. Rather, an alternative theory must explain perceptual organization of speech signals. These experiments will be discussed in Chapter 5.

Speech Perceived through Auditory Prostheses

Auditory prostheses have improved the perception of sound for listeners with hearing loss. Hearing aids can be used to amplify acoustic vibrations when acoustic sensitivity is mildly or moderately diminished. For severe to profound hearing loss, cochlear implants are commonly used and provide the transduction of acoustic vibration to electrical nerve impulse by sending electrical currents to the auditory nerve. A review of the history and technology of cochlear implants is provided in [Wilson and Dorman(2008a)] and [Wilson and Dorman(2008a)]. The functioning of a cochlear implant constrains how acoustic signals are electrically encoded. Encoding the characteristics of speech signals that are perceptually important for language is one aspect of improving results for listeners with auditory prostheses.

To investigate how auditory prostheses can process speech most effectively, simulations and modeling of auditory prostheses have been used to examine listeners with normal hearing. Simulating the conditions of auditory prostheses for listeners with normal hearing has several experimental benefits. The recipients of auditory prostheses are different in the degree of their hearing loss and their devices are custom fitted. Therefore, it is difficult to experimentally control aspects that influence perception across listeners with auditory prostheses. Simulation using listeners with normal hearing ensures each participant is presented the same signals. Pragmatically, a large sample size of listeners with normal hearing is more readily available to participate than listeners with the desired auditory prostheses. Additionally, simulation allows for comparisons of multiple conditions within subjects rather than using separate listening groups for each type of auditory prostheses. Shannon et al. (1995) used multi-band vocoding to simulate cochlear implants and examine speech recognition. However, there are many other types of auditory prostheses currently in use. Future advances in technology for auditory prostheses should continue to be simulated and modeled for listeners with normal hearing.

A variety of auditory prosthesis options are available for listeners with impaired hearing. Listeners may be fitted with a single, unilateral, hearing aid or two, bilateral, hearing aids. Other listeners may be implanted with a unilateral cochlear implant or bilateral cochlear implants. Additionally, listeners with a unilateral cochlear implant may wear a hearing aid on their contralateral ear. Perceptual research has demonstrated that this combination of electric and acoustic stimulation can be beneficial [Dorman and Gifford(2010)]. Advances in implant electrodes and surgical procedures for auditory implants have made it possible to preserve residual hearing in an

implanted ear that would have previously been lost as a consequence of implantation. The preservation of residual hearing makes electric and acoustic stimulation in the same ear possible. An experiment will be presented in Chapter 6 that investigates the perceptual organization of stimuli simulating acoustic and electric stimulation.

CHAPTER 2

Pitch Synchronous Decomposition of Acoustic Characteristics of Speech

2.1 Introduction

As part of the effort to investigate speech communication, various acoustic characteristics of the speech signal have been identified. The contribution of these acoustic characteristics to communication has been used to explain perception. The method to investigate the contribution of different acoustic characteristics has involved using signal processing to isolate each component for inspection. There are many techniques that can be used to separate different information related to acoustic characteristics of speech. Applying various techniques can be used to understand different aspects of perception including: organization, weighting, cognitive processing and linguistic categories.

One way to decompose many naturally produced signals is to represent the signal as a low frequency modulator which modulates the amplitude of higher frequency carriers. A general relationship between signal amplitude $A[t]$ and phase $\phi[t]$ (which can be used to derive signal instantaneous frequency), $x[t] = \mathbb{R}\{A[t]e^{j\phi[t]}\}$ was presented in Loughlin and Tacer (1996). There are an unlimited number of $A[t]$ and

$\phi[t]$ pairs which will satisfy $x[t] = \mathbb{R}\{A[t]e^{j\phi[t]}\}$. One method to calculate a pair that satisfies the general mathematical relationship is based on the analytic signal. Using the Hilbert Transform (HT), $\mathcal{H}(x[t])$, the amplitude envelope is defined as $A[t] = |x[t] + j\mathcal{H}(x[t])|$. The phase of a signal can also be estimated with the HT, $\phi[t] = \arctan(\frac{\mathcal{H}(x[t])}{x[t]})$. This method is not the only method to define signal amplitude and phase, but it has the advantage that it can be used for any signal, whether or not it is speech. Conceptually, separating the low-frequency modulator from the carrier signal is a way the decompose the temporal envelope of the signal.

Conversely, it may also be useful to decompose the spectral envelope from the signal. This can be accomplished by representing many naturally produced signals by an excitation source signal that is filtered in frequency by a resonant chamber or cavity. A signal $x[t]$ could be 'deconvolved' into an excitation source $u[t]$ and spectral filter $h[t]$. Spectrally, these components are related by: $X[z] = U[z]H[z]$. This is a general relationship between components that can be calculated in different ways. O'Shaughnessy (1987) presented one method of decomposing speech signals in this way, called linear predictive coding (LPC). In this method an all-pole filter can be used to model the spectral filter. Rabiner (1997) suggested removing the spectral envelope by using spectral flattening techniques. The possibility of other methods remain and will be considered.

Speech signals have been represented both as a carrier signal modulated in amplitude and as a source signal filtered in frequency based on speech production. In one representation of speech production, the opening and closing of the human vocal tract (low frequency modulator) changes the amplitude over time of the signal produced by vocal fold vibrations (carrier signal) that exit the mouth. In another representation

of speech production, the resonances of the vocal tract created by the configuration of the mouth and tongue (spectral filter) change the amplitude of different frequencies present in the signal produced by vocal fold vibrations (source signal).

2.1.1 Speech Perception Research

Speech perception research has investigated the perceptual contribution of the temporal envelope and the spectral envelope. Listeners with normal hearing have been compared to listeners with perceptual deficits. These perceptual deficits have included sensorineural hearing loss, dyslexia, and auditory neuropathy.

Investigating the Temporal Envelope

The low-frequency modulator has been represented by a temporal envelope in or order to investigate speech perception. The envelope can 'demodulated' from the speech signal, with the residual signal representing the high-frequency carrier. Licklider and Pollack (1948)proposed a method to remove the temporal envelope by using infinite-peak clipping. A different method to isolate the envelope characteristics of speech was presented in Horii et al. (1971). In this method, the original carrier signal was removed and the temporal envelope was used to modulate a noise carrier. Full-wave rectification was used with low-pass filtering to recover the temporal envelope. An extension of this technique was presented by Shannon et al. (1995) to recover the temporal envelope in each channel of a filter bank. Each envelope was then used to modulated band-limited noise filtered for each channel of the filter bank. Smith et al. (2002) presented a technique to recover both the temporal envelope and carrier signal, rather than substituting a noise carrier. This technique called 'auditory chimeras' could be used to combine the temporal envelope from one speech utterance

with the 'demodulated' carrier from a different utterance. In this method, the temporal envelope and carrier signals were recovered from the HT amplitude $A[t]$ and phase $\phi[t]$. Fig. 2.1 shows a diagram of the signal processing used for auditory chimera processing including the filterbank and HT.

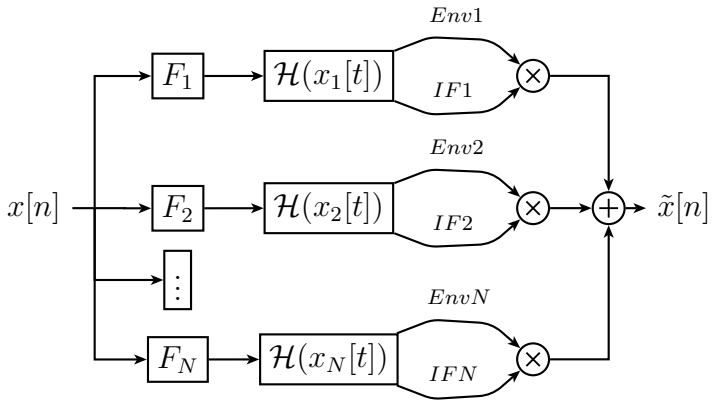

Figure 2.1: Diagram of Auditory Chimera Processing

Listeners with Normal Hearing

Several studies have used these methods to investigate the perceptual importance of the temporal envelope. Smith et al. (2002) compared the perceptual importance of the temporal envelope to the carrier signal. Their conclusion was that for speech stimuli, the temporal envelope information perceptually dominated the carrier information. For music signals, the carrier information perceptually dominated the temporal envelope. Shannon et al. (1995) argued that the temporal envelope was sufficient for speech perception and found near perfect speech recognition for signals represented with primarily temporal envelope cues. This finding went against the conventional argument that the temporal envelope information is unimportant

for speech. Licklider and Pollack (1948) found that removing the temporal envelope from a signal using infinte peak clipping impaired intelligibility by a "suprisingly little" amount.

Although the temporal envelope was found in some cases to be important perceptually for English-speaking listeners, a separate experiment showed that listeners who spoke a tonal language perceptually relied on the the residual carrier signal rather than the temporal envelope [Xu and Pfingst(2003)]. The residual carrier signal was also found to be important perceptually for speech in the presence of background noise [Moore(2008)]. In particular, it was found that the residual carrier signal contributes to a perceptual phenomenon known as "listening in the dips." Separately, it was shown that listeners with impaired hearing have a reduced ability to use the information in the residual carrier signal. This likely contributes to the perceptual difficulty of listening in noise for listeners with hearing loss [Lorenzi, Gilbert, Carn, Garnier, and Moore(2006)].

Listeners with Impaired Hearing

A reason to investigate how speech can be represented with temporal structure is that listeners with impaired hearing have been shown to have strong temporal resolution even in the face of poor spectral resolution. Psychoacoustic [Florentine and Buus(1984), Bacon and Gleitman(1992), Dwyer and Nelson(1992)] and speech perception experiments [Turner, Souza, and Forget(1995)], have demonstrated that moderate to severe sensorineural hearing loss does not impair the temporal acuity of listeners. Results have indicated that temporal resolution is not determined at the level of the cochlea and auditory nerve, but is limited by sites more central in the auditory system [Relkin and Turner(1988), Turner, Relkin, and Doucet(1994)]. It was

argued that "there is no compelling reason to suspect that cochlear damage associated with sensorineural hearing loss will automatically result in reduced temporal acuity" [Turner, Souza, and Forget(1995)]. In fact, if audibility is compensated for, temporal resolution for most listeners with impaired hearing may not be deficient compared to listeners with normal hearing.

One explanation for the inability of listeners with impaired hearing to perceptually attend to the carrier signal is their poorer frequency resolution. Contrasting his research investigating the temporal acuity of listeners with impaired hearing, Turner found significant perceptual deficits in frequency acuity for listeners with impaired hearing compared to listeners with normal hearing [Turner and Nelson(1982)]. Frequency discrimination was not only poorer at high frequencies where sensitivity thresholds were diminished, but also at low frequencies where sensitivity thresholds were normal.

In addition to the research on listeners with sensorineural hearing loss, the perceptual importance of attending to temporal structure has been investigated for listeners with dyslexia [Tallal(1980)] and auditory neuropathy [Zeng, Oba, Garde, Sininger, and Starr(1999)]. In a psychophysical task [Goswami et al.(2002)] and a speech perception task [Goswami et al.(2010)] children with dyslexia performed differently from children without dyslexia for amplitude envelope onset detection. For the speech perception task, children's discrimination of a phonetic contrast, /ba/-/wa/, was measured using synthetic speech stimuli created either by varying the rate of formant frequency change or the rate of amplitude change or rise time. Children with dyslexia showed excellent phonetic discrimination based on formant transition duration, but poor phonetic discrimination based on envelope cues. In a task to assess the

sensitivity to amplitude modulation of acoustic stimuli, adult listeners with dyslexia have been found to have significantly poorer thresholds of amplitude modulation depth than matched control listeners [Menell, McAnally, Stein(1999)]. A similar result was also replicated in children with dyslexia [Rocheron, Lorenzi, Fullgrabe, and Dumont(2002)]. Listeners with auditory neuropathy have poorer temporal gap detection and temporal modulation transfer functions [Zeng, Oba, Garde, Sininger, and Starr(1999)], attributed to a severe impairment in temporal processing abilities.

Investigating the Spectral Envelope

In addition to investigating the low-frequency modulator and high-frequency carrier of speech, researchers have also investigated the perception of the speech spectral filter and the excitation source signal as acoustic characteristics. Sine-wave speech (SWS) is a method to represent speech by its spectral skeleton and remove the source signal. SWS is created by estimating frequencies of formants and synthesizing time-varying tones at these frequencies to replace the original signal [Remez et al.(1981)]. Near-perfect speech recognition has been achieved with SWS given sufficient training and sentence context [Barker and Cooke(1999),Nittrouer and Lowenstein(2010)]. This has also been demonstrated for listeners of a tonal language. Mean recognition of Mandarin sentences with SWS has been reported to be 91.6%, despite poor Mandarin tone recognition (32.7%) with SWS [Feng et al.(2012)]. This would suggest the source signal is perceptually important for Mandarin tone, but is less important for sentence materials with linguistic context.

A separate method of representing the speech signal has been used to contrast the perceptual importance of the spectral filter and source signal [Nittrouer and Tarr(2011),Tarr and Nittrouer(2011),Tarr and Nittrouer(2013)]. In these experiments

two spectrally distant components of synthetic speech were presented in a labeling task. If combined, the components formed a speech-like spectral filter. In the control condition, the two components shared a common harmonic relationship characteristic of a single source signal. In other experimental conditions, the two components had conflicting source signals. If listeners perceptually integrated the two spectrally distant components, this suggested they were overriding the conflicting source signals and giving greater perceptual importance to the spectral filter. If listeners perceptually segregated the two components, this suggested they were overriding the speech-like spectral filter and giving greater perceptual importance to the source signals. The results suggested that child listeners consistently gave greater perceptual importance to the spectral filter. Adult listeners segregated the two components when both were periodic but had conflicting fundamental frequencies. However, adult listeners integrated the two components when one had a periodic source signal and the other had an aperiodic source signal. These experiments will be discussed in Chapter 5.

Summary

In general, this kind of speech perception research has used various methods to separate one acoustic characteristic from a speech signal, with the remaining residual signal representing the original signal without the characteristic of interest. Therefore, removing the temporal envelope resulted in a residual that lacked the low-frequency modulation of speech. Similarly, removing the spectral filter resulted in a residual that lacked the vocal tract resonances of speech. In these cases, the speech signal was represented as either a carrier signal modulated in amplitude or as an excitation source filtered spectrally.

2.1.2 An Alternative Representation

Rather than having one model of speech as a modulator and carrier, and another model of speech as a source and a filter, Rosen (1992) suggested a single decomposition of a speech signal based on temporal structure represented by three parts: *envelope*, *fine structure*, and *periodicity*. The temporal *envelope* was related to the low-frequency fluctuations in the signal. The temporal *fine structure* was related to the formant pattern, or spectral filter in the signal. The temporal *periodicity* was related to the signal's fundamental frequency for voiced speech. It was suggested that these characteristics of the signal reside in distinct frequency regions. The temporal *envelope* was represented by fluctuations between 2 and 50 Hz. The temporal *fine structure* was represented by fluctuations between 600 Hz and 10 kHz. The temporal *periodicity* was represented by fluctuations between 50 and 500 Hz.

Analyzing the temporal fluctuations in amplitude structure by dividing the signal into independent frequency regions is an attempt to represent three aspects of temporal structure specifically for speech. However, it does not provide a method to recover each aspect separately. The periodicity in a signal is not limited to a frequency range of 50 - 500 Hz; rather, harmonics of the signal extend to much higher frequencies and periodicity can even be perceived in the absence of a fundamental frequency [Plomp(1967)]. Similarly, the timbre, quality, or gross spectral structure is not limited to the frequency range of 600 Hz - 10 kHz; rather, it extends to lower frequencies. In particular, the formant pattern of close vowels extends lower than 600 Hz [Peterson and Barney(1952)]. For voiced speech, the timbre, quality, or gross spectral shape is encoded in the speech signal by the relative amplitudes of individual harmonics including the amplitude of the fundamental frequency.

Previous work has been presented to decompose the periodicity and gross spectral structure in speech signals [O'Shaughnessy(1987)]. Instead of dividing the spectrum based on a frequency range, a speech signal $x[t]$ was 'deconvolved' into an excitation source $u[t]$ and spectral shaping filter $h[t]$. The spectral representations of the decomposed parts are related by: $X[z] = U[z]H[z]$. Linear predictive coding (LPC) is one method of decomposing speech signals in this way. Although LPC provides a sufficient method to remove and recover these parts, practical modification of these parts is challenging. For instance, one modification could be sharpening the resonances of the estimated shaping filter $H[z]$. Increasing the magnitude of the filter poles in the z-plane to sharpen the resonances can lead to filter instability.

The frequency divisions for *periodicity* and *fine structure* presented in Rosen (1992) did not 'deconvolve' the source signal, $u[t]$, and spectral filter, $h[t]$. Similarly, by defining temporal *envelope* information as amplitude fluctuations below 50 Hz, this descriptive framework departed from the mathematical definition of the HT envelope. Rosen instead defined temporal structure related to speech production. The HT envelope contains the amplitude structure for all frequencies in a signal over time. For speech signals, the HT envelope contains the amplitude modulations resulting from speakers widening and narrowing their vocal tracts; it also contains the amplitude structure from vibrations of the laryngeal source or 'carrier' signal. Analyzing amplitude fluctuations below 50 Hz aimed to separate the amplitude modulations resulting from speakers widening and narrowing their vocal tracts from the amplitude vibrations of the laryngeal signal.

Alternatively, rather than dividing a signal into distinct frequency regions, recovery of temporal *envelope*, *fine structure*, and *periodicity* could be represented in

relation to the function of an aspect of speech production. Therefore, the temporal *envelope* could be represented by amplitude modulations $A_1[t]$, temporal *fine structure* represented by a spectral filter $h_1[t]$, and temporal *periodicity* represented by a source signal $u_1[t]$. By this representation, the speech signal $x[t]$ should be related to *envelope*, *fine structure*, and *periodicity* such that $x[t] = A_1[t](h_1[t] * u_1[t])$. Furthermore, the general relationship of $x[t] = \mathbb{R}\{A[t]e^{j\phi[t]}\}$ from Loughlin and Tacer (1996) is maintained using this representation of a speech signal. Fig. 2.2 shows a diagram of this model of speech production.

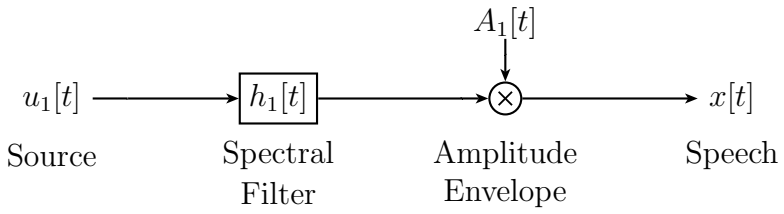

Figure 2.2: Diagram of the Speech Production Model

Loughlin and Tacer (1996) argued that although the Hilbert transform provides a method of uniquely defining $A[t]$ and $\phi[t]$ for any signal, these representations fail to satisfy reasonable physical conditions for the decomposed signals. A additional criticism of the Hilbert Transform is that the analytic signal pair of $A[t]$ and $\phi[t]$ are not reasonable estimates of the specific aspects of speech production to be estimated here. The amplitude envelope $A[t]$ is not recovered from the speech signal independently of the source signal and the spectral filter. As a result, the 'residual' signal

after $A[t]$ has been removed cannot be used to reasonably represent the source signal $u_1[t]$ and the spectral filter $h_1[t]$.

Fig. 2.3 shows an example of the waveform and spectrogram of a speech signal, HT instantaneous frequency, and HT envelope. The HT instantaneous frequency signal has prominent energy at frequencies that had less energy in the original signal, which is best observed in the spectrogram. Because of this, the spectral filter has not be preserved after removing $A[t]$. An additional observation is that the HT envelope contains amplitude fluctuations faster than the syllabic rate of opening and closing of the vocal tract for this utterance and contains fluctuations based on the fundamental frequency of the source signal. By low-pass filtering the HT envelope, amplitude fluctuations can be limited to a desired frequency range. However, the original signal cannot be reconstructed from the HT envelope and instantaneous frequency signal if either has been filtered.

Methodologically, it has been the practice to assume one general (i.e. HT) pair of envelope and instantaneous frequency, then process it for specifically for speech. Rather, a different envelope and carrier could be initially recovered that are specific for speech and do not require additional processing. For these reasons, a new framework for representing acoustic characteristics of speech is presented such that *periodicity*, *envelope*, and *fine structure* in a signal can all be recovered, removed and replaced independently.

2.1.3 Nomenclature

Several terms are defined to be used for the remained of this dissertation. The *Gross Temporal Envelope* (GTE) will refer to the signal structure that is recovered by

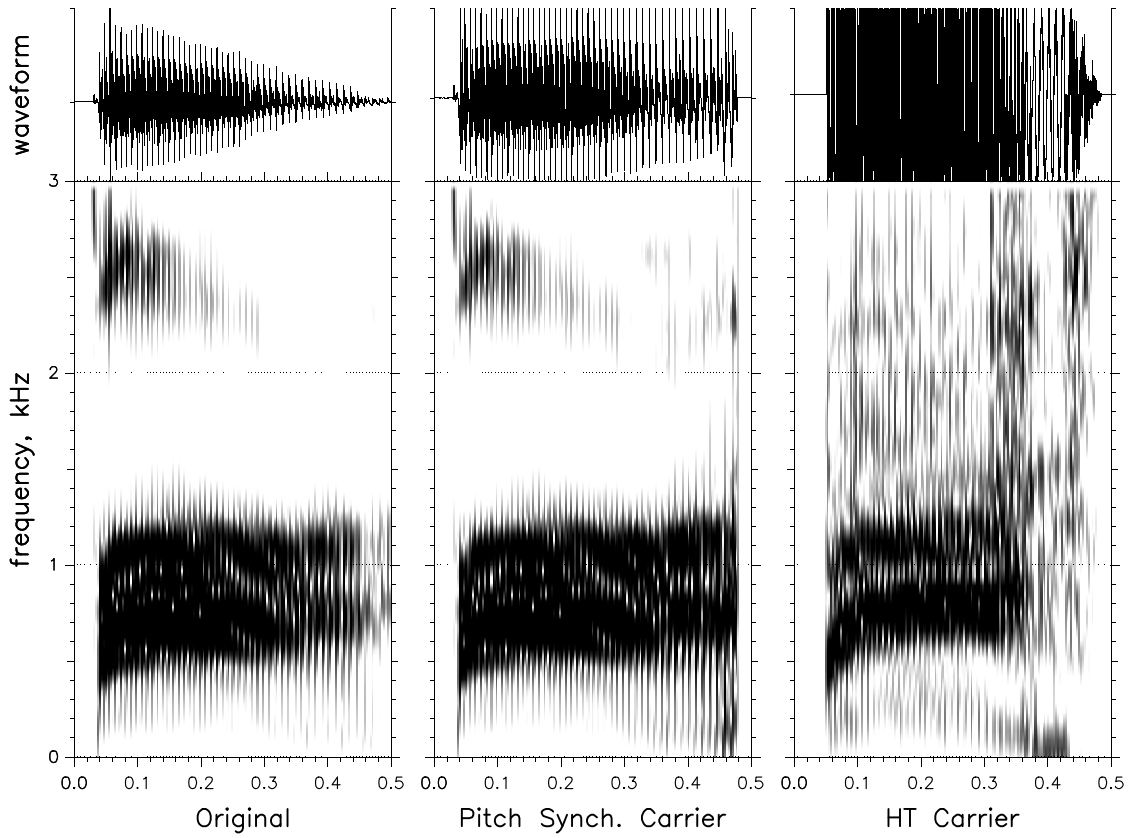

Figure 2.3: Comparison of Speech Signal with HT Instantaneous Frequency and HT Envelope

estimating the syllabic rate of speech. The *Gross Spectral Envelope* (GSE) will refer to the signal structure estimating the resonances of the vocal tract. The *source* will refer to signal structure estimating the laryngeal vibrations of a speech signal. These terms will be qualified by particular methods of estimation such as PS GTE, LPC GSE, etc. A *residual* signal will refer to any signal that has had a temporal envelope and/or spectral envelope removed. For pitch synchronous processing, a residual signal may consistent of source+GSE if the GTE has been removed. A residual signal may also consist of source+GTE if the GSE has been removed. Finally, a residual signal may consist of only the source signal if the GTE and GSE have both been removed. For HT processing, the residual signal will refer to the HT instantaneous frequency signal. The HT envelope will refer to the mathematical definition provided previously. A *temporal envelope* will refer to any representation of the amplitude magnitude over time. Signal *fine structure* will refer to any representation of the spectral magnitude over time.

2.2 Pitch Synchronous Framework for Representing Acoustic Characteristics of Speech

2.2.1 Background

The speech signal is a sequence of sounds created by the vibrations of air moving through the vocal tract. Portions of the speech signal are *voiced*, occurring when a speaker's vocal folds vibrate. Due to the asymmetrical rate of vocal fold vibrations, the produced signal contains harmonics all related to a fundamental frequency, which is perceived as pitch (despite the distinction in perceptual research, the term 'pitch' will be used synonymously with fundamental frequency to remain consistent with previous literature in signal processing). A *pitch period* can be defined for an ideal

speech signal as the time period in which a single cycle of vibration completes prior to repetition. A pitch period will also be defined with a specific signal phase: a cycle beginning at the positive going zero-crossing and ending at the zero-crossing that begins the next signal repetition cycle. This additional criterion is used for consistency during processing. Voiced portions of a speech signal can be divided temporally and processed synchronously to the individual pitch periods.

2.2.2 Pitch Synchronous Processing

There has been several applications that used pitch synchronous (PS) processing for speech signals. It has been used in the analysis of vowels [Mathews, Miller, and David(1961)], text-to-speech synthesis [Moulines and Charpentier(1990)], linear predictive coding [Paliwal and Rao(1981)], independent modification of formant frequencies and bandwidths [Kuwabara(1984)], glottal wave analysis [Alku(1992)], and speaker recognition [Kim and Chung(2004), Zilca, Kingsbury, Navratil, and Ramaswamy(2006)].

Here, PS processing was applied to speech signals to isolate several acoustic characteristics of interest. PS processing was useful in this application because it is directly related to the underlying source signal. By initially analyzing the pitch periods of the signal, the GTE and GSE can be recovered independently of the source. PS processing has been suggested previously in other methods of demodulating the envelope modulations from a speech signal [Li, Nie, Atlas, and Rubinstein(2010)].

An important assumption in the subsequent PS processing techniques was that each pitch period of voiced speech was assumed to be a single period of a periodic signal which extends infinitely over time. A consequence of this assumption was

that the assumptions are satisfied necessary for a Fourier series expansion of the periodic and band-limited signal into a summation of sinusoids. The signal processing methods presented depended on this assumption both mathematically, and also for computational convenience.

The original signal was represented as a sequence of pitch periods:

$$x[t] = \{x_{p1}[t], x_{p2}[t], \cdots, x_{pn}[t]\}$$

that can each be processed separately.

Pitch Period Detection

A first step of pitch synchronous processing is detecting the beginning and ending of individual pitch periods. This analysis should fulfill the signal phase requirements of a pitch period described previously, while using the pitch of the signal to estimate these temporal locations.

There have been many techniques proposed to detect the pitch of speech signals [Gold and Rabiner(1964), Rabiner, Cheng, Rosenberg, and McGonegal(1976), Kadambe(1992), Ying, Jamieson, Michell(1996), Ercelebi(2003)]. A method using signal autocorrelation was found to be successful for the speech stimuli used in the subsequently described experiment. Autocorrelation is a measurement of the similarity in a signal over time, therefore it can be used to estimate repetitions in a signal such as cycles of a pitch period. The use of autocorrelation has been previously examined for its use in pitch detection, and is described in detail in Rabiner (1977).

Several assumptions are appropriate when using signal autocorrelation to detect pitch in a speech signal. First, speech typically has a fundamental frequency between

80-300 Hz [Peterson and Barney(1952)]. Therefore, a pitch period will have a minimum length, $\hat{p}_{min}$ of 3.3 ms and a maximum length, $\hat{p}_{max}$ of 12.5 ms. Second, the instantaneous pitch of a speech signal is variable over time. Therefore, it is advantageous to restrict the signal to a time window for the autocorrelation calculation to a length in temporal proximity to the pitch period to be estimated. A time window of 37.5 ms or three times the maximum possible pitch period length will be used for the autocorrelation calculation, $3 \cdot \hat{p}_{max}$. Third, the temporal ending of one pitch period is the temporal beginning of the subsequent pitch period. Therefore, each pitch period will be estimated in a successive manner starting with the first positive going zero-crossing in the voiced portion of the speech signal.

The following can be used to estimate the length of a pitch period, $\hat{p}$, for a signal, $s[t]$, in which the temporal beginning of the pitch period is at $t = 0$.

$$
\begin{aligned}
\hat{p} &= \underset{\tau \in [\hat{p}_{min}, \hat{p}_{max}]}{\arg\max} \; R_{\bar{s}\bar{s}}[\tau] \\
R_{\bar{s}\bar{s}}[\tau] &= \sum_{t=0}^{3 \cdot \hat{p}_{max}} \bar{s}[t]\bar{s}[t-\tau] \\
\bar{s}[t] &= s[t]w_R[t] \\
w_R[t] &= \begin{cases} 1 & 0 \leq t \leq 3 \cdot \hat{p}_{max} \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

The estimate of pitch period length is determined by the timing of the maximum in the autocorrelation signal during the range of possible fundamental frequency values.

After the estimation of the pitch period length, $\hat{p}$, the temporal location of the end of the pitch period should be determined. This can be done by finding the appropriate zero-crossing location in the signal near the estimate of the pitch period length. After finding the location of the ending of the pitch period, this process can

be repeated with this location as the beginning of the next pitch period and so forth for the remaining portions of 'voiced' speech in the signal.

2.2.3 Gross Temporal Envelope Processing

When a speech signal is produced naturally, the low-frequency amplitude modulations of the GTE are encoded on the same signal as the source signal and GSE. However, it is desirable to represent the GTE independently of the source signal and GSE. Therefore, signal processing should be used to separate these components, but also allow for the original speech signal to be reconstructed after processing. First, a method will be presented to recover the signal's GTE. Next, a method will be presented to recover the GTE, leaving a residual signal representing the source signal and GSE. Finally, different methods will be discussed to replace the original GTE with a new GTE and recombine it with the residual signal.

Representing the GTE

With PS processing, the amplitude (RMS, peak, etc.) of each pitch period can be calculated. The GTE can be constructed by sequencing the measured amplitudes over time. Using this approach, the GTE is represented by the relative amplitude of individual pitch periods over time. The residual (source + GSE) signal of speech is assumed to have a constant temporal envelope during a pitch period. Therefore, temporal amplitude modulations will be attributed to the syllabic rate of speech related to the opening and closing of the vocal tract and not related to vibrations of the source signal. This representation provides a sufficient approximation to the low-frequency temporal *envelope* as defined in Rosen (1992).

Recovering the GTE

One purpose of GTE processing is to decompose a speech signal based on the general relationship, $x[t] = \mathbb{R}\{A[t]e^{j\phi[t]}\}$. First, recovery of the envelope, $A[t]$, will be described. As discussed previously, the PS GTE is represented as the relative amplitude of each pitch period across the signal. Therefore, in order to recover the PS GTE, the speech signal is divided into pitch periods and the amplitude is measures in each pitch period separately. Two conventional methods for measuring amplitude in a signal are the peak amplitude $\alpha_{peak} = \max(\text{abs}(x_p[t]))$ and the RMS amplitude $\alpha_{rms} = \sqrt{\frac{1}{n}({x_1}^2 + {x_2}^2 + \ldots + {x_m}^2)}$ for a pitch period $x_p[t] = \{x_1, x_2, \cdots, x_m\}$. To recover the PS GTE for a signal $x[t]$, each time sample of $A[t]$ should be given the amplitude value measured for the pitch period at the current time.

Fig. 2.4 compares three difference estimates of temporal envelopes. The unfiltered HT envelope, low-pass filtered HT amplitude below 20 Hz, and PS GTE are shown. Although there are discontinuities in the estimated amplitude of the PS GTE at the transitions between pitch periods, this envelope more closely resembles the syllabic rate of speech for this speech utterance than does the unfiltered HT envelope.

Removing the GTE - PS Temporal Amplitude Normalization

Another step of decomposing the speech signal is to determine the instantaneous phase $\phi[t]$, related to the residual (source+GSE) signal. This is accomplished by removing the GTE from the speech signal. The recovered PS GTE can be removed from the signal by using PS Temporal Amplitude Normalization. It will be assumed that the speech residual signal has a consistent, or normalized, amplitude across pitch

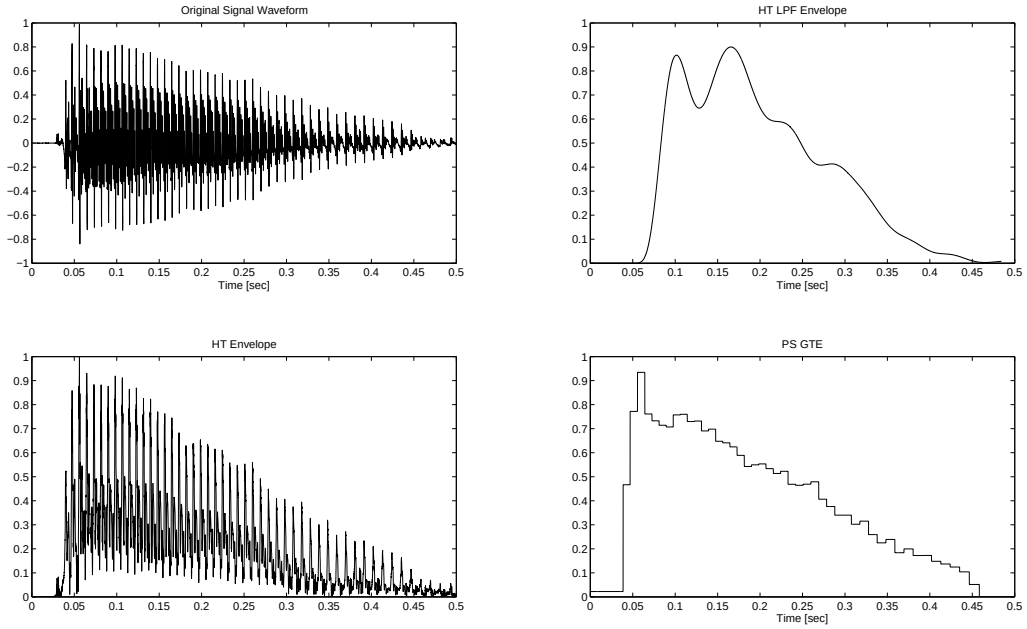

Figure 2.4: Comparison of PS GTE with HT Envelope and HT Low-pass Filtered Envelope

periods. Therefore, signal processing should be used on a speech signal so that each pitch period has the same amplitude.

Normalizing each pitch period for either signal peak or average energy involves scaling the signal's amplitude to achieve the desired level. The scaling is completed by multiplying each individual sample by the appropriate scaling value, which satisfies the relationship between $A[t]$ and $\phi[t]$. This 'point-wise' multiplication can be performed because the assumptions are satisfied for a Fourier series expansion of the periodic and band-limited signal into a summation of sinusoids for voiced speech signals. Therefore, scaling the signal by a constant value maintains frequency, phase, and even relative amplitude of the composite sinusoids.

To normalize a pitch period, $x_p[t] = \{x_1, x_2, ..., x_m\}$, such that the peak has a desired amplitude level, c_{peak}, each sample should be multiplied by the same scaling factor, $s_{peak}[t] = \{s, s, ..., s\}$, such that: $x_{normalized}[t] = s_{peak}[t]x_p[t]$ where $s = \frac{c_{peak}}{\alpha_{peak}}$.

Similarly, to normalize a pitch period, $x_p[t] = \{x_1, x_2, ..., x_m\}$, such that the amplitude achieves a desired RMS value, c_{rms}, each sample should be multiplied by the same scaling factor, $s_{rms}[t] = \{s, s, ..., s\}$, such that: $x_{normalized}[t] = s_{rms}[t]x_p[t]$ where $s = \frac{c_{rms}}{\alpha_{rms}}$.

Applying normalization to each pitch period separately in the voiced portion of a speech signal will result in a residual (source+GSE) signal and is the equivalent of demodulating the PS GTE from the speech signal. Fig. 2.5 compares the results of the PS GTE removal method with the HT instantaneous frequency signal. In this example, the amplitude in each pitch period has been scaled so that the signal peaks are normalized in the PS condition. The PS signal is spectrally similar to the original signal, while the HT instantaneous frequency signal is spectrally different.

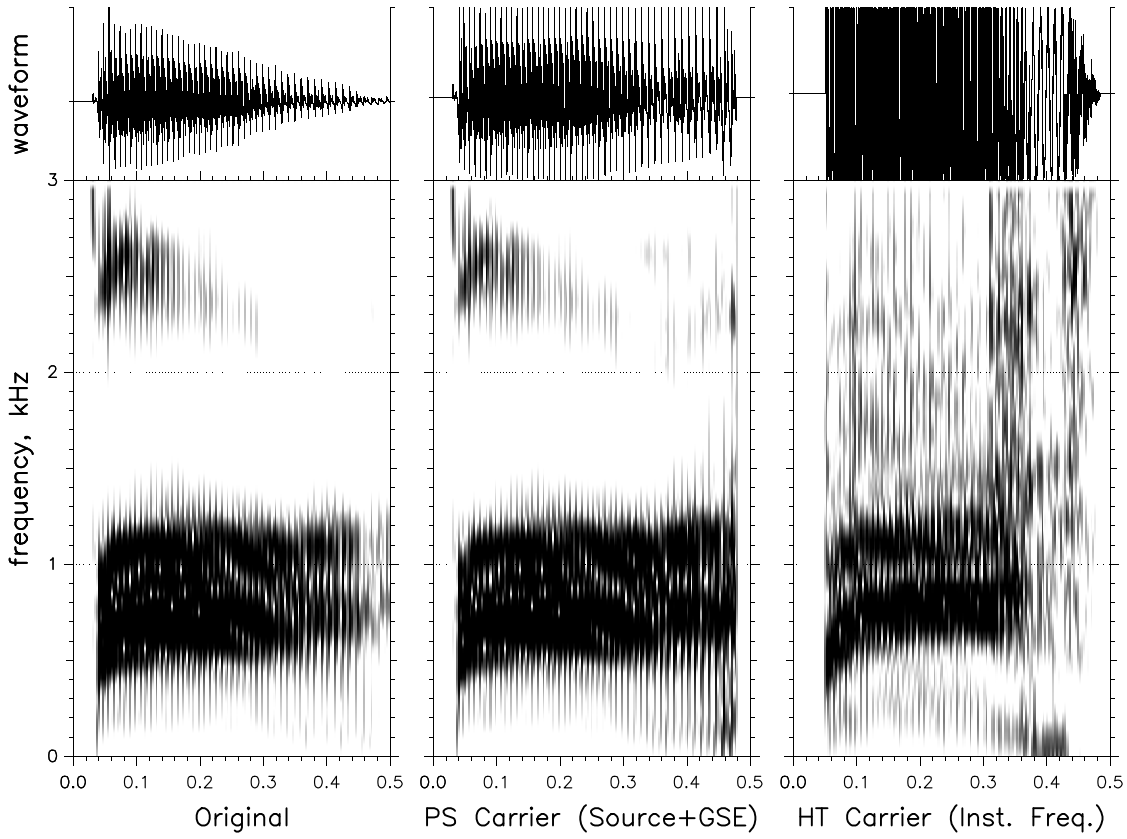

Figure 2.5: Comparison of PS GTE Removal with HT Demodulation

The residual (source+GSE) signal can be used as a frequency signal, $f[t]$ such that $x[t] = A[t]f[t]$. A phase signal can be calculated: $\phi[t] = ln\frac{f[t]}{j}$, in order to recover the $A[t]$ and $\phi[t]$ pair for writing $x[t] = \mathbb{R}\{A[t]e^{j\phi[t]}\}$, as in [Loughlin and Tacer(1996)]. In this sense, the recovered PS GTE and residual signal reconstruct the original signal, and directly estimate the GTE independent of the GSE and source signal without additional processing. Fig. 2.6 shows a diagram of the PS Temporal Normalization processing method.

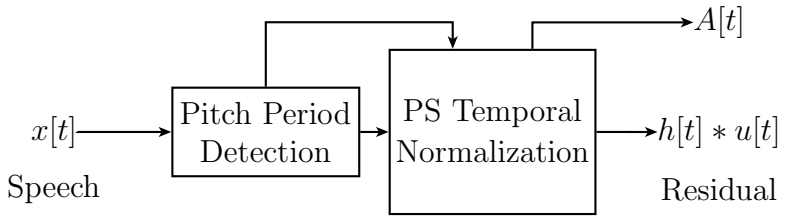

Figure 2.6: Diagram of PS Temporal Normalization

Replacing the GTE

There are a several experimental applications that have involved replacing the GTE for a speech utterance. On example is varying the amplitude rise time for an utterance on a multi-step continuum. PS processing could be used precisely choose amplitude values the initial pitch periods of an utterance. As another example, the amplitude values of the pitch periods from one speech signal could be used to scale the amplitude of the residual signal for a different speech signal. Several experiments have investigated perceptual weighting by interchanging of syllable amplitude envelopes between different speech tokens [Nittrouer and Studdert-Kennedy(1986), Nittrouer,

Lowenstein, and Tarr(2012)]. The technique provides an alternative approach for auditory chimera processing to the previous method using the HT [Smith, Delgutte, and Oxenham(2002)]. This experimental method can be used to investigate perceptual weighting of specific acoustic characteristics. After interchanging the GTE of speech signals, listeners are presented are presented stimuli with conflicting cues and respond based on the cue that is perceptually important.

There are several complications when interchanging the GTE between utterances using PS processing. For natural speech signals, pitch periods across different recordings will be temporally asynchronous, even for the same speaker on the same utterance. Also, utterances may have a different number of pitch periods to process. Therefore, it may not be appropriate to inherit pitch period amplitude values across different signals. Instead, it may be more appropriate to apply a different estimated GTE. Two methods that could be substituted are: a low-pass filtered HT envelope or a low-pass filtered, half-wave rectified amplitude envelope. Provided the GTE is sufficiently low-pass filtered to remove vibrations from the source signal, the consequence should not be significantly worse than typical 'windowing' of a speech signal. Due to the point-wise multiplication of these methods, the GTE must be the same length as the residual signal. If necessary, interpolation methods can be used to change the length of the GTE to match the residual signal.

Fig. 2.7 shows an example of the entire process of PS GTE removal and application of a low-pass filtered half-wave rectified amplitude envelope. Two signals were used from the recordings of two different speakers for the utterance *'ba'*. The signal from the first speaker was used as the GTE, while the signal from the second speaker was used as the residual signal. First, the signal to be used as the residual had the GTE

removed following the PS method described previously. Second, the signal to be used for the GTE was half-wave rectified and low-pass filtered below 20 Hz. Linear interpolation was performed to match the length of the GTE to length of the residual signal. Third, the GTE was temporally shifted to maximize the correlation with the original GTE from the residual signal. Finally, a point-wise multiplication was performed to overlay the GTE on the residual, replacing the original GTE. As desired, the GTE of the final processed signal resembles the GTE of the appropriate original signal, best viewed in the waveform of the signals. Similarly, the spectrogram of the final processed signal closely matches the spectrogram of the original signal used as the residual signal.

2.2.4 Gross Spectral Envelope Processing

Decomposing a speech signal into a GTE and residual (source + GSE) signal is a method to represent the temporal information of speech. Additionally, the GSE could be removed from a residual (source + GTE) signal to represent the spectral information of speech. Rosen (1992) stated that the temporal *fine structure* contains the formant pattern, is related to timbre, or quality, and informs about the spectrum of a sound. The formant pattern of speech is encoded on the signal as a spectral envelope evident by the relative amplitude of individual harmonics of the source signal at a single time instance.

Representing the GSE

For PS processing, the GSE was represented by the relative amplitude of individual harmonics during a single pitch period. It was assumed that the GSE is stationary during a single pitch period. Each pitch period in a signal can have a different GSE.

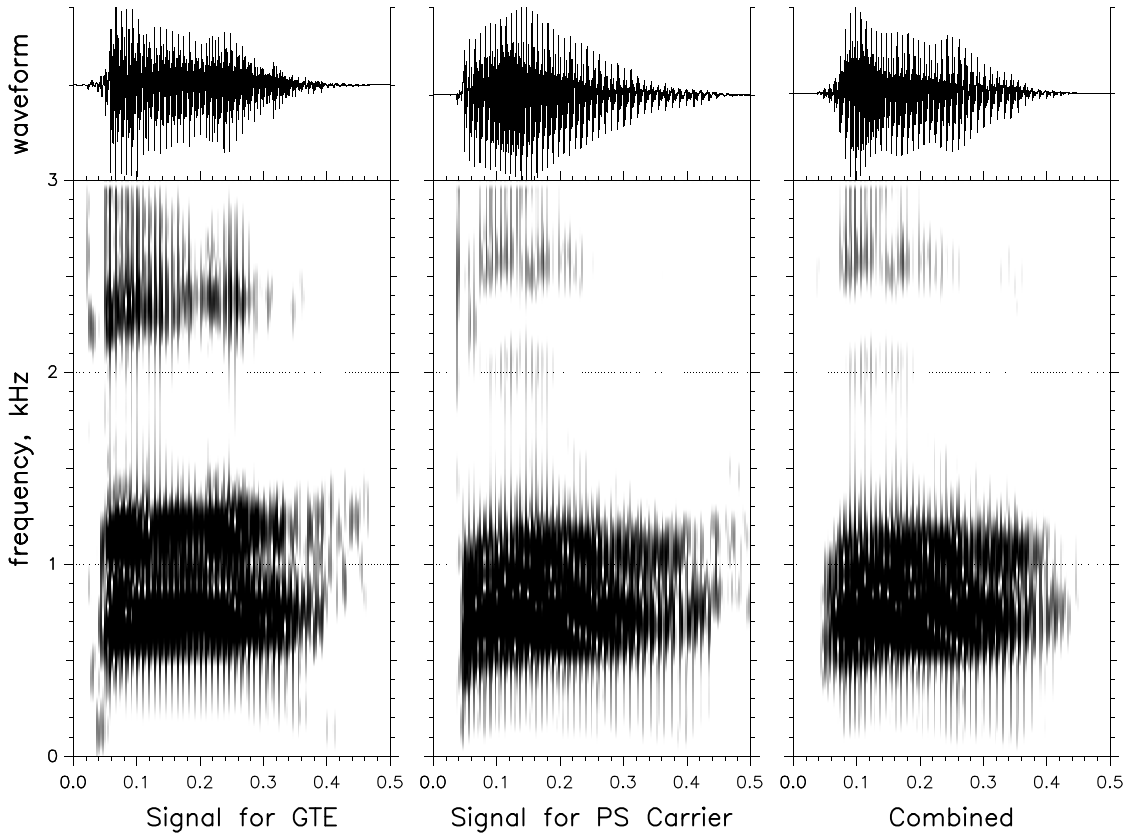

Figure 2.7: GTE Applied to PS Residual (source+GSE) from a Different Speaker

The GSE was represented by the filter necessary to recover the original speech signal from a signal where all the harmonics are the same amplitude. Therefore, the GSE is related to resonances of the vocal tract during speech production. This representation provides a sufficient approximation to the *fine-structure* as defined in Rosen (1992).

The residual (source + GTE) signal of speech was assumed to have a normalized, or flat spectral envelope during a pitch period. A similar assumption has been made for speech signals before. As an example, for LPC source-filter separation, the source signal is estimated to be as spectrally flat as possible, while the filter contains the spectral information [O'Shaughnessy(1987)]. The harmonics of a signal produced by a larynx are not the same amplitude, but rather decrease in amplitude with increasing frequency. Nonetheless, it is convenient to consolidate spectral information about the resonances of the vocal tract and the spectral roll-off of the source signal in the GSE.

Recovering the GSE

One purpose of GSE processing is to deconvolve the spectral information of a speech signal, $x[t] = u[t]*h[t]$, or represented in the spectral domain, $X[z] = U[z]H[z]$. First, recovery of the GSE $H[z]$ will be presented. As discussed previously, the PS GSE is defined as the relative amplitude of each harmonic within a pitch period. Therefore, the first step to recovering the GSE is to divide the signal into individual pitch periods, then separate each harmonic and measure its amplitude. By using PS processing, the duration of each pitch period is known and the fundamental frequency during the pitch period can be estimated. Harmonics in the signal are related to the fundamental frequency by integer multiples. Therefore, the spectrum can be divided to isolate individual harmonics for separate amplitude measurement. A filter bank can be designed for each pitch period with band-pass cut-off frequencies of $n \cdot f0 + \frac{f0}{2}$

for $n = 1, 2, 3, ...$, resulting in cut-off frequencies below the Nyquist frequency. To make the computation of filtering more convenient, each separate pitch period can be assumed to be a single cycle in a periodic signal. This periodic extension of the pitch period allows for more precise filtering than if a single cycle was filtered alone. Ideally, the summation of filters used in the filter bank would create and all-pass filter, $|H_{ap}[z]| = |H_1[z]| + |H_2[z]| + \cdots + |H_n[z]| = 1$. To recover the PS GSE, the amplitude of the signal after it is filtered in each channel of the filter bank should be measured. This amplitude can be used to scale the gain of each filter bank channel. By the linearity property of the z-transform, scaling the amplitude of the separate channels to recover the GSE, $H[z]$, results in: $a_1 h_1[t] + a_2 h_2[t] + \cdots + a_n h_n[t] \Leftrightarrow a_1 H_1[z] + a_2 H_2[z] + \cdots + a_n H_n[z]$.

Removing the GSE - PS Spectral Amplitude Normalization

Another step of decomposing the speech signal is to estimate $u[t]$, related to the residual (source+GTE) signal. The GSE can be removed from a residual (source + GTE) signal such that all harmonics have the same amplitude. Rabiner (1977) presented several types of nonlinear preprocessing which can be used to flatten the spectrum of a speech signal effectively. LPC source-filter separation also can be used to achieve a source signal with a flattened spectrum. PS processing offers an alternative method for processing a signal such that the harmonics have equal amplitude.

After the signal passes through the filter bank described previously, each channel output can be normalized to the same amplitude. Amplitude of each channel can be normalized based on peak or RMS measurements. All of the individual channels can be recombined to produce a signal where all harmonics are similar in amplitude, resulting in a signal with a flat spectrum. This process can be repeated for each

pitch period in a signal. Fig. 2.8 shows a diagram of the PS Spectral Normalization processing method.

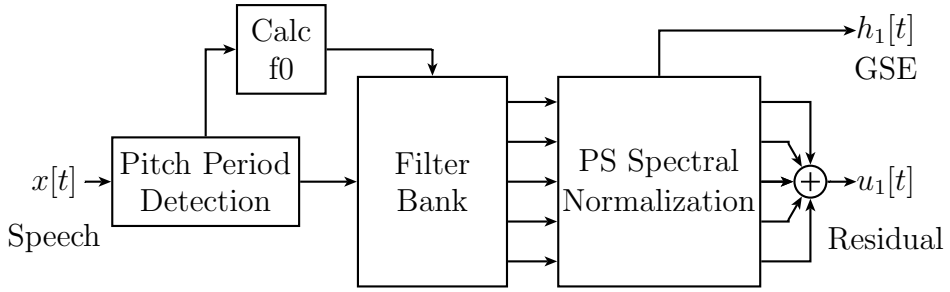

Figure 2.8: Diagram of PS Temporal Normalization

Fig. 2.9 shows an example of the spectrum of a single pitch period, spectrally flattened pitch period, and the frequency response of the LPC GSE and PS GSE.

Replacing the GSE

The original speech signal can be perfectly reconstructed following the GSE removal method previously described. The 'point-wise' multiplication used for PS Spectral Amplitude Normalization can simply be inverted for each channel of the filter bank. There may be several applications when it desired to replace the PS GSE with a different GSE. As discussed previously in regards to processing the GTE, perceptual experiments could be conducted by processing the GSE. There are several complications when using the PS processing for the GSE. In situations where the GSE from one pitch period is applied to a pitch period with a different length, the filter bank with channel divisions based on the signal's fundamental frequency cannot be used. The LPC GSE is one possibility of a filter than can be substituted in each pitch period. This is similar to substituting the low-pass filtered HT envelope to modulate the amplitude of a residual (source + GSE) signal.

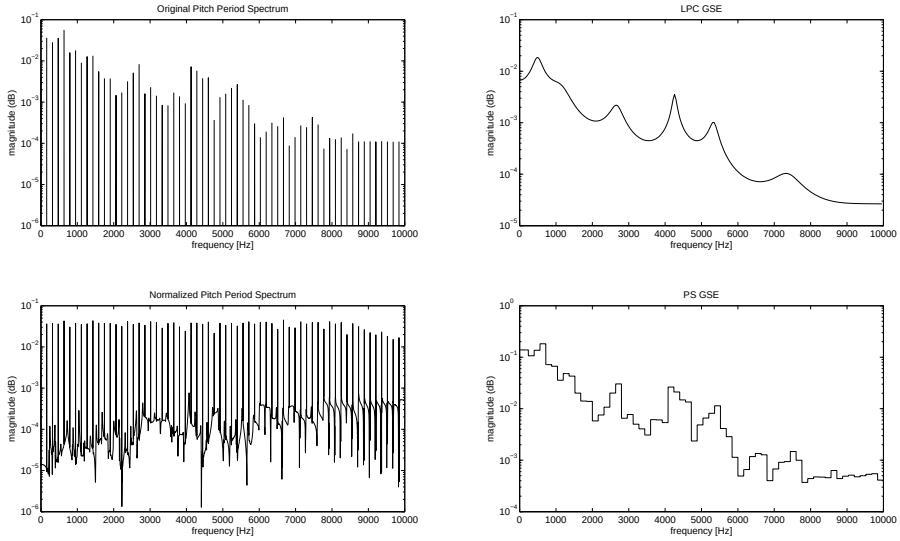

Figure 2.9: Comparison of Spectrum using PS Spectral Amplitude Normalization and Related GSE from LPC and PS processing

2.2.5 Source Signal Processing

The previous sections have discussed how PS processing can be used to recover and remove the GTE or GSE from a speech signal and also apply new envelopes. These techniques can be used to decompose the temporal *envelope* and *fine structure* described in [Rosen(1992)]. Next, techniques to process the temporal *periodicity* based on the source signal will be discussed.

Representing the Source Signal

For PS processing, the source signal of speech is represented by a signal with a constant GTE throughout an utterance and with a flat spectrum in each pitch period. By the first assumption, the source signal is either vibrating at a constant amplitude, or is not vibrating at all. In reality, variation of the amplitude of a speech signal is

due to the varying amplitude of the source signal and the openness of the vocal tract. However, for PS processing, variations in temporal amplitude are consolidate in the GTE. Similarly, the harmonics of the source signal are assumed to have the same amplitude. This assumption is not ecologically valid because the harmonics produced by a human larynx decrease in amplitude with increasing frequency. Nonetheless, it is convenient to assume a flat spectrum for the source signal and consolidate spectral structure in the GSE. This representation provides a sufficient approximation to the *periodicity* as defined in [Rosen(1992)].

Recovering the Source Signal

In order to recover the PS source signal, the GTE and GSE should be removed following the methods described in the previous sections. This residual signal will have a contant GTE throughout an utterance and with a flat spectrum in each pitch period.

Replacing the Source Signal

The source signal cannot be 'removed' because the GTE and GSE modulate and filter a residual signal. The source signal can be replaced is several different ways. Aperiodic noise has been used previously as the residual signal in experiments investigating the role of temporal cues in speech production and could be used to replace the source signal. Similarly, a synthetic speech source signal composed of a tone complex could also be used to replace the original source signal. These other source signals could replace the original source signal by applying the recovered GTE and GSE from the original signal to the synthesized source signal.

Modifications of the original source signal can also be used to create a new source signal. In order to modify the source signal without changing the GTE and GSE, the GTE and GSE must be estimated pre-processing and reapplied post-processing. Otherwise, changes in frequency to the source signal will also change the GSE. By doing so, the formant pattern of the speech utterance would be distorted meaning the *periodicity* would not be processed independent of the temporal *fine structure*. After recovering the source signal, the pitch could be modified and have the original GTE and GSE reapplied.

After the source signal has been recovered independently of the GSE, the *periodicity* of the signal can be modified by 'resampling' a pitch period. Resampling a pitch period can consist of upsampling and downsampling while maintaining the same sampling rate. Upsampling a pitch period while maintaining the same sampling rate will mean that the pitch period will be longer and have a lower frequency. Conversely, downsampling will increase frequency. The fundamental frequency, $f0$, of a pitch period is related to the sampling rate, Fs, and number of samples in a pitch period, n, by $f0 = \frac{Fs}{n}$. Therefore, to achieve a desired fundamental frequency, $\hat{f0}$, a pitch period should be resampled to have a duration, $\hat{n} = \frac{Fs}{\hat{f0}}$.

After a pitch period of the source signal is modified, a GSE can be applied to resynthesize a new speech signal. By changing the fundamental frequency, it is not appropriate to use the PS filter bank technique based on dividing the spectrum to process the amplitude of individual harmonics separately. A GSE estimate by using LPC could be applied. Also, by modifying the duration of the pitch period, the length of the residual signal will be changed. The length of the original utterance can be maintained if the original GTE is applied and the length of the residual signal

is matched to the length of the GTE. If it is desired to maintain the length of the residual signal, a new GTE can be used. In this case, the modified pitch period can either be scaled to the same peak or RMS amplitude of the original pitch period, or the GTE can be interpolated in length to match the new duration of the signal. Fig. 2.10 shows a diagram of the processing method to modify f0.

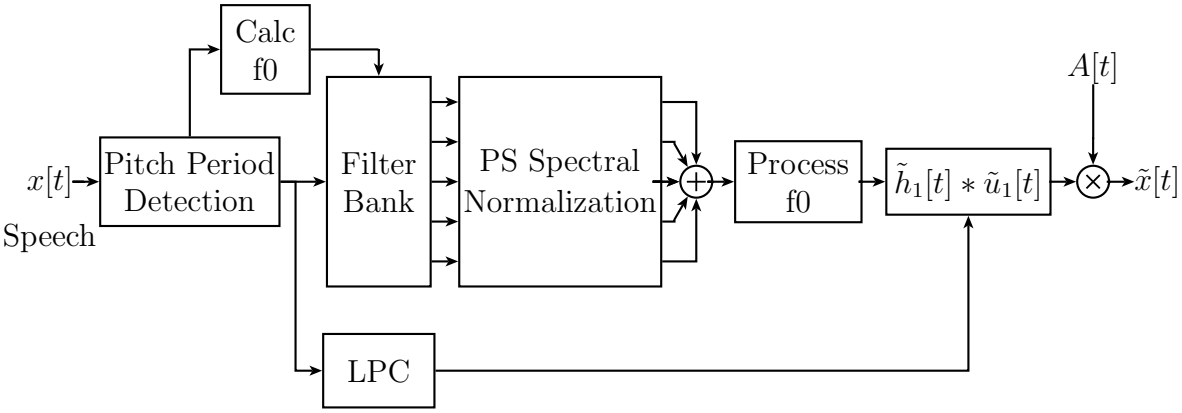

Figure 2.10: Diagram of Processing Method to Modify f0

One reason to process the source signal is to investigate the individual contribution of the fundamental frequency cue to speech recognition. For instance, each pitch period in a signal could be processed to have the exact same duration, eliminating the dynamic change of frequency of the source signal while maintaining the GTE and GSE. Fig. 2.11 shows an example of this processing. In this example, each pitch period the residual signal was upsampled or downsampled to have a duration of 5.6 *ms*, or a fundamental frequency of 180 *Hz*. The GSE in each pitch period of the original sound file was estimated using LPC. The sound file was spectrally flattened prior to performing the processing to change the signal periodicity. Finally, the LPC GSE and PS GTE were reapplied to each pitch period to create the processed stimuli.

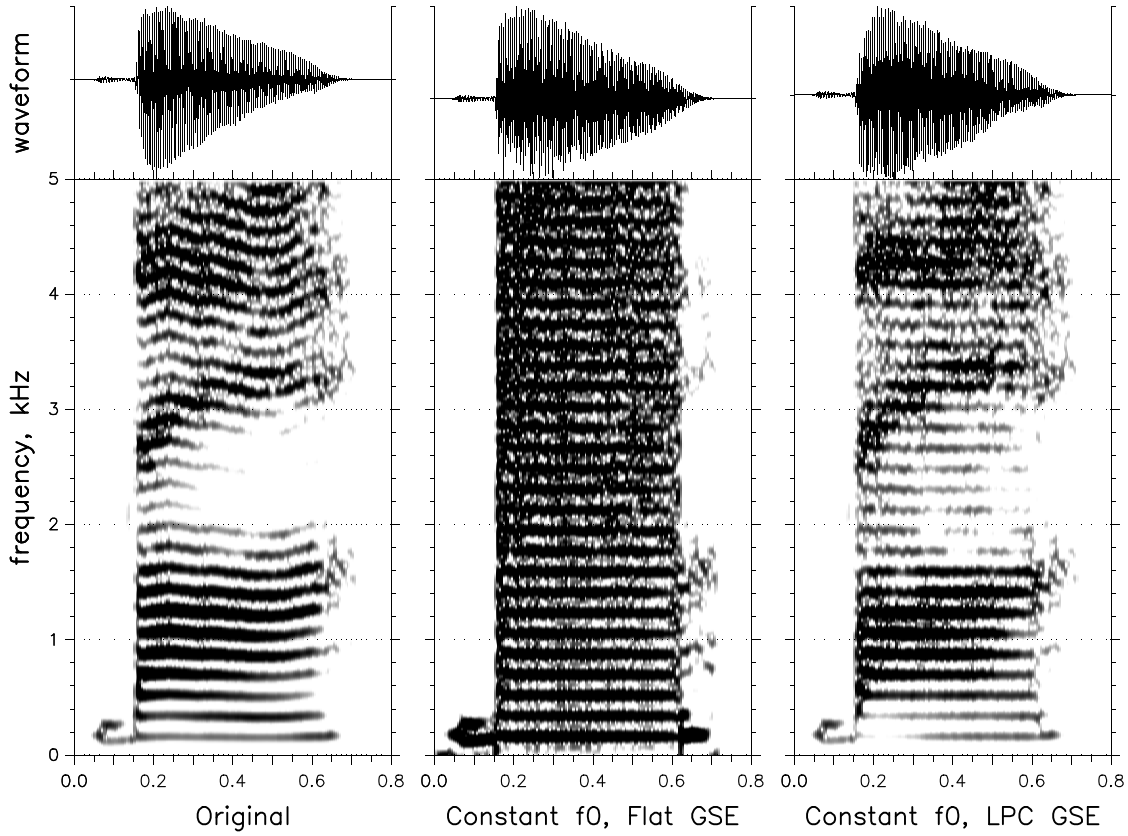

Figure 2.11: Processing the Signal Periodicity to Achieve a Constant f0, While Maintaining the GTE and GSE

2.3 Experiment

A listening experiment was conducted to compare the perception of PS processed speech signals to HT processed speech signals with a filter bank, similar to Auditory Chimeric speech [Smith, Delgutte, and Oxenham(2002)]. Experimental measures were perceived naturalness using a mean opinion score (MOS) on a five-point Likert scale and percent correct (PC) in repeating the provided speech token. Results for each measure were compared to unprocessed stimuli. The goal was to examine whether PS processing method provides a more veridical resynthesis of speech signals than the more commonly used HT method.

2.3.1 Subjects

In this experiment, six adult listeners between the ages of 20 and 25 years were tested. The number of listeners was divided evenly between males and females. It was an criterion of inclusion for each listener to report having normal hearing, speech and language. A hearing screenings of each listener was performed at the frequencies of 0.5, 1.0, 2.0, 4.0, and 6.0 kHz presented at 25 dB HL to each ear separately.

2.3.2 Equipment and Materials

Testing was performed in a sound isolation booth. A computer located in an adjacent room was used to present the stimuli. Stimuli were presented through a Soundblaster digital-to-analog converter and a Samson C-que 8 amplifier, over AKG-K141 headphones. A piece of paper with the numbers one through five with the word 'natural' above the number five and 'unnatural' above the number one was used to collect the MOS.

2.3.3 Stimuli

Speech tokens from the UCLA SPAPL Consonant Vowel (CV) database were used for this experiment. It is available online [Alwan(2013)]. Speech tokens from four different speakers, two male and two female, were used. Forty-eight consonant-vowel syllables were selected for this experiment, based on the pairing of 16 consonants (*b-, ch-, d-, f-, g-, j-, k-, m-, n-, p-, s-, sh-, t-, th-, v-, z-*) with each of 3 vowels (*-a, -ee, -oo*).

There were three stimulus conditions during testing: (1) an unprocessed (control) condition, in which the stimuli were presented as they were originally recorded, (2) a GTE interchanged condition using the PS method to recover the carrier (source + GSE) signal from one speaker, modulated in amplitude with the GTE from a different speaker, (3) a temporal envelope interchanged condition using the HT method to recover a carrier (instantaneous frequency) signal in each channel of a 64-channel ERB perfect-reconstruction filterbank, then amplitude modulate each carrier with a GTE in each channel from a different speaker. For consistency across all conditions, a low-pass filtered version of the half-wave rectified signal was used to create the GTE regardless of which method was used to recover the carrier signal. The cut-off frequency of the low-pass filter was 20 Hz. Linear interpolation was performed such that the length of the GTE was the same length as the carrier signal. Also, the modulating GTE was temporally shifted to maximize the correlation with the original GTE from the carrier signal.

Unvoiced portions at the beginning of tokens were processed differently from the voiced portions for the PS method. Unvoiced portions of speech signals are *aperiodic*, and cannot be temporally divided based on an underlying fundamental frequency.

Therefore, the entire unvoiced portion of the carrier signal was processed jointly by scaling the amplitude such that the RMS energy matched the unvoiced portion of the signal providing the GTE.

Two tokens of each of the 48 syllables were used from each speaker and further processed in the three conditions. The GTE was switched in a 'round-robin' method so that each speaker received an envelope from one other speaker for the transposed envelope conditions. In total, 48 syllables x 4 speakers x 3 conditions x 2 tokens were used to to create 1152 tokens. Fig. 2.12 shows examples of the unprocessed, PS, and HT 64-channel ERB perfect-reconstruction filterbank conditions from the experiment. The GTE from the unprocessed signal in the figure was overlaid on the carrier signal from the same utterance of a different speaker for the two processed conditions displayed. The output of processing with the PS method is a GTE shape similar to the unprocessed signal. Although each channel of the matches the amplitude envelope of the original signal, the wide-band GTE shape is not similar to the unprocessed signal and also displays greater noise across the spectrum.

2.3.4 Procedures

Listeners were instructed that they would be hearing speech signals, that some would sound like a natural voice, and that others would sound unnatural to different degrees. Listeners were asked to respond after the presentation of each individual stimulus by pointing to a number on the 'naturalness' rating scale and repeat what they heard. The 'naturalness' score was recorded for each token. If the consonant and vowel were repeated correctly, the token was recorded as correct. Otherwise the token was recorded as incorrect.

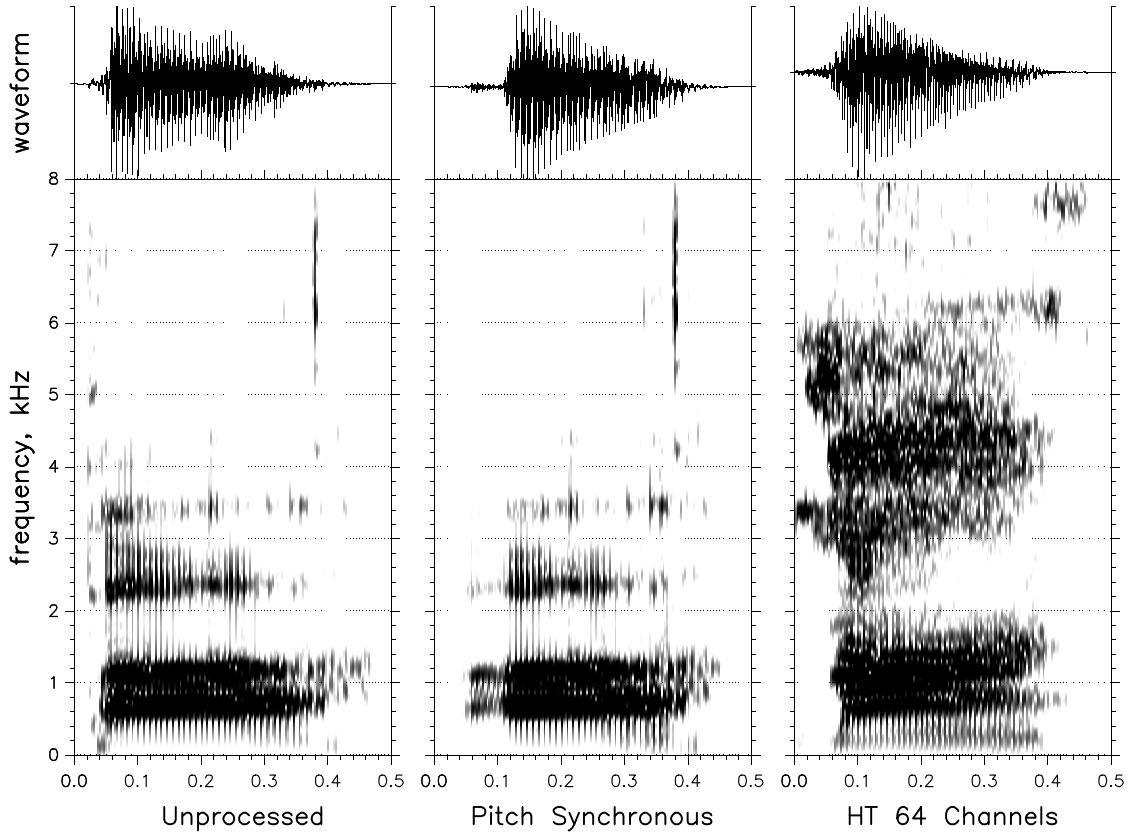

Figure 2.12: Stimuli Conditions for the Perceptual Experiment Comparing PS and HT Processing

Table 2.1: Naturalness Rating Scores (and standard deviations)

Condition	Unprocessed	Pitch Synchronous	HT 64-Channel Filterbank
Average	4.18	4.16	2.23
SD	0.32	0.30	0.25

Table 2.2: Percent Correct Recognition Scores (and standard deviations)

Condition	Unprocessed	Pitch Synchronous	HT 64-Channel Filterbank
Average	98.12 %	91.0 %	56.37 %
SD	1.05 %	2.48 %	5.73 %

Each stimulus was presented once to each subject. Order of stimulus presentation was randomized for each subject such that any condition and CV token could occur at any time.

2.3.5 Results

Table 1 shows the MOS across all participants for each condition. Table 2 shows the PC scores across all participants for each condition. For both measures, the results in the PS processed condition are similar to the unprocessed condition. However, poorer performance was demonstrated for the condition where the signal was decomposed using the HT.

2.4 Conclusions

A PS framework for representing acoustic characteristics of speech has been presented. It was motivated by the descriptions of temporal *envelope, fine structure,* and *periodicity* provided in [Rosen(1992)]. PS processing was compared with the HT method derived from an analytic signal of decomposing a signal into an envelope and instantaneous frequency. PS Temporal Amplitude Normalization was presented to recover and remove the GTE from a speech signal. This technique can be used to produce an amplitude envelope and phase signal pair that satisfies $x[t] = \mathbb{R}\{A[t]e^{j\phi[t]}\}$ while estimating aspects of speech production from a signal. PS processing satisfies the four physical conditions presented in [Loughlin and Tacer(1996)] for representing a signal as $A[t]$ and $\phi[t]$. PS Spectral Amplitude Normalization was presented to recover and remove the GSE from a speech signal. This technique can be used to produce a spectral envelope and source signal pair that satisfies $x[t] = A_1[t](h_1[t] * u_1[t])$. Methods were presented to replace each acoustic characteristic for voiced portions of speech.

There are several benefits of processing the temporal structure of speech using PS techniques. PS representations of a signal's GTE, GSE, and source require no additional processing such as low-pass filtering to be related to aspects of speech production. PS processing can be used to both recover and remove these aspects of speech production independently. Additionally, several methods were presented to replace each individual aspect without perceptually distorting the remaining aspects.

The trade-off of the benefits of PS processing are based on the assumptions. The PS GTE and PS GSE are assumed to be piece-wise constant functions. Similarly, the source signal is assumed to have a constant fundamental frequency during a pitch

period. These assumptions are not ecologically valid, but are necessary to achieve the desired signal decomposition.

There were several applications presented of PS processing that can be used to investigate the perception of the acoustic characteristics of speech production. In the next chapter, an application of PS processing for speech enhancement will be presented.

CHAPTER 3

Speech Enhancement Using Pitch Synchronous Processing

In the previous chapter, signal processing techniques were developed and described for the purpose of investigating the perceptual organization of human listeners. These techniques can be used to help answer questions related to how humans listen. In this chapter, signal processing techniques will be presented as an attempt to enhance the perception of speech.

3.1 Enhancing Speech by Processing the Gross Spectral Envelope

The perception of speech includes a spectral analysis of signals. The basilar membrane in the cochlea of the inner ear contributes to the spectral analysis. Hair cells located at different locations on the basilar membrane are stimulated by different frequencies. High-frequency sounds stimulate hair cells near the base of the cochlea, and low-frequency sounds stimulate hair cells near the apex. Therefore, the spectral analysis performed by the cochlea has been referred to as a frequency-place mapping because different places along the basilar membrane map different frequencies.

The auditory system perceives a speech signal produced by the human vocal tract. The resonances of the vocal tract emphasize different frequencies that are

present in the signal produced by the larynx. The resonances are commonly referred to as formants. The frequencies of formants distinguish between different parts of language. For instance, the stead-state frequencies of formants indicate the different vowels. Similarly, the transitions or change in formant frequencies indicate different consonants. The configuration of a speaker's vocal tract not only produces resonances at particular frequencies, but another characteristic of each resonance is its bandwidth. The bandwidth is determined by how broad or narrow a resonance is, and is conventionally defined as the frequency in which the amplitude of the resonance is 3 dB less than the amplitude at the peak frequency.

It has been the conventional theory that speech perception involves the perception of individual formants. As a result, resonances of the vowel tract can be labeled as separate formants (i.e. F1, F2, and F3). In this theory of speech perception, listeners must detect the frequency of individual formants and map them to a vowel space. One interpretation of this theory involves visualizing the vowel quadrilateral relating F1 and F2 frequencies as high-low and front-back vowels [Peterson and Barney(1952)].

An alternative method of speech perception is that listeners perceive the entire spectral envelope as a single spectral component, and do not perceive separate formants independently. Therefore, vowels and consonants are identified by the gross spectral shape or envelope of the speech signal. This theory has been supported previously, and used to develop a model of vowel perception by analyzing the auditory spectrum shape as a whole, rather than by identifying specific auditory features [Bladon and Lindblom(1981)]. It has also been demonstrated that automatic vowel classification based on gross spectral shape features was superior to that based on formants [Zahorian and Jalali Jagharghi (1993)]. For these reasons, it was was

concluded that the gross spectral shape provides a more complete set of acoustic correlates for vowel perception than formants. Further evidence to support this theory will be presented in Chapter 5.

If the GSE is to be modeled by a single gross spectral shape rather than multiple spectral resonances, it is not appropriate to describe the bandwidth of individual formants. Rather, it is appropriate to describe the peakedness of the GSE in terms of overall flatness and sharpness. The difference in amplitude between the peaks and troughs of the spectral shape indicate how flat or sharp the resonances are occurring. This is related to considering how broad or narrow a resonance is, but rather than describing the characteristic of the resonance as a measurement of frequency based on a fixed amplitude, the characteristic of the resonance is described as a measurement of amplitude independent of frequency across the spectrum.

The perceptual theory of a single spectral envelope can be used to motivate signal processing methods to enhance speech. Rather than pursuing methods to process individual formants, signal processing techniques can be developed to process the entire spectral envelope as a single component. Therefore, the speech signal should not be analyzed to recover individual formants, but rather analyzed to recover a single spectral shape.

It is desired to process the entire spectral filter without processing the signal that was produced by the larynx. Because the spectral filter has already been imposed on the source signal prior to the speech signal being captured for processing, it is necessary to perform a signal processing step to separate the spectral filter from the source signal. After this the spectral envelope can be processed independently of

the source signal. Finally, the processed spectral envelope can be used to filter the unprocessed source signal to create a new speech signal.

3.2 Reasons to Process the Gross Spectral Envelope

There are several reasons to consider processing the GSE as a method to enhance the speech signal for perception. First, listeners with sensorineural hearing loss can have broadened auditory filters. This was concluded because listeners with hearing loss show poorer frequency discrimination at high frequencies where sensitivity thresholds were diminished, but also at low frequencies where sensitivity thresholds were normal [Turner and Nelson(1982)]. Sharpening the spectral envelope of speech could compensate for the decreased spectral resolution for these listeners. Second, when children produce speech, their vocal tract resonances are not as sharp as adult speakers. Processing the spectral envelope of children's speech would be one method to make it resemble adult's speech more closely. Thirdly, when children perceive speech, they weight the dynamic spectral structure of speech as being significant. It was found that children perceptually weight the spectral skeletons in their native language more than the temporal amplitude envelope [Nittrouer, Lowestein, and Packer(2009)]. Sharpening the spectral envelope may help children perceptually attend to the characteristics of a speech that are most important. Finally, when noise interferes with a speech signal, it is more difficult for listeners to perceive the speech signal. Wide-band noise makes it more difficult to perceive speech in part because the peaks and troughs are less pronounced. Sharpening the spectral envelope could be a way to help listeners perceive speech in the presence of noise.

3.3 Previous Methods of Processing the Gross Spectral Envelope

Varying the bandwidth of formants, or the peakedness of the spectral envelope, has been done previously to investigate the impact on perception. Methods of processing have involved both synthetic speech and naturally recorded speech. For synthetic speech, the bandwidth of the formants could be broadened and narrowed by varying synthesis parameters. For naturally recorded speech, the signal was decomposed into a spectral filter and source signal. The spectral filter was processed, and a new speech signal was resynthesized using the source signal and new spectral filter.

Previous work has shown that normal hearing adult listeners and hearing impaired adult listeners with hearing aids have no significant difference in speech intelligibility when formants are narrowed or broadened under conditions of masking [Baer, Moore, and Gatehouse (1993)]. However, other studies have shown that adult recipients of cochlear implants show improved speech intelligibility for synthetic vowels where the formants have been narrowed [Hawks et al. (1997)]. Changing the lowest formant (F1) had the largest and most consistent impact on intelligibility as a result of changing the electrode activation patterns. It has also been hypothesized [Nittrouer and Lowenstein(2010)] that children listeners with hearing impairment, regardless of amplification method, would benefit from formant sharpening because all children listeners pay closer attention to formant transitions than adults when forming a phonetic object during speech perception.

3.3.1 Synthetic Speech

Turner and Van Tasell (1984) investigated the minimum detectable depth of a spectral "notch" between the second (F2) and third (F3) formants of a synthetic vowel-like stimulus. The purpose of the study was to investigate whether the impairment of auditory frequency resolution in sensorineural hearing loss influenced the perception of vowels. In this experiment, synthetic vowels were created to resemble the spectrum of the vowel /ae/. For this vowel, F2 was 1.8 kHz and F3 was 2.4 kHz. A linear-slope spectral notch was centered at 2.12 kHz, and used to shape the amplitude of the harmonics between F2 and F3 at 1.92, 2.04, 2.16, and 2.28 kHz. The notch depth was varied between 0 and 6 dB in increments of 1 dB. This effectively changed the bandwidth of the F2 and F3 formants. Subjects with both normal hearing and impaired hearing were able to discriminate smaller spectral notches than those found in the acoustic spectra of actual vowels. Therefore, it was concluded that auditory bandwidth impairment may not be critical in the perception of vowels.

Although these results could be interpreted to suggest that changing the bandwidth of formants has little perceptual significance, there are several reasons to continue investigating processing the GSE for listeners with sensorineural hearing loss. First, listeners with sensorineural hearing loss were, in fact, sensitive to the depth of the notch. Therefore, changing the sharpness of the spectral envelope would likely be something to which listeners would be sensitive. Second, a discrimination task with steady-state synthetic speech does not test the complexity of listening to naturally produce speech in the presence of real-world background noise. Rather, testing the sensitivity to 'notch' depth for naturally produced speech may differentiate listeners with normal hearing more from listeners with hearing loss. Third, testing a

listener's sensitivity to the notch depth for formants that are close in frequency may not be appropriate. It has been shown that listener's show Center Of Gravity (COG) effects when spectral prominences of close in frequency. In other words, two spectral prominences that are close in frequency can be perceived as a single spectral prominence. The COG effects have been found for spectral prominences as distant as 3.5 Barks [Speeter Beddor and Hawkins(1991)]. The Bark scale is a psychoacoustical scale proposed in [Zwicker(1961)] to represent the first 24 critical bands of hearing [Scharf(1970)]. The frequencies of 1.8 and 2.4 kHz are 11.6 and 14.2 Bark. Therefore, the difference between the spectral prominences is 2.6 Bark, meaning the formants in the experiment are likely on contributing to the perception of a single spectral prominence. Finally, this experiment only examines a listener's ability to discriminate a single 'notch' in the spectrum. Given the support for vowel perception based on recognizing a single broad spectral shape rather than individual formants, it would be more appropriate to consider processing the depth of all the troughs across the entire spectrum.

3.3.2 Naturally Recorded Speech

A method to process the resonances of the GSE across the entire spectrum for naturally recorded speech has been proposed previously [Tarr(2010)]. In order to sharpened or flattened the GSE of a speech signal, the spectral filter needs to be separated from the source signal. A common method of source-filter separation that allows for subsequent re-synthesis is linear predictive coding (LPC). The following is a description of the process to obtain a speech signal that has a sharpened or flattened GSE.

LPC can used to represent the spectral magnitude of speech [O'Shaughnessy(1987), Childers(1987)]. LPC estimates a current speech sample, based on a linear combination of a specified number of 'p' previous samples called poles. The model for LPC analysis consists of an excitation source, $u[t]$, which provides the input to a spectral shaping filter, $h[t]$. Convolving the output speech, $\hat{x}[t]$. Constraints are given to $u[t]]$ and $h[t]$ such that $\hat{x}[t]$ is as close as possible to the original speech signal $x[t]$. The excitation source, $u[t]$, is chosen to have a flat spectral envelope so that the relevant spectral detail is contained in $h[t]$.

The source signal, $u[t]$ and spectral filter, $h[t]$, are assumed to be stationary for a windowed time frame of 'N' samples and assumed to be stationary. Therefore, $h[t]$ can be modeled with constant coefficients, which are updated in each frame. For greater simplicity, $h[t]$ can be modeled as an all-pole filter, which is typically an adequate representation for the vocal tract [O'Shaughnessy(1987)].

LPC Method

The relationship between the speech signal with the source signal and spectral filter is, $\hat{x} = u[t] * h[t]$, or spectrally as $\hat{X}[z] = U[z]H[z]$. The estimation of the source signal and spectral filter from a speech signal is as follows:

$$H[z] = \frac{\hat{X}[z]}{U[z]} = G\frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.1}$$

An inverse or predictor filter, $A[z]$, can be used on speech, $s(n)$, resulting in an error or residual signal, $e(n)$.

$$A[z] = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{3.2}$$

$$e[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k] \tag{3.3}$$

A least squares method can be applied to calculate the a_k poles to minimize the energy in the residual signal.

$$E = \sum_{n=-\infty}^{\infty} e[n]^2 = \sum_{n=-\infty}^{\infty} [x[n] - \sum_{k=1}^{p} a_k s[n-k]]^2 \tag{3.4}$$

Setting the derivative of the energy with respect to a_k equal to zero and recognizing the autocorrelation function results in the minimum residual energy, also called the prediction error.

$$\frac{\partial E}{\partial a_k} = 0 = R[i] = \sum_{k=1}^{p} a_k R[i-k] \tag{3.5}$$

$$E_p = R[0] - \sum_{k=1}^{p} a_k R[k] \tag{3.6}$$

Now that the poles, a_k, have been found, the filter, $H[z]$, can be calculated and used to estimate the source, $U[z]$, by 'inverse filtering' the signal estimate, $\hat{X}[z]$ After the filter, $H[z]$, has been isolated, it can be sharpened or flattened. A new speech signal, $\hat{X}[z]$ can be created from the source signal and the processed spectral filter.

In order to sharpen or flatten the spectral filter, $H[z]$, the original LPC coefficients, $a_{k,i}$, for each frame, i, should be multiplied by some constant, b^k. This performs a 'radial scaling' over the poles. As an example, $a_{1,i}$, would be multiplied by b^1. Iteratively, $a_{2,i}$, would be multiplied by b^2, and $a_{3,i}$, would be multiplied by

b^3, etc. The value of the constant b can be chosen such that the bandwidth of the spectral resonances are processed on a frequency scale.

The relationship between a desired change in frequency (Hz) and the amount 'b' to scale the poles is as follows. The desired change in Hertz ($BWexp$) should first be related to a normalized change in bandwidth (ΔBW) depending on the sampling rate (Fs).

$$\Delta BW = 10^{-7} \cdot Fs \cdot BWexp \tag{3.7}$$

A nonlinear scaling is performed to transform the normalized change in bandwidth (ΔBW) to radians.

$$x = \tan \frac{\Delta BW}{2} \tag{3.8}$$

Finally, the amount 'b' to scale the poles is calculated by the ratio:

$$b = \frac{1 - x}{1 + x} \tag{3.9}$$

LPC Example

The following is a demonstration of the LPC Algorithm to broaden the peaks by 25 Hz. A sampling rate of 22050 samples/sec will be used for demonstration. An original signal of speech is a man's voice uttering the sentence: 'Find girls these clouds' will be used. The waveform is shown in Fig. 3.1. This is a typical sentence for testing recognition threshold as proposed in [Nittrouer and Lowenstein(2010)].

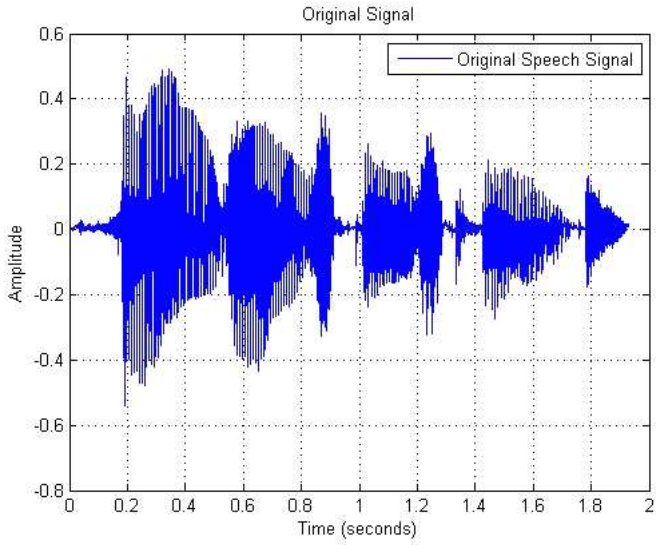

Figure 3.1: Original Speech Signal - 'Find girls these clouds'

The LPC coefficients are first estimated for each frame of time from the speech signal and used for inverse filtering. After inverse filtering, the input source, $U(t)$, is recovered. When the signal in a frame of time is designated as 'silence/noise,' the source will be set to zero. This will not impact the re-synthesized signal because for these frames, the input speech is passed to the output unprocessed; therefore the input source, $U(t)$, does not need to be calculated for these frames. The input (source) estimate for this speech signal is shown in Fig. 3.2.

After the source signal is recovered, it should be used to re-synthesize speech through filtering the source signal with the flattened filters. For comparison, the re-synthesized signal and the original speech have been plotted together in Fig. 3.3 to display how the formant flattening impacts the waveform.

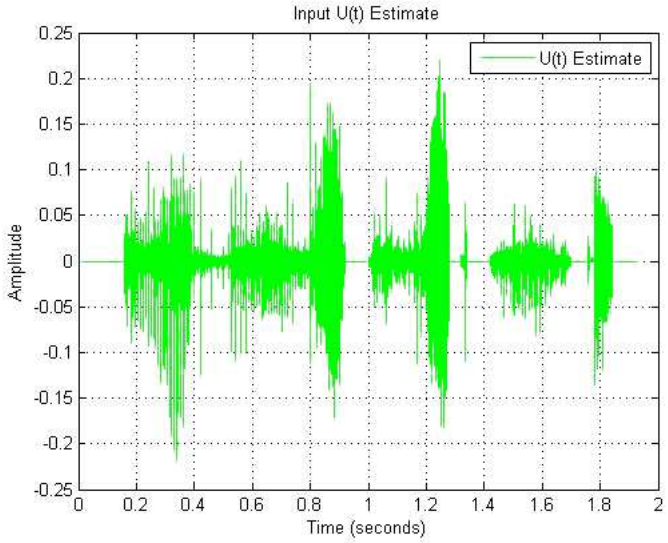

Figure 3.2: Input Source Estimate using LPC coefficients

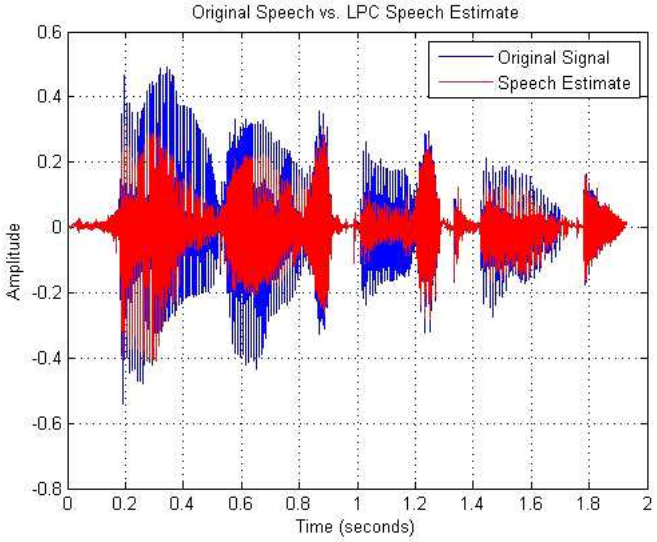

Figure 3.3: Comparison of Original Speech and LPC Speech Estimate

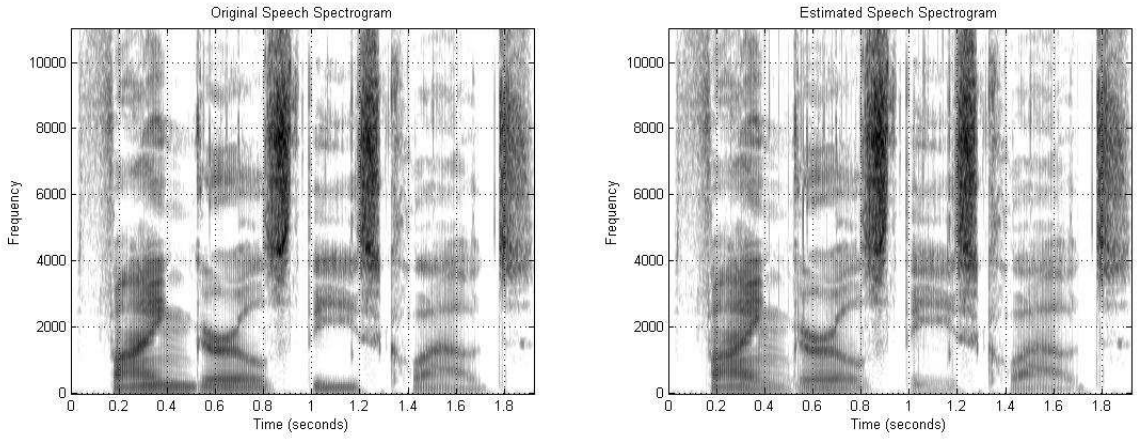

Figure 3.4: Spectrograms of Original Speech and Estimated Broadened Speech

A plot of the spectrogram of the 'Original Speech' and the 'Processed Estimated Speech' in Fig. 3.4 show how the formants have been flattened.

To further demonstrate the effectiveness of the broadening algorithm, a single frame is plotted in Fig. 3.5 to show how the formants have been flattened.

3.3.3 Issues with Using LPC

This algorithm to process the GSE of speech using LPC has several problems when applied to natural speech. First, when sharpening the spectral filter, it is possible for the system to become unstable. The radial scaling of the filter coefficients is the equivalent of increasing the magnitude of the poles in the z-plane. If the magnitude of the poles is greater than one, the system will be temporarily unstable. The consequence of system instability can result in large changes in amplitude in the re-synthesized speech signal. Stability constraints can be imposed when processing the filter coefficients, however this will limit the effectiveness of the algorithm. Second,

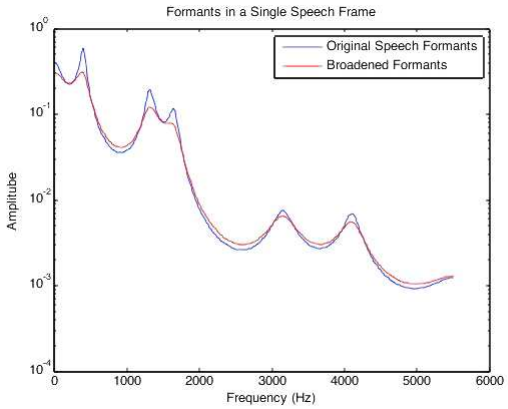

Figure 3.5: Original and Broadened Formants in a Single Speech Frame

even if stability constraints are imposed on the filter, it is possible to sharpen the spectral resonances beyond what is typical of speech. Perceptually, this may cause 'ringing' in the re-synthesize speech signal and cause it to sound unnatural. Finally, when the GSE is processed using this LPC algorithm, the GTE will likely be distorted during processing. While attempting to investigate the merits of processing the GSE, it is important to keep many other acoustic characteristics of the signal unchanged.

For these reasons, a new method of sharpening and flattening the GSE of speech was developed such that the GSE is processed independently of the GTE and source signal.

3.4 Pitch Synchronous Processing of the GSE

In Ch. 2, PS Spectral Normalization was discussed as a method to process the GSE separately of the source signal and GTE. Several of the same processing steps will be applied in this method to shape the resonances of the GSE. The important difference is that the harmonics of the source signal will not be spectrally normalized.

Rather, the harmonics will be processed relative to an estimation of the spectral roll-off of the source signal in each pitch period. This is necessary aspect of the processing in order to maintain the spectral roll-off characteristic of the source signal, while processing on the spectral resonances.

3.4.1 Pitch Synchronous Processing Method

The following is a description of a method using PS processing that can be used to change the shape of the GSE's resonances. This method is based on flattening or sharpening the resonances of the GSE. The GTE envelope of the original signal is maintained post-processing. Similarly, several characteristics of the source signal including the frequency of f0 and the spectral roll-off are maintained post-processing.

The first step in this method to process the signal in a PS manner is to detect the locations of the beginning and ending of each pitch period. Then the signal can be divided into separate pitch periods for processing based on these temporal locations. A method for detecting the location of the beginning and ending of each pitch period was presented in section 2.2.2.

After each pitch period has been separated, f0 should be estimated as the inverse of the length of the pitch period. Next, the signal of the pitch period can be duplicated and concatenated to create a quasi-periodic version of the original signal. Then, a filterbank can be created based on the frequency of f0 with channel frequency divisions of $n \cdot f0 + \frac{f0}{2}$ for $n = 1, 2, 3, ...$, resulting in cut-off frequencies below the Nyquist frequency. Next, the quasi-period signal can be separated into individual channels by the filterbank, with one harmonic of the signal in each channel. The quasi-period

version of the signal was created to allow for higher order filtering to decrease crosstalk between filterbank channels.

When the harmonics of the signal have been separated into individual channels, the amplitude of each channel can be measured to create a PS GSE as presented in section 2.2.4. Next, the spectral roll-off of the source signal can be estimated from the PS GSE. This is performed such that the spectral roll-off of the processed signal will closely match the spectral roll-off of the original signal.

It was found to be more convenient to estimate the spectral roll-off of the source signal by transforming the GSE to the logarithmic domain rather than using the linear domain. By using the logarithmic domain, the spectral roll-off of the source signal can be reasonably estimated using linear or quadratic regression. When the spectral roll-off of the source signal was estimated in the linear domain, it was necessary to use cubic or higher-order regression. In many cases, these estimations produced at GSE with negative values which is not appropriate. Fig. 3.6 shows the estimate of the source signal spectral roll-off in the logarithmic domain and the estimate transformed to the linear domain.

The amplitude of each harmonic could be scaled to the amplitude of the estimated spectral roll-off of the source signal. This would result in an estimate of the source signal independent of the GSE, or in other words, a signal where the resonances of the GSE have been maximally flattened. Fig. 3.7 show the estimate of the source signal by scaling the amplitude of each harmonic to match the estimated spectral roll-off.

Typically, it is not desired to maximally flatten the spectrum of the speech signal, but rather partially flatten the GSE. Therefore, after the spectral roll-off of

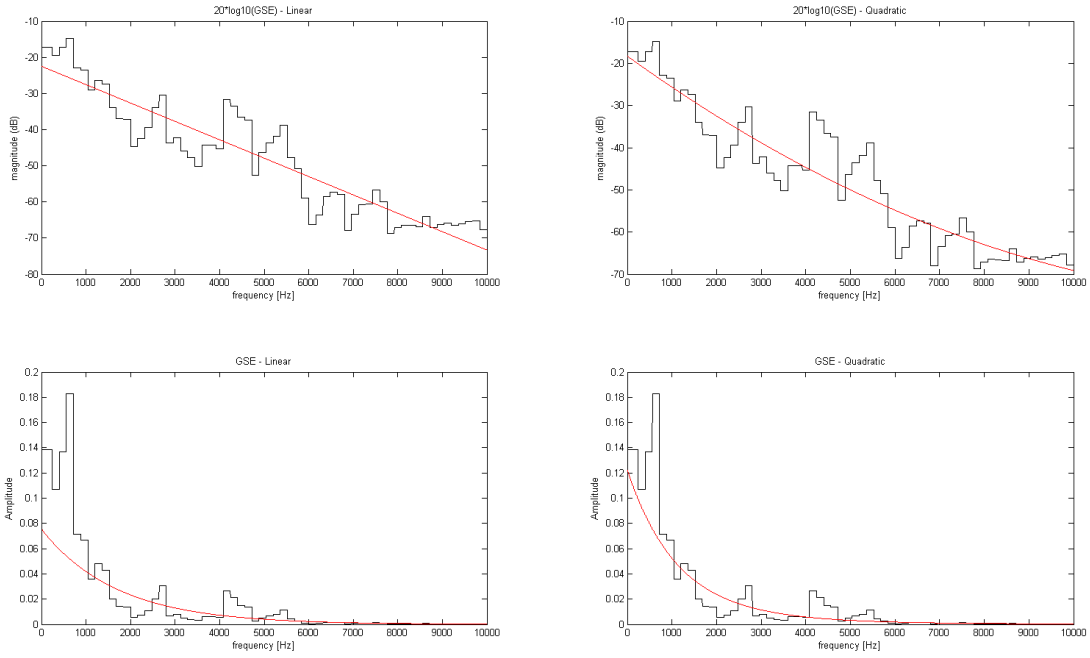

Figure 3.6: Fitting a Line to Estimate the Spectral Roll-off of the Source Signal

the source signal has been estimated in the logarithmic domain, the difference can be calculated between the estimated spectral roll-off and the amplitude of the GSE in each channel of the filterbank. This calculated difference indicates the resonances of the vocal tract. If the amplitude of the channel is greater than the estimated spectral roll-off, then it is assumed that this channel is contributing to a peak in the GSE. Conversely, if the amplitude of the channel is less than the estimated spectral roll-off, then it is assumed that this channel is contributing to a trough in the GSE.

In order to flatten the GSE relative to the estimated spectral roll-off, the difference or excursion between the GSE and estimated spectral roll-off of the source signal is decreased by a scaled amount. The 'peak' channels where the amplitude

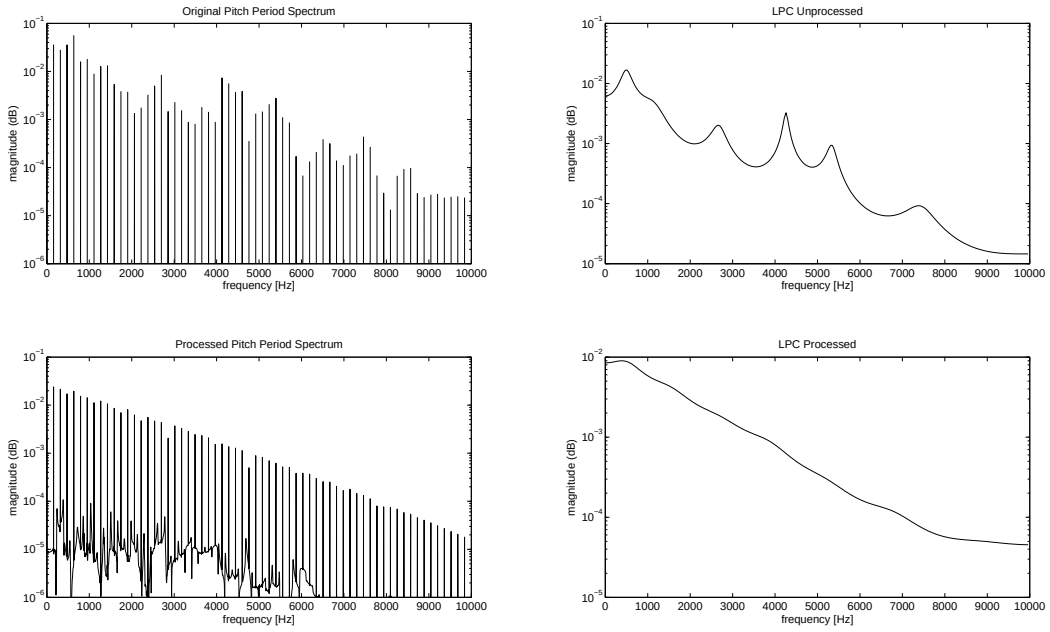

Figure 3.7: Estimate of the Source Signal Independent of the GSE

is greater than the estimated spectral roll-off of the source signal can be decreased in amplitude. The 'trough' channels where the amplitude is less than the estimated spectral roll-off of the source signal can be increased in amplitude. Essentially, the opposite is performed in order to sharpen the GSE.

As an example, suppose it is desired to flatten the GSE. The amount of how much to flatten the GSE can be selected such that the resonances of the GSE are half as sharp (or any proportion of sharpness). In other words, the differences or excursions between the GSE and estimate spectral roll-off of the source signal should be made to be half as much. The amount of change in amplitude in each channel, $\Delta[k]$ in each channel, k, can be calculated based on the initial amplitude in the channel, $a[k]$, the amplitude of the estimated spectral roll-off of the source signal, $s[k]$, and

the desired proportion of sharpness, p, with $\Delta a[k] = (1-p)(a[k] - s[k])$. The desired amplitude of the channel, $\tilde{a}[k]$ can then be calculated, $\tilde{a}[k] = a[k] - \Delta a[k]$. This method can be used to change the sharpness by proportions less than 1 to flatten the GSE or by values greater than 1 to sharpen the GSE. Fig. 3.8 shows an example of flattening the GSE. LPC spectral estimates are displayed to compare the LPC GSE in the unprocessed and processed signals. The PS GSE for the unprocessed and processed signals are displayed to compare the amplitude in each channel of the filterbank.

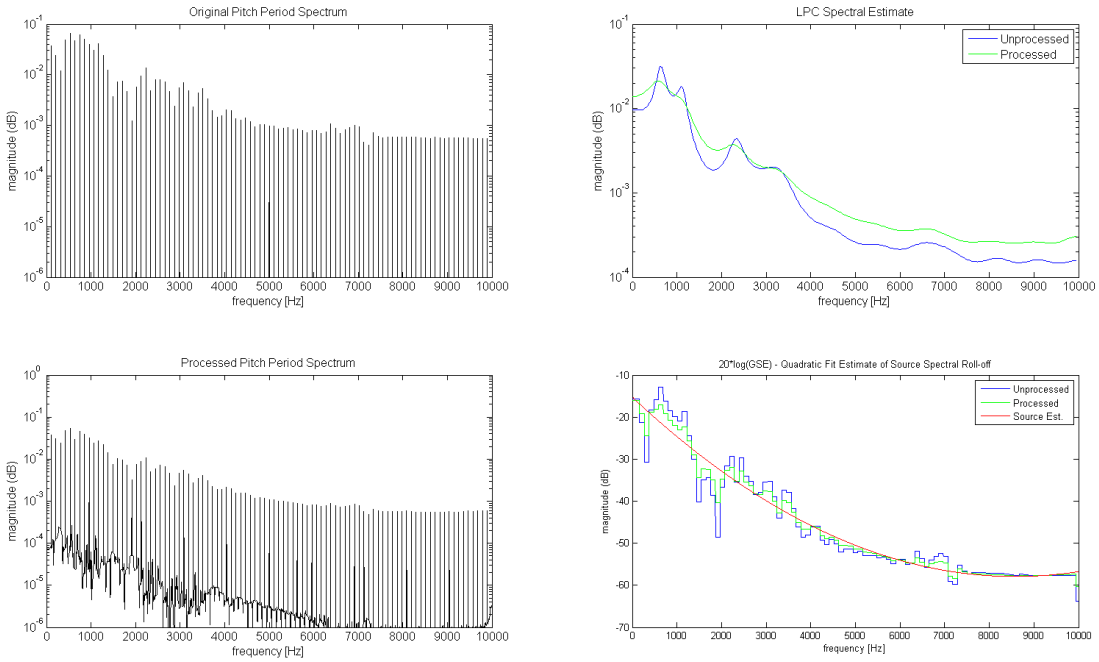

Figure 3.8: Decreasing the Difference Between the Amplitude of each Filterbank Channel Relative to the Source Signal Spectral Roll-off

Finally, after all harmonics have been scaled in amplitude to the appropriate values, the channels of the filterbank can be recombined. The overall amplitude of the GTE from the original signal in this pitch period can be reapplied to ensure it is maintain post-processing. This process can then repeated for the remaining pitch periods in a speech utterance. Fig. 3.9 shows the waveform and spectrogram of an unprocessed and flattened speech signal. The GTE is is maintain after processing. Spectrally, the resonances are not as pronounced.

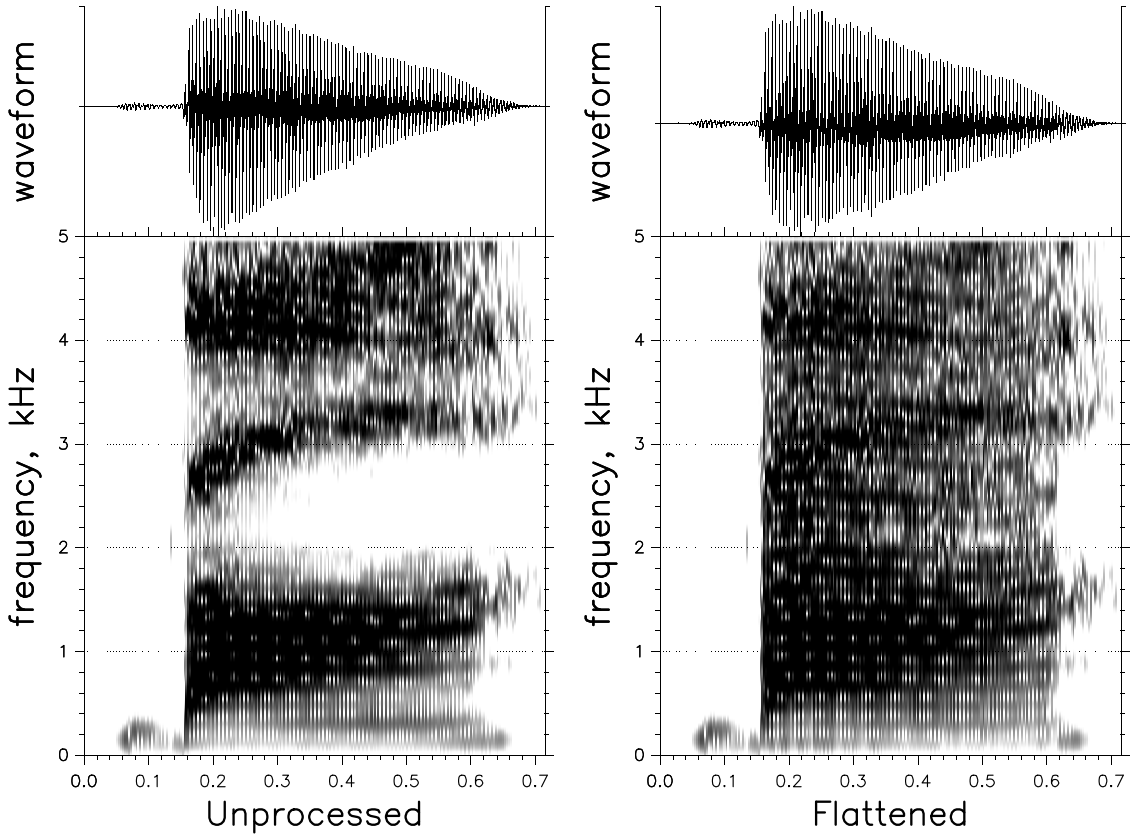

Figure 3.9: Spectrogram Comparison - Unprocessed and Flattened GSE

A similar procedure can be used when sharpening the GSE. Fig. 3.10 shows an example of sharpening the GSE. LPC spectral estimates are displayed to compare the LPC GSE in the unprocessed and processed signals. The PS GSE for the unprocessed and processed signals are displayed to compare the amplitude in each channel of the filterbank.

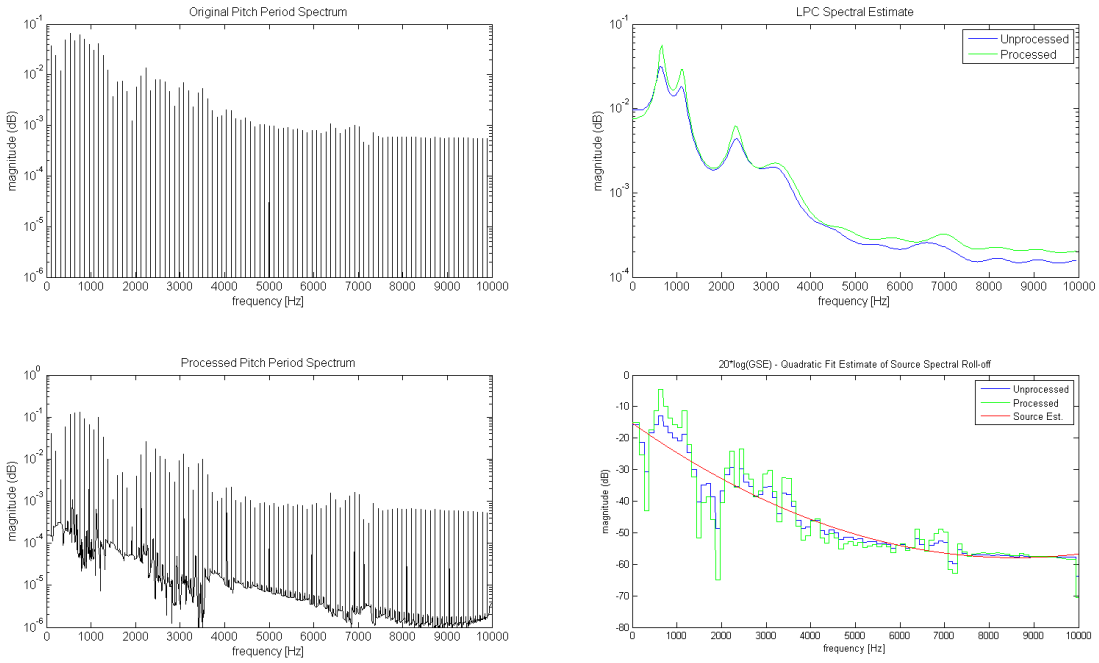

Figure 3.10: Increasing the Difference Between the Amplitude of each Filterbank Channel Relative to the Source Signal Spectral Roll-off

Fig. 3.11 shows the waveform and spectrogram of an unprocessed and sharpened speech signal. The GTE is is maintain after processing. Spectrally, the resonances are more pronounced.

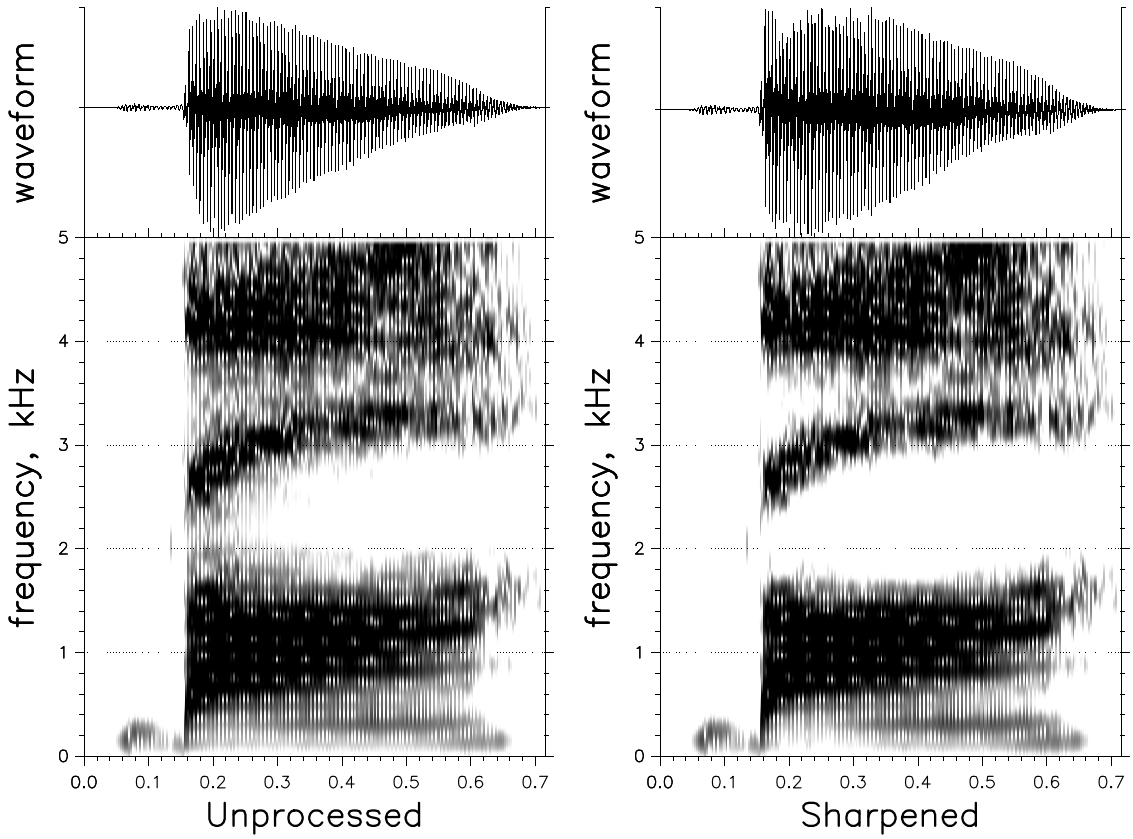

Figure 3.11: Spectrogram Comparison - Unprocessed and Sharpened GSE

CHAPTER 4

The Perceptual Importance of the Gross Temporal Envelope

In previous chapters, signal processing techniques were presented to represent speech as a gross temporal envelope, gross spectral envelope, and source signal. By representing these characteristics independently, inspection and manipulation can be performed to investigate the perceptual importance of each in different listening situations. In this chapter, the perceptual importance of the gross temporal envelope will be compared with the gross spectral envelope for listeners with normal hearing and listeners with cochlear implants (CI). It is hypothesized that these groups of listeners may not use the same perceptual weighting strategies related to the gross temporal envelope and the gross spectral envelope. This difference in perceptual weighting may explain differences in speech recognition for these listening groups. Additionally, understanding the perceptual weighting strategies of listeners with CI may lead to opportunities to improve perceptual outcomes.

4.1 The Perceptual Importance of the Gross Temporal Envelope for Listeners with Normal Hearing

The gross temporal envelope, as an acoustic characteristic of a speech, provides perceptual cues to a listener. One perceptual cue based on the gross temporal envelope

is the amplitude rise time (ART) for a syllable. The gross spectral envelope, as a different acoustic characteristic of speech, provides different perceptual cues to a listener. One perceptual cues based on the gross spectral envelope is the rate of formant transition or formant rise time (FRT). Listeners can make use of these perceptual cues to make distinctions between different utterances. One example is the stop-glide distinction between the syllables /ba/ and /wa/. The syllable /ba/ has a fast ART and a fast FRT. The syllable /ba/ has a slow ART and a slow FRT. Initially it appears that these two cues are redundant. One question that could be asked is whether listeners -perceptually attend to both cues, or if they prefer to attend to one cue because the other is redundant. This is the notion of perceptual weighting.

In the first experiment, the perceptual weighting strategies of listeners with normal hearing with be analyzed using the /ba/-/wa/ distinction. Additionally, sensitivity to ART and FRT was analyzed separately form the /ba/-/wa/ distinction. It is hypothesized that the perceptual weighting strategies of listeners may not be simply explained by sensitivity.

Several studies have already examined perceptual weighting strategies for the stop-glide contrast. Nittrouer and Studdert-Kennedy (1986) processed the ART of natural /ba/ and /wa/ syllables such that tokens of each were given the structure of the other, while preserving original formant structures. Results showed that adults based their phonemic decisions almost entirely on the FRT. Walsh an Diehl (1991) used synthetic speech tokens and replicated the findings of Nittrouer and Studdert-Kennedy (1986). Similar methods to these previous experiment will be used in the current experiment. potential effects of f0 were examined by using both natural and synthetic stimuli. In this case, both natural and synthetic stimuli will be used.

4.2 Method

4.2.1 Listeners

Fifty participants in two age groups were included in the first experiment. Twenty adults between 18 and 40 years of age and 30 children ranging in age from 4 years; 3 months to 5 years; 11 months participated. All listeners (or their parents, in the case of children) reported normal speech, language, and hearing. All listeners passed hearing screenings consisting of the pure tones of .5, 1, 2, 4, and 6 kHz presented at 25 dB HL to each ear separately. Parents of the children reported that their children were free from significant histories of otitis media, defined as six or more episodes during the first three years of life.

Tests of language were administered to each age group as an inclusionary criterion. Children were given the Goldman Fristoe 2 Test of Articulation [Goldman and Fristoe(2000)] and were required to score at or better than the 30th percentile for their age in order to participate. Childrens scores ranged from the 30th percentile to greater than the 87th percentile, with a mean of the 59th percentile (SD = 18). Adults were given the reading subtest of the Wide Range Achievement Test 4 [Wilkinson and Robertson(2006)] and all demonstrated better than a 12th grade reading level.

4.2.2 Equipment and materials

Stimuli were recorded in a sound-isolating room using a Shure KSM studio microphone, a Tube MPStudio V3 amplifier, and an Echo Gina 3G digital audio converter, using Adobe Audition software. All testing took place in a sound-isolating booth, with the computer that controlled stimulus presentation in an adjacent room.

Hearing was screened with a Welch Allyn TM262 audiometer using TDH-39 headphones. Stimuli were stored on a computer and presented through a Creative Labs Soundblaster card, a Samson headphone amplifier, and AKG-K141 headphones. The experimenter recorded responses with a keyboard connected to the computer.

For the labeling tasks, two drawings (on 8 x 8 in. cards) were used to represent each response label: a picture of a baby for /ba/, and a picture of the ocean (water) for /wa/. The experimenter explained to the listener that they were being used to represent the response labels because babies babble by saying /ba/ and babies call water /wa/.

For the discrimination tasks, a cardboard response card with a line dividing it into two halves was used with all listeners during testing. On one half of the card was two black squares representing the same response choice. On the other half was one black square and one red circle representing the different response choice. Ten other cardboard cards were used for training with children. On six cards were two simple drawings, each of common objects (e.g., hat, flower, ball). On three of these cards the same object was drawn twice (identical in size and color) and on the other cards two different objects were drawn. On four cards were two drawings each of simple geometric shapes: two with the same shape in the same color and two with different shapes in different colors. These cards were used to ensure that all children knew the concepts of same and different.

4.2.3 Stimuli

Five sets of stimuli were created: one set of natural speech and four sets of synthetic stimuli. Two of the sets of synthetic speech were used for a labeling tasks, and two sets of nonspeech stimuli for discrimination tasks.

Natural speech stimuli

Three types of natural speech stimuli were created from recordings of /ba/ and /wa/ syllables: (1) syllables with original, unprocessed temporal envelopes; (2) syllables with temporal envelopes transposed within the same category of syllable (/ba/ onto /ba/ or /wa/ onto /wa/); and (3) syllables with temporal envelopes switched between the two types of syllable (/ba/ onto /wa/ and /wa/ onto /ba/).

A male speaker was recorded producing ten tokens each of /ba/ and /wa/. These tokens were digitized at a 44.1 kHz sampling rate with 16 bit resolution. Five tokens of each were selected which matched duration as closely as possible. Acoustic measurements were made of each token, using TF32 software [Milenkovic(2004)]. For these measurements, the vowel was defined as the whole vocalic portion of the syllable, from the release of closure for /ba/ or the release of constriction for /wa/. Six measurements were made. (1) F0 was measured for the first three pitch periods after vowel onset. (2) F1, F2, and F3 were measured for the first two pitch periods after vowel onset. (3) F1, F2, and F3 were measured at the start of vowel steady state, defined as the point where F2 no longer rose more than 10 Hz from one pitch period to the next, and was calculated using 26-pole LPC analysis at each pitch period. (4) Vowel duration was measured from vowel onset to the end of the vowel, which was defined as the zero crossing where energy from the formants higher than F1 was

fully attenuated. (5) Formant rise time (FRT) was measured from vowel onset to the start of the vowel steady state. (6) Amplitude rise time (ART) was measured, using the following procedure. The amplitude peak of the syllable was found and RMS amplitude was calculated over the 5 pitch periods with the amplitude peak as the center, using WavEd software [Neely and Peters(1992)]. RMS was then computed for individual pitch periods preceding the amplitude peak. The first pitch period with an RMS value equal to or greater than 80% of the peak value was labeled as the end of the amplitude rise. ART was calculated as the duration between vowel onset and the end of amplitude rise.

To eliminate any additional distinctive cues between the tokens, prevoicing was removed from the /wa/ tokens and the burst was removed from the /ba/ tokens. These tokens will be referred to as unprocessed in this report. After these stimuli were created, the next step was to create the processed stimuli. The gross temporal envelope (GTE) was interchanged (either transposed or switched) between tokens using the three subsequent steps to create transposed and switched stimuli. The general technique of interchanging GTEs used pitch synchronous processing and was presented in 2.

First, the GTE was removed from the target token so that a new envelope could be applied without interaction from the envelope of that target. A process of pitch period normalization was used in which every individual sample in one pitch period was scaled by the same amount so the maximum amplitude peak across all pitch periods was uniform. To find the individual pitch periods of the target, autocorrelation functions with 40-ms windows starting at the beginning of each pitch period were used, in a three step process. (1) Measuring from the start of one pitch period, the

start of the next pitch period was estimated to be between 5.6 and 14.3 ms later, assuming f0 to be between 70 and 180 Hz, which is typical for a male speaker. (2) The peak of the autocorrelation function within that window was used to constrain further where the start of the next pitch period would be. (3) Then the nearest positive-going zero crossing was identified as the exact start. After every pitch period was identified, each was scaled separately.

Next, the GTE was extracted from a second, or model, token. Due to variations of f0 and utterance length across tokens, there could be no temporal alignment of individual pitch periods in the separate model tokens. Therefore, the GTE from each model token was measured by half-wave rectifying the original signal and low-pass filtering using a cutoff frequency of 20 Hz.

Third, the extracted GTE of the model was overlaid on the normalized target using a point-wise multiplication. Prior to the multiplication, the longer of the two signals was truncated at the end to match the length of the shorter signal. Because tokens had been selected to be similar in length, truncation only involved a pitch period or two, when necessary. After this was done, the ART of the target with the new envelope was measured using procedures already described to ensure that the envelopes of the transposed or switched tokens matched the envelopes of the tokens that served as models. For two tokens there were found to be slight deviations. In those cases, individual pitch periods were adjusted so the ART of the target matched the model ART precisely.

The transposed stimuli were created in round-robin fashion: /ba/ 1 was the model imposed on /ba/ 2 as the target; /ba/ 2 was the model imposed on /ba/ 3 as the target; and so forth. The same procedure was used with the /wa/ stimuli.

The switched stimuli were created so that /ba/ 1 was the model and /wa/ 1 was the target, and vice versa; /ba/ 2 was the model and /wa/ 2 was the target, and vice versa; and so forth.

These manipulations resulted in a total of 30 stimuli of six types: Five unprocessed /ba/ tokens; 5 transposed /ba/ tokens (with envelopes from different /ba/ tokens); 5 switched /ba/ tokens (with /wa/ envelopes); 5 unprocessed /wa/ tokens; 5 transposed /wa/ tokens (with envelopes from different /wa/ tokens); and 5 switched /wa/ tokens (with /ba/ envelopes).

Two tokens of each type were played during every block of testing, creating blocks of 12 stimuli. There were ten blocks presented. Each time the software selected a token of a particular type to play, it did so randomly, but did not replace that token to the pool until each token in the pool had been played. This process resulted in each stimulus being played a total of 4 times during testing, so 20 responses were collected to each type. In all, 120 stimuli were presented (6 types X 5 tokens X 4 repetitions).

Synthetic speech stimuli

There were two sets of synthetic speech stimuli created for a labeling task. Onset and steady state formant values averaged across the /ba/ and /wa/ tokens were used to determine formant values for a 9-step /ba/-/wa/ continuum created using a Klatt synthesizer (Sensyn). The tokens were 370 ms in duration. An f0 of 100 Hz was used throughout. Starting and steady-state frequencies of the first two formants were the same for all stimuli. The time to reach steady-state frequencies was varied between stimuli. F1 started at 450 Hz and rose to 760 Hz at steady state. F2 started at 800 Hz and rose to 1150 Hz at steady state. F3 was kept constant at 2400 Hz. FRT

varied along a 9-step continuum from 30 ms to 110 ms, in 10-ms steps. Matlab was used to overlay ART after synthesis using Sensyn. The envelope created by Matlab started at zero amplitude and rose to the maximum amplitude at the end of the rise time, which varied along a 7-step continuum from 10 ms to 70 ms in 10-ms steps.

The synthesis procedures described above were combined to make two sets of stimuli, one which varied in FRT and one which varied in ART. For the FRT stimulus set, the most /ba/-like ART (10 ms) and most /wa/-like ART (70 ms) were applied to each stimulus along the 9-step /ba/-/ba/ FRT continuum, resulting in 18 FRT stimuli (9 FRTs X 2 ARTs). For the ART stimulus set, stimuli were created with the most /ba/-like FRT (30 ms) and most /wa/-like FRT (110 ms). Then the seven ARTs were applied to each stimulus, resulting in 14 ART stimuli (2 FRTs X 7 ARTs). Fig. 4.1 shows synthetic stimuli for which FRT and ART signaled phonemic identity in a consistent manner. top). Fig. 4.2 shows stimuli for which they were set to signal phonemic identity in a contradictory manner. Inspection of the waveforms confirms that ART was implemented as described. During testing, stimuli were played 10 times each in blocks of however many stimuli there were, so that listeners heard a total of 180 FRT and 140 ART stimuli, in two separate conditions.

Nonspeech stimuli

Two sets of nonspeech stimuli were created for the discrimination task. One set that was more speech-like in quality and one set that was not speech-like at all. As in Miyawaki et al. (1975), it was hypothesized that listeners would be sensitive to ART when it is not heard as part of a speech signal, but fail to attend to it in making judgments about speech. Evidence for that position would be obtained if listeners

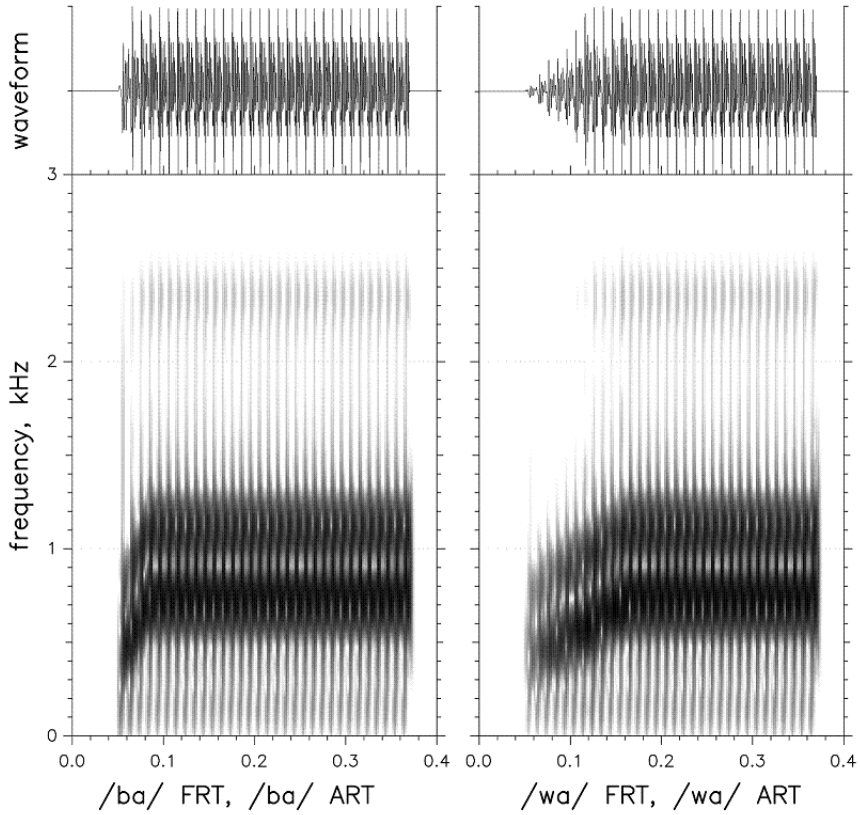

Figure 4.1: Synthetic Stimuli with Consistent FRT and ART Cues

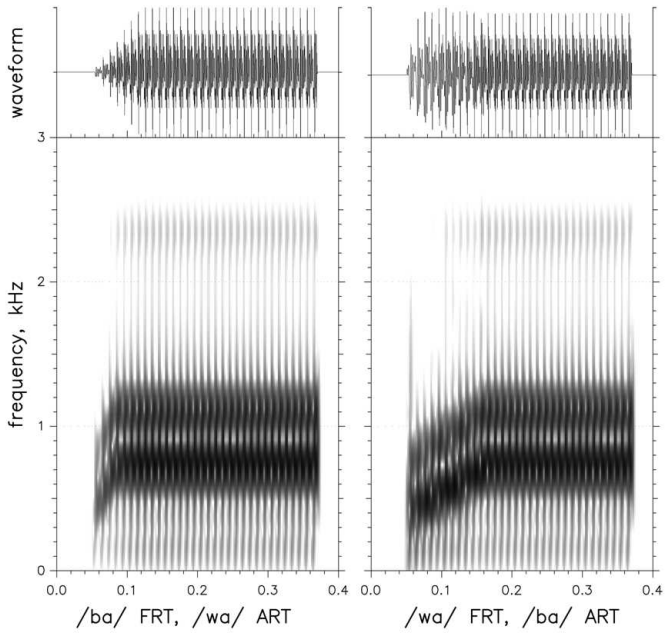

Figure 4.2: Synthetic Stimuli with Contradictory FRT and ART Cues

recognized smaller differences in ART with the completely nonspeech signals than they did with the more speech-like signals.

The more speech-like set of stimuli was synthesized with Sensyn, using steady-state formants. The frequencies were 500 Hz for F1, 1000 Hz for F2, and 1500 Hz for F3. These frequencies are typical values for modeling the resonances of a male vocal tract with a quarter wavelength resonator, although they do not represent any vowels in English. The f0 was 100 Hz. Total duration was 370 ms. Onset amplitude envelopes (similar to the envelopes used for the other synthetic speech stimuli) were overlaid on these signals. ART ranged from 0 to 250 ms in 25 ms steps, resulting in 11 stimuli. The other set of stimuli consisted of sine waves synthesized using Tone [Tice and Carrell(1997)]. Sine wave tones were substituted as the frequencies of the formants in the Synsen stimuli. The sine wave stimuli had the same duration

and ARTs as the Sensyn stimuli. This resulted in 11 stimuli. The stimulus with the 0-ms ART was always the standard (A), and every stimulus (including the standard) was played as the comparison (X). During testing, each stimulus was compared to the standard 10 times, so that listeners heard a total of 110 formant stimuli and 110 sine wave stimuli.

4.2.4 Procedures

During the testing session, the screening procedures (hearing screening and the WRAT or Goldman-Fristoe) were administered first. The five test conditions were ordered so that one of the synthetic labeling tasks was first, then one of the discrimination tasks. The labeling task with the natural stimuli was always presented third, followed by the other discrimination task and finally the other synthetic labeling task. As a result, there were four possible orders of presentation. Adults completed all tasks in the single session, and children completed the first three tasks in the first session and the last two tasks in the second session.

For the labeling tasks, the experimenter introduced each picture for the two tokens. Ten live-voice practice trials were presented in which the listener pointed to the picture and named it. Requiring the listener to point to the picture and say the syllable ensured that they were correctly associating the syllable and the picture. Then the listener heard the five unprocessed exemplars of the two tokens over headphones, and was instructed to respond in the same way. Listeners were required to respond to nine of the ten exemplars correctly to proceed to testing. For synthetic stimuli, training was provided using the most /ba/-like (30 ms FRT, 10 ms ART) and /wa/-like (110 ms FRT, 70 ms ART) stimuli. Listeners heard five

presentations of each endpoint, and had to respond to nine out of the ten correctly to proceed to testing. Feedback was provided for two rounds of training, and listeners were given up to two rounds without feedback to reach criterion. If listeners were not able to respond to nine of ten endpoints correctly by the third round, testing for that condition was not done. During testing in each of the three stimulus conditions, listeners needed to respond with 80% or better accuracy to the endpoints in order to have their data included in the analysis.

Different dependent measures were analyzed for the natural and synthetic stimuli. For natural stimuli, the percentage of stimuli of each type given the label of the original (target) syllable served as the dependent variable. Arcsin transformations were used for statistical analysis because results were close to 100%. For synthetic stimuli, the percentage of /wa/ responses across each continuum served as the dependent variables.

An AX procedure was used for the discrimination tasks. In this procedure, listeners compare a stimulus, which varies across trials (X), to a constant standard (A).The A stimulus was the one with a 0-ms ART for both the sine wave and formant stimuli. The interstimulus interval between standard and comparison was 450 ms. The listener responded by pointing to the picture of the two black squares and saying same if the stimuli were judged as being the same, and by pointing to the picture of the black square and the red circle and saying different if the stimuli were judged as being different. Both pointing and verbal responses were used because each served as a check on the reliability.

Children were required to recognize the drawings on the pictures as same or different prior to beginning the listening experiment. All listeners were presented with

five pairs of stimuli that were identical and five pairs of stimuli that were maximally different, in random order before testing with stimuli in each condition. Listeners were asked to report whether the stimuli were the same or different and were given feedback. Next, these same training stimuli were presented, and listeners were asked to report if they were the same or different, only without feedback. Listeners needed to respond correctly to nine of the ten training trials without feedback in order to proceed to testing. During testing in each of the two stimulus conditions, listeners needed to respond correctly to at least 16 of these physically same and maximally different stimuli (80%) to have their data included in the final analysis.

The discrimination functions of each listener formed cumulative normal distributions, and probit functions were fit to these distributions. From these fitted functions, distribution means were calculated and served as difference thresholds. These thresholds were the 50% points on the discrimination functions.

4.3 Results

All 20 adults met the training and testing criteria to have their data included in the study for all five tasks. All 30 children met the inclusion criteria for the labeling task with natural syllables. For the two labeling tasks with synthetic syllables, 27 of the 30 children met the inclusion criteria for the FRT continua and 29 of the 30 children met the inclusion criteria for the ART continua. Only 12 of the 30 children met the inclusion criteria for the formant stimuli in the discrimination task. Those 12 and an additional 5 children met the inclusion criteria for the sine wave stimuli.

Fig. 4.3 shows percent original responses for adults and children for the unprocessed, transposed, and switched natural stimuli. Adults and children responded to

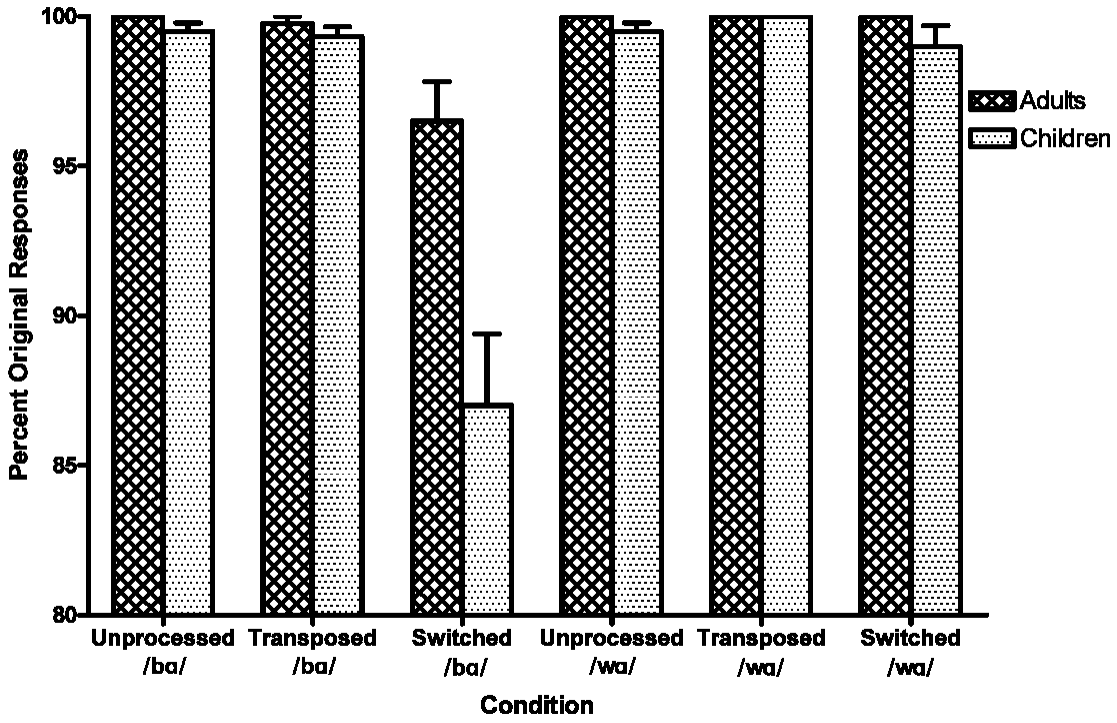

Figure 4.3: Percent Original Responses for Adults and Children for Unprocessed, Transposed, and Switched Stimuli

the unprocessed and transposed stimuli with the original response (/ba/ for /ba/ and /wa/ for /wa/) 99.33 to 100% of the time. The switched /wa/ stimuli (/wa/ with /ba/ envelopes) were also heard as /wa/ nearly all of the time: 100% for adults, and 99% for children. Switched /ba/ stimuli (/ba/ with /ba/ envelopes) showed many more /wa/ responses, especially from children. Two of the 20 adults (10%) and 13 of the 30 children (43%) responded to switched /ba/ with more than 10% /wa/ responses.

Table 4.1 and Table 4.2 show results from two-way ANOVAs with age as the between-subjects factor and stimulus type (unprocessed, transposed, or switched) as the within-subjects factor, for /ba/ and /wa/ stimuli separately. For stimuli created

Table 4.1: Statistical Outcomes of the ANOVA for Natural Unprocessed, Transformed, and Switched /ba/ stimuli

Original /ba/				
Effect	*df*	*F*	*p*	η^2
Stimulus type	2,96	30.28	$< .001$	.39
Age	1,48	9.77	.003	.17
Stimulus type x Age	2,96	5.71	.005	.11

from the original /ba/ syllables, the main effects of stimulus type and age were statistically significant, as was the Stimulus type x Age interaction. The largest amount of variance was explained by stimulus type (2 = .39). Because a significant interaction was found, one-way ANOVAs with age as the factor were done for each of the three stimulus types separately to locate the source of that interaction. Only switched /ba/ showed a significant effect: F(1, 48) = 9.14, p = .004. These statistical findings support the claim that the only natural stimuli that were not labeled entirely according to formant trajectories were the switched /ba/ syllables. Labeling of these stimuli was influenced slightly by ART, and that influence was greater for children than adults. Stimuli created from the original /wa/ syllables were labeled entirely according to formant trajectories, by adults and children alike.

Fig. 4.4 shows results for the FRT continua (top) and the ART continua (bottom) for adults (squares, solid lines) and children (circles, dashed lines). The results for the FRT continua show that adults and children responded similarly. When these labeling functions are compared to those in Fig. 4.5 from a sibilant-vowel study, responses for both groups resemble those of adults in that earlier study where functions are steep. This pattern indicates that listeners responded largely based on the cue

Table 4.2: Statistical Outcomes of the ANOVA for Natural Unprocessed, Transformed, and Switched /wa/ stimuli

Original /wa/				
Effect	*df*	*F*	*p*	η^2
Stimulus type	2,96	1.09	NS	
Age	1,48	3.33	.074	.07
Stimulus type x Age	2,96	1.09	NS	

manipulated along the continuum represented on the x axis. In this case, that was FRT. However, functions from this experiment are less separated based on the second cue (ART in the current experiment) than those of adults in Fig. 4.5. This pattern indicates that the second cue in this experiment was not weighted strongly.

Two-way ANOVAs were performed on the slopes and phoneme boundaries, with age and ART as main effects to more closely examine results for this FRT continuum. For slopes, the main effects of age and ART were not significant, but the Age x ART interaction was, $F(1, 45) = 4.71$, $p = .035$ ($\eta^2 = .10$). This outcome reflects the fact that adults labeling functions had similar slopes across ARTs, and childrens labeling function for the /ba/ ART had a similar slope to those of adults. However, childrens labeling function for the /wa/ ART (filled circles) is slightly shallower, and that likely accounts for the significant interaction. As with natural stimuli, children were somewhat affected by the /wa/ ART, even when stimuli had obvious /ba/ FRTs. For phoneme boundaries, the effect of ART was significant: $F(1, 45) = 17.74$, $p < .001$ ($\eta^2 = .28$), but neither age nor the Age x ART interaction was significant. These outcomes mean that listeners responded differently to stimuli with /ba/ and /wa/ ART, but the effect was similar for children and adults.

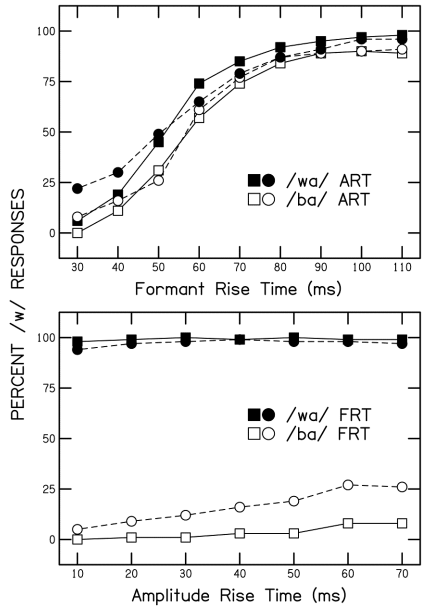

Figure 4.4: Results for the FRT and ART Continua for Adults and Children

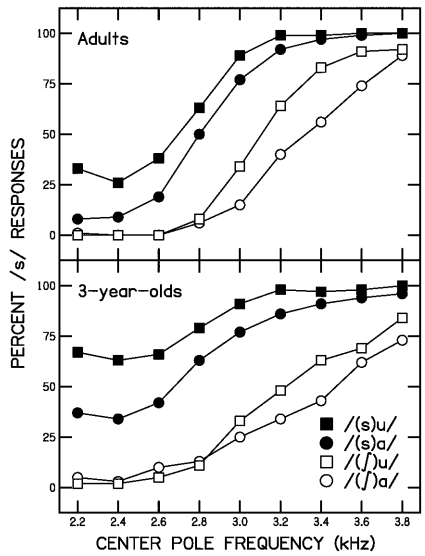

Figure 4.5: Labeling Functions from a Silbilant-Vowel Study

Results for the ART continua are shown at the bottom of Figure 4.4. Listeners appear to have assigned no weight at all to ART when the formants were /wa/-like (filled symbols). Adults also did not weight ART strongly when the formants were /ba/-like (open symbols), but children did to some degree. Around 25% of the stimuli with /ba/ FRT and the longest ART were labeled as /wa/ by children. This matches the result found for natural tokens where some children labeled some switched /ba/ syllables as /wa/, and what was found for the FRT continuum with the /wa/ ART, where children's function was shallower than others because some stimuli at the /ba/ FRT endpoint were labeled as /wa/.

The percentage of /wa/ responses across all steps on the ART continuum was computed for the /ba/ and /wa/ FRT continua separately. Arcsin transforms were used on the percentages. A two-way ANOVA was done with age as the between-subjects factor and FRT as the within-subjects factor. The effect of age was significant, $F(1, 47) = 6.87$, $p = .012$ ($\eta^2 = .13$), as was the effect of FRT, $F(1, 47) = 1591.26$, $p < .001$ ($\eta^2 = .97$). The Age x FRT interaction was also significant, $F(1, 47) = 25.93$, $p < .001$ ($\eta^2 = .36$). These results confirm that listeners in both groups weighted FRT heavily in their phonemic decisions, and that children showed some small weighting of ART, which adults did not do.

From the labeling results using natural and synthetic stimuli, it is clear that all listeners weighted FRT strongly in their stop-glide decisions. Neither adults nor children weighted ART strongly at all. The remaining question is whether listeners are sensitive to this acoustic structure or not, and is considered in the discrimination task.

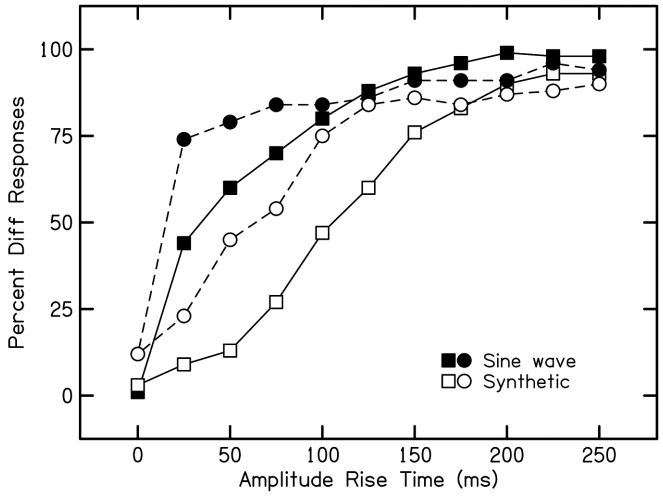

Figure 4.6: Discrimination Functions for Sine-wave and Synthetic Stimuli

The discrimination functions for sine wave (filled symbols) and synthetic (open symbols) stimuli for adults (squares) and children (circles) are shown in Fig. 4.6. It seems both adults and children were sensitive to ART. The children included in the data presented here represent only those children who could perform the task with these stimuli, and on average they appear to have been slightly more sensitive to ART than adults were. However, there were some children who were excluded because they did not meet the inclusionary criterion.

A two-way ANOVA with age as the between-subjects factor and stimulus type as the within-subjects factor was performed on distribution means from adults and the 12 children who met criteria for participation with both types of stimuli. It was found that the effect of age was significant, $F(1, 30) = 4.95$, $p = .034$ ($\eta^2 = .14$), as was the effect of stimulus type, $F(1, 30) = 48.63$, $p < .001$ ($\eta^2 = .62$). The Age x Stimulus type interaction was not significant. Thus, it can be concluded that listeners

were more sensitive to ART for sine wave stimuli than for formant stimuli. In other words, listeners were less sensitive to ART for speech-like sounds.

4.3.1 Discussion

The purpose of this study was to examine the labeling of stimuli in the stop-glide distinction by adults and children. Two objectives were investigated by the current study. The first objective was to examine whether adults and young children differ in terms of how they weight the acoustic cues to the stop-glide contrast. Outcomes revealed that both ages groups of listeners based their decisions about whether syllables started with stops or glides almost entirely on the rate of formant change. That similarity across age groups matches earlier findings showing that adults heavily weight formant transitions in phonemic decisions, and children demonstrate the same strategies.

One age-related difference was observed, when the conflicting cues were presented, as was the case for switched /ba/ syllables. In that case, adults were able to attend to the formant cue only. Children, on the other hand, were slightly changed their labeling decisions by ART. Nittrouer and Crowther (2001) suggested that children are obliged to integrate acoustic structure for speech signals, even when that structure provides conflicting information about phonemic identity. That suggestion may explain this result in the current study because children showed less consistency in their perceptual weighting when rate of formant transition and rate of amplitude change cued different phonemic decisions. The effect was small, and only found if the rate of formant change was rapid. It was not found for the switched /wa/ syllables

where the rate of formant change was gradual. Therefore, in general, responses of listeners in both age groups were based on the rate of formant frequency change.

The second objective was to examine if sensitivity to the acoustic property of ART, as measured for nonspeech sounds, could explain the extent to which listeners weight that property in making the stop-glide distinction. In this current experiment, Adults and children showed mean discrimination thresholds of less than 50 ms. Therefore, they were highly sensitive to variation in ART, but did not use this acoustic property in their phonemic decisions. Thus, further evidence was found to support the position that sensitivity to acoustic properties does not indicate the role those properties contribute in phonemic decisions. Here, evidence of separate organizational strategies for speech and nonspeech stimuli have been demonstrated. As a consequence, the ability to make decisions about the auditory qualities of nonspeech signals does not predict how the properties of those signals will be used in the perception of speech.

This experiment also served to demonstrate the usefulness of the pitch synchronous temporal envelope processing method presented in 2. For the natural stimuli with transposed envelopes (/ba/ onto /ba/ or /wa/ onto /wa/), recognition accuracy was near perfect. Additionally, all stimuli were reported as sounding 'speech-like' by participants.

Of course, finding that samples of some listener populations fail to weight amplitude rise time in making decisions regarding this manner contrast does not necessarily mean that all other populations of listeners will similarly disregard it. The current study demonstrated that adults and children with age-appropriate speech perception failed to attend to amplitude rise time, and work in Goswami et al. (2010) has shown

that children with dyslexia do not attend to this property, either. However, there is one other population of listeners for whom the question needs to be explored of how well amplitude structure supports phonemic decisions, and that is listeners with cochlear implants. The processing strategies of cochlear implants affect the quality of the signal properties delivered. Where formants are concerned, change over time is only represented when frequencies cross channels, so small changes in formant frequencies are missing. Because children typically depend so strongly on the time-varying spectral structure of the speech signal, it is important to determine how well they might be able to use amplitude structure instead when spectral structure is impoverished in this way. Cross-linguistic studies of weighting strategies have revealed that listeners depend on acoustic cues that are most relevant in their native language. For example, adults whose native language is English rely on the length of the vocalic portion before vocal-tract closure in decisions regarding the voicing of final obstruents [Chen(1970), Peterson and Lehiste(1960), Raphael(1972)]. However, listeners whose native language either does not include syllable-final obstruents or does not make a length distinction based on voicing fail to weight this acoustic property strongly [Crowther and Mann(1992), Crowther(1994), Flege and Wang(1989)]. Nonetheless, individuals can modify their weighting strategies as they gain experience with a new, second language [Miyawaki et al.(1975)]. Thus the hypotheses could be posed that perhaps adults who receive cochlear implants modify existing weighting strategies once they get those implants and perhaps children with implants develop strategies that involve weighting amplitude rise time strongly. Future investigations will need to explore those hypotheses.

4.4 The Perceptual Importance of the Gross Temporal Envelope by Listeners with Cochlear Implants

In the first study, listeners with normal hearing were found to make phonetic decisions based on the rate of change in formant frequency and not based on the rate of change in the temporal envelope. This finding does not mean that all populations of listeners perform in a similar manner. Listeners with cochlear implants (CI) are another population of listeners for whom the question of how well amplitude structure supports phonemic decisions needs to be explored. The processing strategies of CI devices affect the quality of the signal properties delivered. Where formants are concerned, change in frequency over time is only represented when frequencies cross channels. Therefore, small changes in formant frequencies may not be perceived. Because children typically depend so strongly on the time-varying spectral structure of the speech signal, it is important to determine how well they might be able to use amplitude structure instead when spectral structure is impoverished in this way. It has been demonstrated that individuals can modify their weighting strategies as they gain experience with speech [Miyawaki et al.(1975)]. Thus the hypotheses could be posed that perhaps adults who receive CIs modify existing weighting strategies once they are implanted. It could also be the case that children with CIs develop strategies that involve weighting ART strongly. This new experiment using listeners with CIs and similar procedures to Experiment 4 will investigate these questions.

4.4.1 Listeners

Twenty-one adults who wore CIs and were between 18 and 62 years of age participated in the study. All listeners were native English speakers and had varying types

of hearing loss and ages of implantation. All participants had CI-aided thresholds measured by certified audiologists within the 12 months prior to testing. Mean aided thresholds for the frequencies of .25 to 4 kHz were better than 35 dB hearing level for all participants. Six participants had bilateral implants, five used a hearing aid on the ear contralateral to the CI in everyday settings, and ten did not use additional amplification. Processing strategies varied between listeners and included Cochlear Advanced Combined Encoder (ACE), Cochlear Spectral Peak (SPEAK) strategy, and Advanced Bionics HiRes Fidelity 120.

Audiometric testing was performed on nonimplanted and implanted ears to measure residual hearing in those ears. None of the participants had pure-tone average (PTA) thresholds in the nonimplanted or implanted ear better than 68 dB HL for the frequencies of 0.5, 1 and 2 kHz. This was done to determine whether either ear needed to be plugged during testing. Due to the magnitude of hearing loss for all participants, ear plugging was not necessary for any participant during the presentation of stimuli.

4.4.2 Equipment and Materials

All testing took place in a sound-isolated booth. The computer that controlled stimulus presentation was located in an adjacent room. Audiometric testing was done with a Welch Allyn TM262 audiometer using TDH-39 headphones. Stimuli were stored on a computer and presented through a Creative Labs Soundblaster card, a Samson amplifier, and a Roland MA-12C powered speaker. For the labeling and discrimination tasks, the experimenter recorded responses with a keyboard connected to the computer.

The same testing materials from Experiment 4 were used in this experiment including pictures and cards.

4.4.3 Stimuli

The four sets of synthetic stimuli from 4 were used in this experiment. The two sets of synthetic speech stimuli were used in a labeling task. The two sets of non-speech synthetic stimuli were used in a discrimination task.

4.4.4 Procedures

All participants were tested while wearing a single CI. The participants who typically wore hearing aids did not wear them during testing to avoid acoustic hearing. For bilateral CI users, testing was performed using device which was implanted first. Twelve participants were tested through CIs on their right ears, and 9 were tested through CIs on their left ears. All stimuli were presented at 68 dB SPL, measured at ear level and distance, via a speaker positioned one meter from the participant at 0 degrees azimuth.

Audiometric testing of residual hearing in the implanted and nonimplanted ears was performed first.

Participants were tested during two sessions of 60 minutes each on two different testing days. The labeling tasks were tested during the first session. The first labeling task (FRT continuum or ART continuum) was performed, and then followed by the alternate labeling task, with the order of labeling tasks alternated between participants. Discrimination tasks were performed during the second testing session, with task order randomized among participants. Five participants were unable to return for the second testing session.

The same labeling and discrimination tasks from Experiment 4 were used in the this experiment. The only difference was that listeners with CIs were required to obtain 70% correct response in the labeling task to be included, rather than the 80% requirement for listeners with normal hearing. Similar training procedures were also used to ensure uniform experience with the stimuli for each listening group.

4.4.5 Analyses

As in Experiment 4, the proportion of /wa/ responses given to each stimulus on each continuum was used in the computation of these weighting factors. In this experiment, values for the continuous property were normalized to values between 0 and 1 to match settings on the binary property. Weighting factors were calculated for both the FRT cue and the ART cue, for each stimulus set, and averaged across conditions. These FRT and ART weighting factors were used in subsequent statistical analyses. Logistic regression was performed to calculate measures of the perceptual weights for the FRT and ART cues.

To analyze the discrimination functions of each listener, average d' values for each listener in each condition were computed [Holt and Lotto (2005),Macmillan and Creelman(2005)]. The d' value was selected as the discrimination measure because it reduces response bias. The d' value is defined in terms of z-values based on a Gaussian normal distribution which converts values into standard deviation units. The d' value was calculated at each step along the continuum as the difference between the z-value for the "hit" rate (proportion of different responses when A and X stimuli were different) and the z-value for the "false alarm" rate (proportion of different responses when A and X stimuli were the same).

4.5 Results

The data for one participant was excluded because they were unable to recognize nine out of ten of the natural speech tokens correctly Twenty participants were included in the labeling tasks. Sixteen participants underwent testing on the discrimination tasks.

Initially, two-sample t tests were performed to see if side of implant influenced scores for FRT weighting factor, ART weighting factor, d' values for FRT continua, or d' values for ART continua. No significant differences were found, allowing listeners with an implant in their left ear to be grouped together with listeners with an implant in their right ear.

Additionally, potential effects of typically using one CI, two CIs, or a CI plus hearing aid were examined. One-way ANOVAs found no differences in group means for FRT weighting factor, ART weighting factor, d values for FRT continua, or d values for ART continua based on whether participants used one CI, two CIs, or a CI plus hearing aid. Therefore, data were combined across all participants in subsequent analyses, regardless of typical device use.

Means and standard deviations of FRT and ART weighting factors, and d values for FRT and ART continua are shown in Table 4.3. For comparison, the listeners from Experiment 4 had a mean FRT weighting factor across the two stimulus conditions of 9.52 ($\pm$ 3.10). Those same listeners had a mean ART weighting factor of 1.10 ($\pm$ 1.34). Based on the labeling functions and weighting factors of the CI users in this study, these listeners showed different weighting strategies from those observed for adults with normal hearing in the Experiment 4. Additionally, increased variability was seen in FRT and ART weighting factors among listeneres with CIs. There was

Table 4.3: Means and standard deviations of FRT and ART Weighting Factors, and FRT and ART *d*' Values

Measure	M	SD
FRT weighting factor	3.00	2.33
ART weighting factor	2.85	1.74
d' FRT Synthetic	2.34	1.18
d' FRT Sine-Wave	1.54	1.32
d' ART Synthetic	2.59	1.26
d' ART Sine-Wave	2.54	1.12

not a significant correlation between FRT and ART weighting factors, which suggests that individuals with CIs do not necessarily attend to one cue at the exclusion of the other.

As with listeners with normal hearing in Experiment 4, it was important to contrast perceptual weighting with discrimination sensitivity for listeners with CIs. Sensitivity is necessary in order for a cue to be weighted in a perceptual decision, even if the cue is weighted as less important. Table 4.3 shows means and standard deviations of d values for FRT discrimination and ART discrimination for both formant stimuli and sine-wave stimuli. For comparison, a d value of 0 suggests no ability to discriminate the cue, a d value of 2.33 suggests the individual could discriminate 50% of different stimuli as different, and a d value of 4.65 suggests 100% discrimination of different stimuli as different. Overall, participants showed significant variability, but could discriminate about 50% of FRT and ART cues. The one exception was that FRT cues in the sine-wave condition were almost impossible to discriminate for listeners with CIs. Discrimination scores of less than 5% were achieved for FRT differences in the sine-wave condition.

4.6 Discussion

The experiment presented here was conducted to examine whether adult CI users demonstrate different perceptual weighting strategies than listeners with normal hearing. The method to investigate this question used a categorical labeling task for the /ba/-/wa/ contrast and discrimination tasks based on ART and FRT. It was found in Experiment 4 that normal-hearing adults heavily weight spectral structure when listening to speech signals, and are sensitive to both ART and FRT in the discrimination tasks. On the other hand, listeners with CIs did not show the same consistent perceptual weighting of spectral structure when listening to speech signals. It was suggested that this was related to the sensitivity of temporal and spectral structure in signals. Discrimination sensitivity for listeners with CIs was poorer for both ART and FRT. As a result, sensitivity to FRT, as assessed with the discrimination tasks, could explain perceptual weighting of FRT cues to some degree. These findings may simply confirm that sensitivity to FRT is necessary for weighting FRT strongly. In order words, individuals would have to be sensitive to a spectral cue to be able to use it in a phonemic decision. Listeners who were more sensitive to FRT were able to adjust their perceptual strategies to be more similar to listeners with normal hearing.

The results of these experiments have clinical implications. Research with CI users has focused on improving signal presentation by implants to deliver the necessary cues for speech understanding. The findings of this study show that sensitivity to acoustic structure is only one aspect of speech recognition. Listeners must also learn optimal weighting of the various components. Auditory training protocols should help CI users learn to shift their weighting strategies to those that are most effective in speech perception. Shifts in weighting strategies have been seen as a result of

auditory training using nonspeech stimuli (Holt & Lotto, 2006), and it is likely that CI users could benefit from training using speech stimuli.

4.7 Conclusion

Listeners with normal hearing and listeners with cochlear implants use different perceptual weighting strategies when listening to speech signals. Listeners with normal hearing are very sensitive to changes in a signal's temporal envelope and spectral envelope. However, they perceptually weight the spectral envelope more than the temporal envelope when listening to speech signals. Listeners with cochlear implants are less sensitive to changes in a signal's temporal envelope and spectral envelope. The perceptual weighting strategies for listeners with cochlear implants are much less consistent than listeners with normal hearing. This is likely a result of the diminished sensitivity to these acoustic characteristics.

CHAPTER 5

The Perceptual Importance of the Spectral Envelope

5.1 Coherence Masking Protection: Introduction

In previous chapters, the perceptual importance of the gross temporal envelope was compared with the gross spectral envelope for listeners with normal hearing and listeners with cochlear implants (CI). In this chapter, the perceptual importance of the spectral envelope will be compared with the source signal for listeners recognizing speech in the presence of noise. Based on the theory of Auditory Scene Analysis (ASA), listeners rely on the harmonic structure of signals to recover speech in the presence of noise. It will be demonstrated that the harmonic structure of speech is not a necessary condition for listeners to recover speech in the presence of noise. Furthermore, children rely less on the harmonics structure of speech related to source information and rely strongly on the spectral envelope for recognizing speech in noise.

5.1.1 Auditory Scene Analysis

In *Auditory Scene Analysis* [Bregman(1990)], the task of hearing in complex auditory environments is addressed in detail. Informally known as the cocktail-party problem, perceptual research has investigated how the auditory system analyzes a

mixture of sounds. With several interfering signals occurring simultaneously, listeners form perceptual streams to isolate the signals. In order to form perceptual streams, listeners segregate and integrate various components of a mixture of sounds. In other words, listeners allocate regions to objects, or parts to wholes, in order to perceptually describe the physical cause that led to an acoustic event.

Several principles of Gestalt psychology were applied in development of the theory of Auditory Scene Analysis (ASA). These principles were used to explain how the brain created mental patterns by forming connections between elements of sensory input. The principle of *Similarity* explained that sounds of similar timbre are grouped together. The principle of *Proximity* explained that sounds that were apparent in both time and frequency were grouped together. Conversely, the further apart in frequency two simultaneous components are, the less likely they are to fuse. The principle of *Closure and Belongingness* explained that fragmented views of events could be connected in plausible ways as a way of dealing with missing evidence.

Two categories of stream segregation were described by ASA. First, primitive based stream segregation categorized innate or unlearned segregation methods. The harmonicity principle of stream segregation was argued to be a primitive method of stream segregation because several physical characteristics of a signal remain constant and are dealt with by every human everywhere. The harmonicity principle explained that a simultaneously present set of partials are perceptually assigned to the same auditory stream if they are all harmonics of the same fundamental. This was argued because when a harmonically structured sound changes over time, all the harmonics in it will tend to change together in frequency, amplitude, and direction, in order to maintain a harmonic relationship.

The second category of stream segregation were any methods that required perceptual learning and the development of perceptual knowledge of familiar patterns or schemas. These methods were called schema based stream segregation. In these methods of stream segregation, listeners search for confirming stimulation in the auditory input that will match their stored knowledge of acoustic patterns.

The series of experiments described below were conducted to test the principles of ASA as they relate to speech recognition in the presence of noise. A focus of the experiments was to investigate the primitive principle of harmonicity because it represents the perceptual importance of source information in a speech signal. This was contrasted with the perceptual importance of the spectral envelope of speech as a perceptually organizing mechanism. As a consequence, the notions of primitive and schema principles of speech were also investigated.

5.1.2 Perceiving Auditory Objects

Another focus of the series of experiments described below was to investigate how listeners perceive vowels. Two models of perception have been used to describe the analysis of vowels. In one model, listeners recover isolated details of a signal separately, and must construct a vowel percept by combining the individual parts. In the second model, listeners analyze a speech signal by attending to the entire spectral envelope as a single object without attending to individual components. Investigating these models of vowel perception may lead to better insight into how listeners perceive auditory objects.

Traditionally, the approach to studying human speech perception has primarily focused on asking how individual details or 'cues' support the recovery of phonetic

structure. Experimental methods have used synthetic syllables in which all acoustic elements are held constant across a series of stimuli at settings providing ambiguous information about phonetic identity, with one exception. That one signal cue is manipulated across the series, which will produce labeling between two phonetic categories. These experiments typically focus on manipulating vocal tract resonant frequencies because this property has been shown to underlie phonetic structure. All stimuli are presented to listeners for phonetic labeling multiple times, and perceptual categories are derived from listeners' responses. This line of investigation has been useful in helping to investigate small, well-defined details of the signal. However, this approach assumes the perception of individual components of a speech signal. This assumption may be valid during controlled experiments using isolated synthetic stimuli, but may not explain the complexity of speech perception in real-world situations.

In order to understand the complexity of speech perception, it is important to examine how the various components of the complex signal are integrated into coherent streams. Several experiments that have examined the question of how discrete components of the speech signal are fused in perception have reliably revealed that they coalesce such that any one is no longer available for individual inspection, except under very special circumstances. One example of this is the duplex perception paradigm [Liberman, Isenberg, and Rakerd(1981), Mann, and Liberman(1983), Whalen and Liberman(1987)]. In this paradigm, the constant components of a synthetic syllable are presented to one ear, and the informative cue is presented to the other ear. This presentation evokes two distinct percepts: a fused percept of the constant and informative cues, and the informative cue by itself. Results demonstrate that listeners can make two phonetic judgments. The first judgement is about the

fused percept. The second judgment is about the acoustic qualities of the isolated informative cue, such as whether it is rising or falling in frequency. This suggests that listeners can perceptually organize signals into a single auditory object and also recover individual components.

Given that listeners recover fused auditory objections, the next question is whether listeners automatically fuse signal components to analyze a single object, which would suggest that the principles governing that coherence are innate. Alternatively this perceptual strategy might be acquired through years of experience listening, and require the learning. Morrongiello et al. (1984) found that 5-year-old children did not demonstrate the signal coherence that was a hallmark of adults' perception. Instead these children demonstrated a need to learn how to fuse separate signal properties as adults do. Nittrouer and Crowther (2001) attempted to replicate the result with adults and children using different stimuli, but instead found that 5-year-old children were unable to attend to individual elements of speech signals. Adults actually demonstrated an ability recover the separate acoustic elements of speech signals under some conditions. This conclusion has been supported from other experiments [Carney, Widin, and Viemeister(1977),McMurray, Tanenhaus, and Aslin(2002)]. Evidence from these studies would suggest that child listeners attend to more global structure in a speech signal, but learn to attend to certain individual components of the signal as they gain more listening experience.

The contradictory conclusions of Morrongiello et al. (1984) and Nittrouer and Crowther (2001) suggest that additional investigation is needed to understand the perception of auditory objects. Given the variability of the results using the experimental paradigm of Morrongiello et al. (1984) with different stimuli, a different

paradigm may produce more consistent results. It has been found that children and adults assign different perceptual weights to the acoustic cues defining phonetic categories [Nittrouer(1992), Nittrouer, Manning, and Meyer(1993), Nittrouer(2004)]. Coherence masking protection is one paradigm that decouples perceptual coherence from the informativeness of the signal, and is therefore useful to investigate auditory object perception for different ages.

5.1.3 Coherence Masking Protection

Coherence Masking Protection (CMP) is an experimental paradigm in which a low-frequency component is labeled accurately in poorer signal-to-noise levels when combined with a high-frequency cosignal, rather than presented alone. Similar procedures have been used to study auditory grouping for non-speech signals [Hall and Grose(1990)] and speech signals [Grose and Hall(1992)]. The procedures used in the experiments discussed below were presented in Gordon (1997). In these procedures, the signal intensity was measured for listeners to provide correct labels 79.4 percent of the time for voiced speech stimuli modeled after the vowels /ɪ/ and /ɛ/. In the F1-only condition, only the first-formant target was presented with interfering noise. The noise was low-pass filtered to the same frequency range as F1, below 1000 Hz. In the full-formant condition, a constant F2/F3 cosignal was presented with synchronous onset and offset to F1 at a level 12 dB down from F1. Adults thresholds improved by 3.2 dB in the full-formant condition over the F1-only condition, even though that cosignal provided no additional information regarding vowel identity and was outside the critical band of F1. Therefore, this procedure demonstrates protection from masking as one benefit of perceptual coherence of acoustic components. Additionally,

the procedure examines perceptual coherence without variation in how phonetically informative stimuli are.

Finally, the experiments described below were also designed to examine what principle might account for any patterns of perceptual integration or segregation. It was specifically asked if perceptual coherence across vowel formants seems to be based on all formants sharing a common harmonic structure. This question was addressed by Experiment 5.3 in which the F1 target did not share harmonic structure with the F2/F3 cosignal. In Experiment 5.4, the harmonic structure of the signal was removed to analyze coherence for sine-wave tones. In Experiment 5.5, perceptual coherence was investigated for sine-wave tones with a harmonic cosignal. Finally, in Experiments 5.6 and 5.7, the necessity of signal periodicity in CMP was explored using aperiodic cosignals.

CMP was used to examine perceptual coherence of speech signals by adults and children who were 8 or 5 years of age. Including children provided an test of the hypothesis that children rely more than adults on perceiving complete auditory objects. This hypothesis derived from the findings of Nittrouer and Crowther (2001). It proposes that children are more obliged to fuse speech components than are adults.

It is also possible that children might demonstrate weaker perceptual coherence than adults. If listeners perceive individual components of a signal separately, the challenge of integrating F1 with the F2/F3 cosignal might be viewed as perceptually sophisticated for children. Listening experience may be necessary before these components can be fused perceptually.

5.2 Experiment 1: Synthetic Speech

In this experiment, the procedures of CMP from Gordon (1997) using synthetic stimuli were replicated. Similar to Gordon's experiment, adults were included as listeners. Additionally, two age groups of children were included as listeners in this experiment.

5.2.1 Listeners

Three age groups of subjects were tested in this experiment. Nive-five listeners participated in the experiment in total: 25 adults between the ages of 18 and 25 years; 32 children between 8 years, 0 months and 8 years, 11 months; and 37 children between 5 years, 2 months and 5 years, 11 months. A hearing screening was completed at the time of testing. All participants passed hearing screenings of the frequencies of 0.5, 1.0, 2.0, 4.0, and 6.0 kHz presented at 25 dB HL to each ear separately. All participants (or in the case of children, their parents on their behalf) reported having normal hearing, speech and language. Children were required to have less than six episodes of otitis media before the age of 3 years.

5.2.2 Equipment and Materials

Testing took place in a sound-isolating booth. A Welch Allen TM-262 audiometer and TDH-39 headphones was used for hearing screenings. The computer that controlled stimulus presentation and recorded responses was located in an adjacent room. Stimuli were presented from the computer, through a Soundblaster digital-to-analog converter, amplified by a Samson Q5 headphone mixer, and AKG-K141 headphones.

Listeners were presented two pictures on cardboard (6 in. x 6 in.) so that they could point to the picture representing their response choice after each stimulus presentation. One picture was of a dog biting a womans leg (bit), and the other was of a man with playing cards in his hands and stacks of poker chips in front of him (bet).

5.2.3 Stimuli

A Klatt synthesizer, Sensimetrics "SenSyn," was used to create the synthetic speech stimuli. A sampling rate of 10-kHz and 16-bit digitization was used. Target stimuli were 60 ms long, which included 5-ms on and off ramps. Stimuli were modeled on the vowels /ɪ/ and /ɛ/, with three steady-state formants. F1 was 375 Hz for /ɪ/ and 625 Hz for /ɛ/. F2 and F3 were the same for all stimuli: 2200 Hz and 2900 Hz, respectively. The formant bandwidths were 50 Hz for F1, 110 Hz for F2 and 170 Hz for F3. The fundamental frequency (f0) was 125 Hz.

A digital low-pass filter was used to create the F1-only stimuli and the low-frequency portion of the full-formant stimuli. The two stimuli described above were low-pass filtered with attenuation starting at 1000 Hz, a transition band to 1250 Hz, and 50-dB attenuation above that. A digital high-pass filter was used to create the cosignal for the full-formant stimuli from the /ɛ/ token. The high-pass filter has a stop-band with 50-dB attenuation below 1000 Hz, a transition band up to 1250 Hz, and a pass band above 1250 Hz. For the full-formant stimuli, this high-pass portion was combined with the low-pass F1-only portions using synchronous onsets and offsets. The F2/F3 cosignal was 12 dB lower than the F1 target. Fig. 5.1 shows smoothed spectra of the /ɪ/ and /ɛ/ full-formant stimuli.

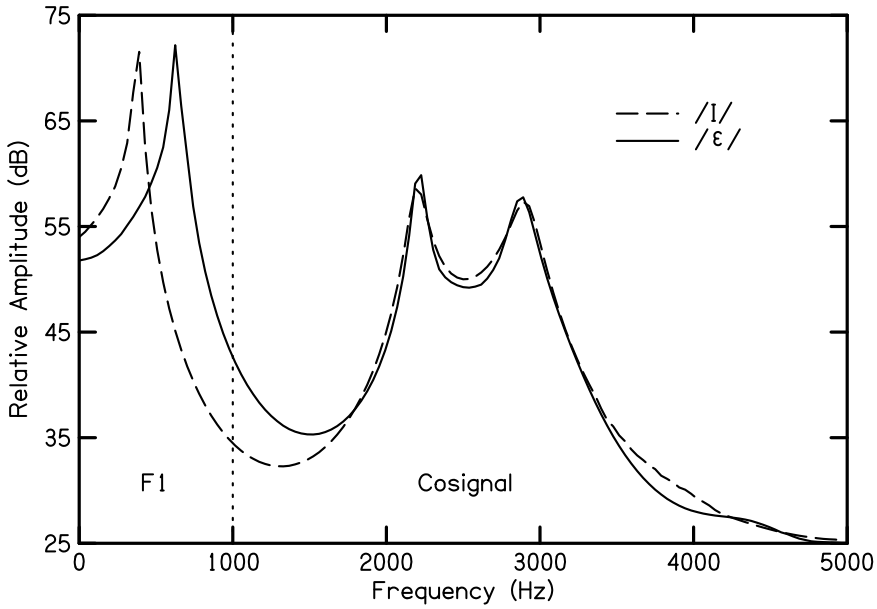

Figure 5.1: Smoothed Spectra of the /ɪ/ and /ɛ/ Full-formant Stimuli

To create the masking noise, 600 ms of flat-spectrum white noise was generated with a random-number generator in Matlab. The noise was low-pass filtered below 1000 Hz in the same manner as the F1-only stimuli.

5.2.4 Procedures

Testing was completed in a single session. Feedback was provided during training to ensure that listeners reliably labeled stimuli and could perform the task. Feedback was not provided during testing.

An initial set of general training was performed to introduce the task. During general training, the 60-ms full-formant stimuli were presented without noise. They were told to pointing to the picture corresponding to the stimulus and say the word /bit/ and /bet/ for the vowels /ɪ/ and /ɛ/. Training consisted of a total of 50 tokens (25 of each) in a random order at 74 dB SPL.

Next, a set of condition-specific training was provided to match the condition of testing. Training for the F1-only condition was always completed after general training because F1-only was the first condition of testing. Training consisted of presenting 50 of the F1-only stimuli at 74 dB SPL without noise. A pre-test was then completed. As soon as the listener responded correctly to nine out of ten consecutive presentations, the pre-test stopped. If 50 stimuli were presented without the listener ever responding correctly to nine out of ten consecutive presentations, that listener did not complete the adaptive testing section for that particular condition. The condition-specific training and the pre-test were repeated before testing with the full-formant stimuli, using full-formant stimuli.

An adaptive procedure [Levitt(1971)] was used to vary the signal-to-noise ratio (SNR) to find the ratio at which each listener could provide the correct vowel label 79.4 percent of the time. The level of the signal varied, while the amplitude of the noise was held constant at 62 dB SPL. The initial signal level started at 74 dB SPL. After three consecutive correct responses, the level of the signal decreased by 8 dB. That progression, or "run," of decreasing signal level by 8 dB after three correct responses continued until the listener made one labeling error. When the error occurred, the level of the signal increased by 8 dB. That shift in direction of amplitude change is termed a "reversal." Signal amplitude continued to increase for each incorrect response until the listener responded with three correct responses. This event began another reversal. For the first two runs (one with decreasing amplitude and one with increasing), signal level changed by 8 dB when appropriate. For the next two runs, signal level changed by 4 dB. For the remaining twelve runs, level changed by 2 dB.

The mean signal level at the last eight reversals was used as the threshold. Order was randomized by the experiment software. No feedback was provided.

After testing was completed in each condition, listeners heard ten stimuli without noise at 74 dB SPL. In order for a listener's data to be included, they were required to respond correctly to nine out of ten tokens. Listeners were required to meet both the pre- and post-test inclusionary criteria. This ensured that the adaptive tracking procedure was not affected by a listener's inability to know the vowel labels.

5.2.5 Results

Several participants failed to meet either the pre- or post-test criterion described above. One adult (4%), six 8-year-olds (19%), and fourteen 5-year-olds (38%) had their data excluded from the results. These listeners failed to meet criterion for the F1-only condition in all cases. The remaining participants were 24 adults, 26 8-year-olds, and 23 5-year-olds with data to be included in the analyses.

Comparison of current results to Gordon (1997)

Because children participated in this experiment, methods for the current experiment differed slightly from those of Gordon (1997). Therefore, the first step in comparing these data was to see if the magnitude of the CMP effect was similar for adults across the two experiments. Table 5.1 shows labeling thresholds for all groups. Mean thresholds (and SDs) in Gordon's experiment were 58.5 dB (2.3 dB) for the F1-only condition and 55.3 dB (2.1 dB) for the full-formant condition. Therefore, adults in Gordon (1997) showed 3.2 dB of masking protection. Adults in the current experiment showed 3.3 dB of masking protection. Although thresholds were slightly higher in the current experiment, the magnitude of masking protection was equivalent.

Table 5.1: Means (and standard deviations) of labeling thresholds for Synthetic Stimuli

Synthetic Speech					
		F1-only		Full-formant	
Age	n	M	SD	M	SD
Adults	24	61.2	3.4	57.9	1.4
8-yr-olds	26	65.0	4.1	58.8	1.1
5-yr-olds	23	70.2	3.8	61.1	2.9

Age effects

Means (and SDs) of differences (in dB) between the F1-only and full-formant stimuli for adults, 8-, and 5-year-olds were 3.3 (3.5), 6.2 (3.8), and 9.2 (3.7), respectively. Those differences represent the CMP effect for each age group. A two-way Analysis of Variance (ANOVA) was performed on the thresholds shown in Table 5.1, with age as a between-subjects factor and number of formants (F1-only or full-formant) as within-subjects factors. Both main effects were found to be significant: formants, $F(1, 70) = 208.18$, $p < .001$, and age, $F(2, 70) = 39.38$, $p < .001$. Thresholds were generally higher for younger rather than for older listeners and for the F1-only than for the full-formant stimuli, confirmed by the ANOVA. Also, the Age x Formants interaction was significant, $F(2, 70) = 15.05$, $p < .001$. This interaction indicates that the magnitude of the formant effect increased with decreasing age.

Matched t-tests were performed on the difference in thresholds between the F1-only and full-formant stimulus conditions for each age group separately to see if CMP effects were significant. In all cases, the CMP was significiant: adults, $t(23) = 4.57$, $p < .001$; 8-year-olds, $t(25) = 8.28$, $p < .001$; and 5-year-olds, $t(22) = 11.96$, $p < .001$.

5.2.6 Discussion

The results of this experiment replicated the results of the CMP paradigm used in Gordon (1997) for adult listeners. Additionally, children demonstrated CMP with a larger magnitude than adults.

This experiment was conducted to investigate the perceptual importance of auditory objects being represented as broad spectral shapes. It appears listeners had poorer recognition when speech was being represented by a narrow-band limited spectral prominence. When the cosignal of F2 and F3 was added to the target formant, listeners had better recognition. This suggests that these components were combined and used together, or in other words, cohered in the same perceptual stream, in order to better recognize speech.

There are several questions remaining to be investigated. First, why do the spectral components cohere? After all, the formants in the cosignal provide no acoustically distinguishing information on their own. In this sense, the cosignal may only be adding additional information to distract the listener from perceiving F1. It could be the case that the shared harmonicity of the cosignal with F1 directs the listener's attention to the relevant information in F1. Second, is the perceptual 'target' represented by the acoustically distinguishing F1 or an entire vowel object comprised of a broad spectral shape with several resonances? In other words, does the cosignal help the listener attend to the F1 'target' or does the combination of F1, F2, and F3 better represent a vowel's broad spectral shape than F1 alone? Additional conditions of this experiment were conducted to answer these questions and explain CMP.

5.3 Experiment 2: Disharmonic Stimuli

In Experiment 5.2, it was suggested that children showed evidence of more strongly fusing disparate spectral components of the signal than adults did. This finding seems to refute the idea that part of perceptual learning involves discovering how to fuse related signal components to form auditory objects. Children at the age of 5 already performed this perceptual integration. This finding might initially suggest that CMP is a consequence of a primitive principle of stream segregation, with an obvious candidate being harmonicity. One hypothesis to explain this results is that the cosignal contributes to the recognition of the F1 component of the signal because both components have a f0 of 125 Hz. This second experiment was undertaken to test that hypothesis.

5.3.1 Listeners

Twenty-five adults, 27 8-year-olds, and 20 5-year-olds participated. New listeners were recruited for this study, but all met the same criteria as in Experiment 5.2.

5.3.2 Equipment and materials

The same equipment and materials as those used in Experiment 5.2 were used in this experiment.

5.3.3 Stimuli

The same F1-only stimuli from Experiment 5.2 were used in this experiment. F1 was either 375 Hz (for /ɪ/) or 625 Hz (for /ɛ/), both with an f0 of 125 Hz. The full-formant stimuli were created in the same way as in that first experiment, except that the F2/F3 cosignal had an f0 of 175 Hz.

5.3.4 Procedures

Training was the same as in Experiment 5.2, with general training using the full-formant stimuli from Experiment 5.2 with a shared harmonic structure across components. Condition-specific training and pre-testing was conducted with the stimuli from this experiment. Adaptive testing was also the same as in Experiment 5.2. Half of the listeners started with the F1-only stimuli, and half started with full-formant stimuli.

5.3.5 Results

The same inclusionary criteria from Experiment 5.2 were used in this experiment. Several participants failed to meet either the pre- or post-test criterion: five adults, seven 8-year-olds, and seven 5-year-olds. Table 5.2 shows mean thresholds (and SDs) for the F1-only and full-formant stimuli for each age group. Thresholds were similar across the two experiments for each listener group for the F1-only stimuli. As found in Experiment 5.2, it appears that thresholds were slightly higher for children than for adults overall, but children continue to show larger CMP effects. In fact, adults did not demonstrate CMP effects at all for the full-formant stimuli in this experiment.

In the current experiment, the mean CMP effect was 7.1 (3.1) for 5-year-olds and 5.1 (3.5) for 8-year-olds. A two-way ANOVA performed on thresholds revealed significant main effects of age, $F(2,50) = 23.25$, $p < .001$, and number of formants, $F(1,50) = 80.83$, $p < .001$, as well as a significant Age x Formants interactions, $F(2,50) = 25.69$, $p < .001$. This significant interaction indicates that the magnitude of the CMP effect decreased with increasing listener age, as was observed in the first experiment. Matched t-tests were performed on differences in thresholds between the

Table 5.2: Means (and standard deviations) of labeling thresholds for Disharmonic Synthetic Stimuli

Disharmonic Synthetic Speech					
		F1-only		Full-formant	
Age	n	M	SD	M	SD
Adults	20	60.4	1.9	60.7	1.9
8-yr-olds	20	65.8	4.4	60.7	2.7
5-yr-olds	13	70.2	3.6	63.3	3.0

F1-only and full-formant. Significant CMP effects were found for 5-year-olds, $t(12) = 8.29$, $p < .001$, and 8-year-olds, $t(19) = 6.42$, $p < .001$, but not for adults ($p > .10$).

5.3.6 Discussion

This experiment was conducted to examine the magnitude of the CMP effect for each age group when stimuli lack a common harmonic structure between the low-frequency F1 component and the high-frequency cosignal. It was found that disruption in harmonicity due to change in the f0 of the F2/F3 cosignal was sufficient to eliminate the CMP effect for adults. It remained intact for children, which means that the principle of harmonicity could not explain children's coherence of signal components in speech perception. Therefore, something else must explain children's strong and seemingly obligate tendency to integrate spectral components in speech signals such that they form a unitary auditory percept. In summary, the first two experiments revealed that children demonstrate CMP effects greater in magnitude than those of adults. Therefore, at least one primitive principle of auditory grouping (harmonicity) cannot explain this strong perceptual integration in children. However, the mechanism that does underlie children's strong tendency to fuse components of

speech signals is not discernible from these two experiments. A third experiment was designed to explore another possibility, that children fuse signal components based on a strategy in which elements with any harmonic structure are integrated regardless if they are related to the same f0.

5.4 Experiment 3: No Harmonicity

5.4.1 Listeners

Subjects that participated in Experiment 5.2 also participated in this experiment.

5.4.2 Equipment and materials

The same equipment and materials were used in this experiment as in the first two experiments.

5.4.3 Stimuli

The sine-wave speech stimuli were created with TONE [Tice and Carrell(1997)]. A sine-wave tone was synthesized for each of the three formants found in the synthetic speech stimuli from Experiment 5.2. The F1-only stimuli consisted of a single tone at 375 and 625 Hz for the vowels /ɪ/ and /ɛ/, respectively. The cosignal was comprised of two signals at 2200 and 2900 Hz. The amplitude of the cosignal was adjusted so that the tones of F2 and F3 were 12 dB lower than the F1 tone.

Table 5.3: Means (and standard deviations) of labeling thresholds for Sine-wave Stimuli

Sine-Wave Speech					
		F1-only		Full-formant	
Age	n	M	SD	M	SD
Adults	24	56.5	1.5	57.9	1.7
8-yr-olds	24	59.1	4.6	60.2	3.7
5-yr-olds	22	67.0	5.0	65.5	4.8

5.4.4 Procedures

The same procedures as those used in Experiment 5.2 were used in this experiment. General training using the full-formant synthetics speech stimuli from Experiment 5.2 was presented. Condition-specific training consisted of practice using 50 presentations of the sine-wave speech without background noise with feedback. The pre-test followed, adaptive testing, and post-test followed.

5.4.5 Results

Means (and standard deviations of labeling thresholds in this experiment are shown in Table 5.3. For the sine-wave condition, the main effect of age was significant, again reflecting generally higher thresholds for younger listeners. However, the main effect of number of formants was not significant, indicating that thresholds were similar for the F1-only and full-formant stimuli. Finally, the Age x Formants interaction was significant, reflecting different trends across these stimulus types for the three groups of listeners. For these stimuli, means (and SDs) of differences (in dB) between F1-only and full-formant stimuli for adults, 8-, and 5-year-olds were -1.4 (2.3), -1.1 (2.5), and 1.5 (4.8), respectively.

Because the source of this last interaction is not immediately obvious from examining Table 5.3, matched t-tests were performed for each age group separately, comparing thresholds for F1-only and full-formant stimuli. Results showed that adults and 8-year-olds actually had significantly lower thresholds for F1-only than for full-formant stimuli: for adults, t(23) = 3.00, p = .007; and for 8-year-olds, t(23) = 2.14, p = .044. Five-year-olds showed no significant difference (p ¿ .10) in thresholds for these two types of stimuli when they were sine waves. That result is notable because 5-year-olds showed the greatest difference between the F1-only and full-formant stimuli when they were synthetic speech.

It could be the case that listeners failed to show CMP for the sine-wave speech stimuli because thresholds may have been close to a psychoacoustic limit for the F1-only stimuli, and so thresholds for the full-formant stimuli could not be any lower. However, that suggestion is contradicted by results from Gordon (2000) showing that some adults demonstrated thresholds as low as 53 dB SPL in this experimental paradigm.

5.4.6 Discussion

This experiment was conducted to test the hypothesis that a primitive-based principles do not explain children's strong tendency to fuse components of the speech signal. This was accomplished by using a stimuli that lack the harmonic structure of speech. The results suggest that listeners did not show CMP when a sine-wave cosignal was used. This likely that listeners do not perceive these sine-wave stimuli as speech, and instead use a non-speech mode of perceptual organization. In the next

experiment, a synthetic cosignal will be used with a sine-wave F1. It is hypothesized that this stimuli will facilitate a speech-like percept for listeners.

5.5 Experiment 4: Hybrid Stimuli

The purpose of this third experiment was to further examine conditions under which CMP might be observed for adults and children. The primary hypothesis addressed was that a schema-based principle of perceptual organization might explain children's strong tendency to fuse spectral components of speech signals. Specifically the question was asked if children fuse signal components when they recognize them as being part of a single speech signal.

To test this hypothesis, a procedure developed by [Gordon(2000)] was again used. In this procedure, a target signal that explicitly lacks the qualities of speech is presented in combination with a speech-like cosignal to see if that target is recruited into the speech percept, which would produce CMP effects. Because the target lacks the typical qualities of speech, integration of signal components, if observed, could not be attributed to properties of the signal itself. If that integration is observed, it provides evidence that a schema-based principle underlies the kind of integration that leads to CMP. The alternative possibility is that the non-speech target would be segregated perceptually from the speech cosignal and used by itself to assign labels, a strategy that would not result in CMP.

Following the procedures from Gordon (2000), stimulus design in this third experiment used low-frequency sine waves as the F1 targets, combined with the synthetic speech F2/F3 cosignal of Experiment 1. Sine waves lack speech-like qualities themselves, so they meet the criterion for this experiment. Replicating Gordon's

procedures, those sine wave targets were the same frequencies as the F1 targets in Experiments 1 and 2: 375 Hz and 625 Hz. These frequencies are harmonics of the 125-Hz f0 used to generate the F2/F3 cosignal. Extending procedures of Gordon, low-frequency sine waves other than the 375-Hz and 625-Hz tones of the previous experiments were also used. These other tones were deliberately selected to be out of alignment with the harmonic structure of the F2/F3 synthetic speech cosignal. This manipulation was used to further test the hypothesis that harmonicity might explain, at least to some extent, the perceptual integration of target and cosignal that leads to CMP: If CMP effects are present (or greater) when the sine wave targets are harmonics of the f0 of the cosignal, but not otherwise, then harmonicity could be invoked to explain the phenomenon, at least to some extent. If CMP effects are similar in magnitude regardless of whether the target is or is not a harmonic of the cosignal, then harmonicity could not be attributed with explaining any of the effect.

Another way in which procedures in this experiment differed from those of Gordon (2000) had to do with the labels that listeners were asked to apply to the sine wave targets. Gordon had the adults in that study label these targets as 'high' or 'low,' but they applied the vowel labels (/ɪ/ and /ɛ/) to the hybrid, full-formant stimuli. Even though these labels for non-speech tones seem natural and obvious to adults, they are actually abstract. Children learn them through (even rudimentary) musical training, which all 5-year-olds in this study may not have had. Partly for that reason, but also to keep procedures consistent across experiments, listeners used the same vowel labels for the sine wave targets as for the full-formant stimuli in this experiment. Although it might have been a bit unnatural for older listeners to assign a phonetic label to a non-speech tone, it was considered preferable for the youngest

children. In any case the pre- and post-tests provided the opportunity to identify and dismiss listeners who had difficulty assigning phonetic labels to these non-speech signals.

5.5.1 Listeners

A total of sixty-nine listeners participated in this experiment: 20 adults, 27 8-year-olds, and 21 5-year-olds. All participants met the criteria for participation described in Experiment 5.2.

5.5.2 Equipment and materials

The same equipment and materials were used in this experiment as in the previous experiments.

5.5.3 Stimuli

Six sets of stimuli were used in this experiment. Three sets of F1-only sine-wave signals were included along with three sets with those sine waves combined with F2/F3 cosignals. The cosignal in this experiment was taken from Experiment 5.2, so all had a harmonic structure based on an f0 of 125 Hz. The amplitude of the cosignal was adjusted so that the resonant peaks of F2 and F3 were 12 dB lower than the F1 tone.

Each set of the F1-only conditions consisted of a stimulus for the vowel /ɪ/ and /ɛ/. One set of the F1-only conditions consisted of the F1 tones used in Experiment 5.4 (375 Hz and 625 Hz), and that condition will be referred to as the mid-F1 condition. The purpose of the other two sets was to shift the frequency of the sine waves

away from values harmonically related to the 125-Hz f0 of the cosignal, while maintaining F1 values that would reasonably be expected to lead to /ɪ/ and /ɛ/ percepts. In one set, the frequency of the sine waves was moved to 438 Hz (/ɪ/) and 688 Hz (/ɛ/). The stimulus sets with those values will be described as the high-F1 condition. In the final set, the low-F1 condition had sine waves of 337 Hz and 587 Hz for /ɪ/ and /ɛ/, respectively.

In summary, there were low-, mid-, and high- stimulus conditions, and within each of those conditions there were F1-only and full-formant stimuli. The full-formant stimuli paired a sine-wave F1 with a F2/F3 cosignal consisting of synthetic speech.

5.5.4 Procedures

The same procedures as those used in Experiment 5.2 were used in this experiment. The order of presentation was randomized among the six conditions in the following manner. One of the six conditions was selected to be the first for a participant such that order of the six stimulus sets was evenly distributed across listeners within each age group. Condition and number of formants was then alternated across presentations such that repetitions were not made. As an example, if that first stimulus set happened to be the low F1-only condition, then the second set had to be a full-formant condition and it had to be one of the other F1 conditions (mid or high). The third stimulus set had to return to the number of formants presented in the first set, and the remaining F1 condition was used. The fourth through sixth stimulus sets alternated through F1 conditions in the same order as the first three had, presenting the stimuli with one or three formants that had not been presented for that condition in the first round.

General training was provided as in Experiment 5.2 with full-formant stimuli having the same harmonic structure across target and cosignal. Before testing with each of the six stimulus sets, condition-specific training using 50 presentations in no background noise was provided. The pre-test, adaptive testing, and post-test followed.

5.5.5 Results

Eight 8-year-olds and six 5-year-olds were excluded from data analysis because they were unable to label nine out of ten items correctly in either the pre- or post-test for one of the conditions. In total, data were included for 20 adults, 19 8-year-olds, and 15 5-year-olds.

Based on the results of Experiment 5.4, listeners did not perceive the F1-only sine-wave stimuli as speech. The full-formant stimuli, on the other hand, were described by listeners as unambiguously sounding like speech.

Table 5.4 shows means (and SDs) for each age group for each stimulus type. A three-way ANOVA was performed on these thresholds, with age as a between-subjects factor and condition (low-, mid- or high-) and number of formants (one or three) as within-subjects factors. Table 5.5 shows results of the ANOVA analysis. The main effects of formants and age were significant, indicating that thresholds generally increased with decrease age. Also, thresholds were generally lower for full-formant than for F1-only stimuli. The main effect of condition was not significant, suggesting that thresholds were similar across conditions of low, mid, and high. The Age x Formants interaction was significant, reflecting the apparent finding that the magnitude of the CMP effect decreased with increasing age. The Age x Condition

Table 5.4: Means (and standard deviations) of labeling thresholds for Hybrid Stimuli

Hybrid Speech													
		Low F1				Mid F1				High F1			
		F1-only		Full-formant		F1-only		Full-formant		F1-only		Full-formant	
Age	n	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Adults	24	57.6	2.3	56.7	1.0	57.5	1.1	55.6	1.4	57.0	1.4	56.3	0.8
8-yr-olds	19	63.0	5.7	57.6	2.0	63.8	6.4	57.9	4.2	60.6	3.9	57.4	0.9
5-yr-olds	15	68.7	5.9	61.7	4.4	69.8	5.7	60.7	5.0	70.2	5.8	59.5	5.1

Table 5.5: Statistical Outcomes of Three-way ANOVAs performed on Thresholds for Hybrid Stimuli

Effect	*df*	*F*	*p*
Age	2,51	36.80	$< .001$
Condition	2,102	2.78	.067
Formants	1,51	95.67	$< .001$
Age x Condition	4,102	1.60	NS
Age x Formants	2,51	18.93	$< .001$
Condition x Formants	2,102	2.24	NS
Age x Condition x Formants	4,102	4.66	.002

interaction was not significant, indicating that age-related differences in thresholds were comparable across conditions of low, mid, and high F1.

The CMP effect varied across age groups. For adults, a small, positive CMP effect was found, especially for the mid-F1 condition. For both 5- and 8-year-olds, the CMP effect was highly significant ($p < .001$). This suggested that the cosignal

and sine-wave signal were cohering for listeners. Children were found to rely on the cosignal more than adults.

5.5.6 Discussion

This experiment was conducted to test the hypothesis that a schema-based principle may explain children's strong tendency to fuse components of the speech signal. This was accomplished by using a F1 signal that lacks the acoustic qualities of speech. By presenting it synchronously with a signal that possessed those speech-like qualities, the opportunity was afforded listeners to recruit that signal into the speech-like percept. If they did, CMP effects would be observed. If listeners perceptually segregated that target signal from the cosignal, no CMP would be seen.

Results of this experiment showed that in all three conditions children demonstrated substantial CMP effects. In most cases these effects were similar in magnitude to what was observed for speech stimuli in Experiments 5.2 and 5.3. That was true even when the harmonicity of the stimuli was disrupted. Consequently it may be concluded that children group these disparate spectral components together based on the expectation that the signal is speech; harmonicity is not necessary. The one exception to this conclusion was the high-F1 condition, where 8-year-olds showed a diminished CMP effect. However, that result was not due to these children having especially high thresholds for full-formant stimuli; thresholds were similar for these stimuli across the three conditions. Rather, that result was obtained because 8-year-olds' thresholds were lower for the F1-only stimuli in that high-F1 condition than in the other two conditions. It is not clear why that would be, but that particular outcome does not negate the general finding that children showed evidence of CMP with

these hybrid stimuli similar in magnitude to what they showed for synthetic speech. As long as the full-formant stimuli could be recognized as speech-like, children showed the effect.

For adults, findings were different. These experienced listeners were found to have greatly reduced CMP for all conditions in this third experiment, compared to findings with synthetic speech stimuli in Experiment 5.2. Unlike children, adults did not strongly incorporate that non-speech target into a unitary percept, even though they reported hearing these stimuli unambiguously as speech-like. Of the three conditions, the effect was most apparent for the mid-F1 condition in which signals preserved a harmonic relationship across target and cosignal. However, even there it was reduced from Experiment 5.1. There is no obvious reason why the results of this experiment differ from those of Gordon (2000) who did not find any diminishment in effect for these stimuli from what was observed with synthetic speech. Nonetheless, the trends are clear: Adults can much more easily than children be deterred from perceptually integrating signal components so strongly that any one component can not be segregated and independently examined. Put another way, children exhibit stronger perceptual coherence for speech signals than adults, a trend that has been previously reported [Nittrouer and Crowther(2001)]. Furthermore, the primitive principle of harmonicity appears to explain CMP in adults' responding, at least to a small extent. Children, on the other hand, seemed to group formants together into unitary percepts if they were recognized as originating from a common generator.

5.6 Experiment 5: CMP with a Shaped-Noise Cosignal

In the previous experiment, it was shown that a harmonic cosignal facilitated CMP for listeners in different age groups. It was uncertain from that experiment whether CMP was achieved because the cosignal had a harmonic structure or if it was achieved because the cosignal had a 'speech-like' spectral resonances. This experiment was conducted to determine which characteristics of the cosignal are necessary to facilitate CMP. In other words, did listeners perceive the stimuli as speech because of the speech-like harmonic structure or the speech-like spectral envelope. In order to investigate these hypotheses, the harmonic structure of the cosignal was removed, but the spectral envelope was maintained. In Experiment 5.3 with sine-wave tones, both the harmonic structure and the spectral envelope were removed. The spectral envelope will be maintained in this experiment by using wide-band noise as the cosignal. This noise will be filtered to the frequency range of the cosignal, as well as filtered to have the same spectral resonances as the synthetic speech cosignals.

5.6.1 Listeners

All subjects in this experiment met the same inclusionary criteria discussed in Experiment 5.2. A total of sixty-two listeners were tested in this experiment: 21 adults between the ages of 18 and 39 years; 20 children between 8 years, 0 months and 8 years, 11 months; and 21 children between 5 years, 2 months and 5 years, 11 months. Subjects from this experiment also participated in Experiment 5.7

5.6.2 Equipment and materials

The same equipment, materials and testing facilities were used in this experiment as were used in Experiment 5.2.

5.6.3 Stimuli

As a measure of reliability, the F1-only and full-formant synthetic stimuli from Experiment 5.2 were included in this experiment. An additional full-formant condition was used that included a noise cosignal. The noise cosignal was synthesized from white noise generated using a random number generator in Matlab. The white noise was band-pass filtered between 1000 and 5000 Hz using identical filters to Experiment 5.2. An estimate of the spectral-shaping filter was recovered from the synthetic speech stimuli using LPC source-filter separation. This spectral-shaping filter was applied to the band-pass filtered noise. Thus the cosignal had the identical speech-like shape as the constant-formants stimuli, but without the harmonic source. The amplitude of the shaped-noise cosignal was adjusted so that the resonant peaks of F2 and F3 were 12 dB lower than the F1 resonant peak. This condition tested the hypothesis that listeners integrate components consisting of periodic sources.

Just as before, all stimuli were embedded in low-pass filtered masking noise that was 600-ms long during the adaptive testing.

5.6.4 Procedures

Similar to Experiment 5.2, the hearing screening was completed at the beginning of the session. The order of presentation of stimulus conditions was randomized across listeners. The starting condition was randomly selected from the two cosignal conditions (full-formant synthetic or full-formant shaped-noise), with the stipulation

Table 5.6: Means (and standard deviations) of labeling thresholds for Stimuli with a Shaped-Noise Cosignal

Shaped-Noise Cosignal Speech					
		F1-only		Full-formant	
Age	n	M	SD	M	SD
Adults	20	61.4	1.5	58.9	2.0
8-yr-olds	20	65.6	4.2	61.7	3.2
5-yr-olds	14	67.7	4.0	61.7	5.0

that each starting condition was evenly distributed across subjects in each age group. The second condition was always the F1-only stimuli. Finally, the untested cosignal condition was completed.

All listeners completed general training at the beginning of the experiment, followed by the same four steps from Experiment 5.2: condition-specific training, pre-test, adaptive testing, and post-test.

5.6.5 Results

Several participants' data were excluded from the analysis of results for failing to meet the inclusionary criteria. One adult (5%) and seven 5-year-olds (33%) failed to meet either the pre- or post-test criteria described in Experiment 5.2. After data for these participants were excluded, data remained from 20 adults, 20 8-year-olds, and 14 5-year-olds. Table 5.6 shows thresholds for all groups and conditions used in this experiment.

Comparison of current results to Experiment 5.2

Mean thresholds (and SDs) from Experiment 5.2 for the F1-only condition were 61.2 (3.4), 65.0 (4.1), and 70.2 (3.8) for adults, 8-year-olds, and 5-year-olds, respectively. These results were replicated in this experiment with thresholds of 57.9 (1.4), 58.8 (1.1), and 61.1 (2.9), for the same groups in the same order. These group means differ by no more than 0.6 dB across experiments, except for 5-year-olds' thresholds for the F1-only condition. In the current experiment, this mean threshold is 2.5 dB lower than in Experiment 5.2. For each age group, t tests were performed on thresholds from each condition used in both experiments (i.e., F1 only and full-formant synthetic). All were non-significant with p ¿ .10, except for 5-year-olds' thresholds for the F1-only condition, t(35) = 2.21, p = 0.034. This one significant outcome did not affect the current experiment because the 2.5 dB difference across experiments in this one threshold is modest compared to the size of the CMP effect itself for 5-year-olds: 9.1 dB in Experiment 5.2 and 7.2 dB in this experiment. These comparisons for the conditions used in both experiments reveal good reliability for the test measures.

Age and condition effects

A two-way Analysis of Variance (ANOVA) was performed on the thresholds shown in Table 5.6, with age as the between-subjects factor and condition (F1-only, full-formant synthetic, and full-formant noise) as the within-subjects factor to examine overall age and condition effects. The main effect of age was significant, $F(2, 51) = 13.37$, $p < .001$, $\eta^2 = .344$. Post hoc comparisons showed that adults' thresholds were lower overall than those from both groups of children ($p < .001$), but thresholds were similar across children's groups ($p > .10$). The main effect of condition was also

significant, F(4,204) = 78.67, $p < .001$, $\eta^2 = .607$. In addition, the Age x Condition interaction was significant, F(8, 204) = 4.52, $p < .001$, $\eta^2 = .151$. This last outcome indicates variability in the pattern of performance across conditions for each age group. That variability was explored by examining outcomes across conditions for each age group separately, according to the predictions made in the Introduction.

Testing the Hypotheses

One-way, repeated-measures ANOVAs with condition as the factor were performed on data for each age group separately. The post hoc comparisons derived from those ANOVAs served as the outcomes used to evaluate whether or not listeners showed CMP in the full-formant noise condition.

For adults and 5-year-olds, thresholds for the shaped-noise stimuli were lower than for the F1-only stimuli and equivalent to those of the full-formant synthetic stimuli. For 8-year-olds, the evidence is mixed. Thresholds were significantly lower in the full-formant noise condition than in the F1-only condition, supporting the notion that the shaped-noise cosignal was cohering. However, 8-year-olds had higher thresholds for the full-formant noise condition than for the full-formant synthetic condition. This suggested that the shaped-noise cosignal did not provide as large of a benefit as the synthetic cosignal. In summary, 8-year-olds showed some CMP for the full-formant noise stimuli (3.9 dB); the magnitude of that CMP was just not as great as the CMP measured for the full-formant synthetic stimuli.

5.7 Experiment 6: CMP with a Flat-Noise Cosignal

In the previous experiment, it was shown that a shaped-noise cosignal was sufficient to facilitate CMP. This experiment was conducted to determine what characteristic of the shaped-noise cosignal was necessary to facilitate CMP. The ASA primitive principle of onset/offset synchrony would suggest that listeners were able to recover F1 from the interfering noise because the cosignal marker the interval in time to attend to the target. A conflicting hypothesis is that the speech-like spectral shape of the cosignal helped listeners recover the auditory object of a vowel. To test these hypotheses, the speech-like spectral envelope was removed from the cosignal. The remaining signal was spectrally flat noise which was temporally synchronous with the F1 component. It is possible that listeners integrate the cosignal with the F1 component because they share temporal synchrony. However, it is also possible that listeners integrate the flat-noise cosignal with the masking noise because it has a similar spectral shape and similar underlying source signal.

5.7.1 Listeners

Subjects that participated in Experiment 5.6 also participated in this experiment. This allowed for direct comparison between the stimuli with a shaped-noise cosignal and the stimuli with a flat-noise cosignal.

5.7.2 Equipment and materials

The same equipment, materials, and testing facilities were used in this experiment as the previous CMP experiments.

5.7.3 Stimuli

Two conditions of stimuli were used in this experiment: synthetic F1-only from Experiment 5.2, and the synthetic F1 stimuli combined with a flat-noise cosignal. The flat-noise was created by band-pass filtering white noise between 1000 and 5000 Hz. The same high-pass filter from Experiment 5.2 was used. The flat-noise cosignal was then combined with the F1 synthetic vowels for /ɪ/ and /ɛ/. The flat-noise cosignal was adjusted in amplitude to match the RMS amplitude of the shaped-noise cosignal in Experiment 5.6.

This condition investigated whether listener's integrated the cosignal because it shared a synchronous onset/offset with F1. If temporal synchrony is sufficient to evoke CMP, then listeners would show masking protection even when the cosignal is spectrally flat noise. If listeners show masking protection for the shaped-noise condition from Experiment 5.6, but not the flat-noise condition, then it can be concluded that temporal synchrony is not sufficient to explain CMP. Rather it would seem that listeners integrate the various spectral components because they share a 'speech-like' spectral shape. Furthermore, this condition examines whether listeners use the cosignal to attend to F1 by itself, or if the combined F1 and cosignal better represent a 'speech-like' perceptual object spanning across the spectrum.

5.7.4 Procedures

The order of presentation of stimulus conditions was randomize across listeners. Half of the listeners started with the F1-only stimuli and the other half started with the flat-noise stimuli. All listeners completed general training using the full-formant synthetic stimuli from Experiment 5.2. Next, the following four steps were completed

Table 5.7: Means (and standard deviations) of labeling thresholds for Stimuli with a Flat-Noise Cosignal

Flat-Noise Cosignal Speech					
		F1-only		Full-formant	
Age	n	M	SD	M	SD
Adults	20	61.4	1.5	59.1	2.3
8-yr-olds	20	65.6	4.2	64.9	4.9
5-yr-olds	14	67.7	4.0	65.4	5.6

in each of the testing conditions: (1) condition-specific training, (2) a pre-test, (3) adaptive testing, and (4) a post-test.

5.7.5 Results

The inclusionary criteria from Experiment 5.2 were used in this experiment. One adult failed the post-test for the F1-only condition. Seven five-year-olds failed the pre-test or post-test for at least one condition. After excluding the data for these participants, 20 adults, 20 8-year-olds, and 14 5-year-olds remained. Table 5.7 shows thresholds for all groups and conditions used in this experiment.

A one-way, repeated-measures ANOVAs with condition as the factor were performed on data for each age group separately. The post hoc comparisons derived from those ANOVAs served as the outcome used to evaluate whether listener's had shown CMP. It seems for adult listeners that the flat-noise cosignal was sufficient to facilitate CMP. However for children, the flat-noise condition was not significantly different than the F1-only condition. Therefore, CMP did not occur for children with the flat-noise cosignal.

The major outcome of this study was that young children showed greater coherence masking protection than adults for the speech signals used here, and were less readily perturbed from integrating spectral components in that way than were adults. The effect for children was not restricted to conditions in which all components shared a harmonic relationship, or even to conditions in which all components had harmonic structure. It was more related to listener characteristics than to signal characteristics. These outcomes support the hypothesis that young children are more strongly obliged than adults to fuse spectral components when those components are recognized as being part of a speech signal.

CHAPTER 6

The Perceptual Importance of Periodicity

In previous chapter, the perceptual importance of the source signal was compared with the gross spectral envelope when speech signal is interfered by noise. In this chapter, the perceptual importance of periodicity, related to the source signal, will be demonstrated for signals that are spectrally degraded. It is hypothesized signal periodicity can help listeners perceptually organize spectrally degraded signals to better recover auditory objects. The application of this hypothesis is for the treatment of patients with cochlear implants. Electric signals presented through a cochlear implant are spectrally degraded. Recently, an effort has been made to investigate the possibility of combining any residual acoustic hearing with the electric stimulation to improve patient outcomes. Two examples of this are bi-modal stimulation using a unilateral cochlear implant with a contralateral hearing aid, and hearing-preservation hybrid implants that present both acoustic and electric signals in the same ear. Typically, residual hearing is poor for candidates of cochlear implants, and may only consist of very-low frequencies. Nonetheless, even if the residual acoustic hearing does not support any speech recognition on its own, it may help a listener perceptually organize the spectrally degraded electric signal.

The experiment described below focused on questions related to the combination of the kinds of signals received with combined electric-acoustic stimulation, but specifically when the acoustic stimulation would only be very low frequency. This situation differs from what typically exists in the case of either bimodal or hybrid stimulation: In these situations, patients often have auditory thresholds of 60 dB hearing level or better in frequencies up to at least 500 Hz, but possibly as high as 1 kHz. That means that these patients generally can hear the first formant (F1) of speech signals through their hearing aid. It also means they can generally hear five to eight harmonics of the fundamental frequency. Access to these two properties of the speech signal can help with speaker identification, signal segregation in noisy environments, and cues to both vowel identity and consonant manner. The current study simulated listening conditions in which a traditional cochlear implant would be used, with a low-frequency cut-off of 250 Hz, and the only acoustic signal that could be heard was lower than 250 Hz. This situation meant that typically only the lowest harmonic was available to listeners in this study. As would be expected, this single harmonic is not interpretable as any kind of linguistic unit by itself.

An essential proposition of the work conducted here was that an explanation for the broad variability in speech recognition outcomes observed for users of traditional implants might rest with how well patients organize the degraded signal they receive through those implants. It has been known for quite some time that signals that are inherently non-speech can be organized perceptually so that phonetic qualities can be recovered [Risberg and Agelfors(1982)]. Referred to here are signals that lack the periodic structure imposed on speech by glottal pulsing or the broad bandwidths associated with vocal-tract resonances. The most commonly recognized demonstration

of the principle at stake involves sine-wave speech. In the first use of these speech analogs, Remez et al. (1981) processed speech signals to preserve only the center frequencies of each of the lowest three formants, and presented those frequencies as time-varying sine waves. When no instructions were given to study participants, most reported hearing whistles or bird chirps or some other nonspeech signal. When participants were instructed that they would be hearing sentences that they were to transcribe, however, most were able to recognize the original sentences, indicating that they were able to integrate the separate sine waves into unitary phonetic objects. Thus, it is fair to conclude that listeners were imposing organization on the sensory information they were receiving.

There were a total of four hypotheses in this experiment. First, it was hypothesized that the addition of the low-frequency signal would improve speech recognition. Second, stimuli in the experiment were presented in diotic and dichotic configurations to simulated different auditory prostheses, and it was hypothesized that there would be a benefit of the low-frequency signal in both configurations. Third, stimuli in the experiment consisted of individual words materials and four word sentence materials, and it was hypothesized that the low-frequency signal would improve recognition of the sentence materials more than the words materials. This was expected because the low-frequency signal could contribute in a more global organization sense for sentences than for words. Finally, children and adults were included as listeners, and it was hypothesized that younger listeners would show a greater benefit than adults. This is because children could benefit more than adults when the low-frequency signal if they need extra help perceptually organizing the spectrally degraded signals.

The experiment presented below was designed to evaluate whether the addition of a very low-frequency signal to an implant simulation (using noise vocoding) would improve speech recognition. In all, four hypotheses were tested:

1. Adding the very low-frequency component of the speech signal to an implant-simulated signal would improve speech recognition.

2. The advantage would be greater in magnitude when both signal components were presented diotically, rather than dichotically.

3. The advantage would be greater in magnitude for words presented in sentences rather than in isolation.

4. Children would demonstrate a greater advantage than adults.

To test those hypotheses, adults and children with normal hearing were asked to recognize sentence and word materials. Four-channel noise-vocoded analogs were created from these speech materials, after filtering with a low-frequency cut-off of 250 Hz. The signal component below 250 Hz was presented simultaneously with the vocoded signal in half of the presentations. Stimuli were presented either dichotically or diotically. The outcomes of this work should have significant clinical implications by facilitating our collective ability to answer the who, what, when, and why questions surrounding the electric-acoustic stimulation advantage observed for patients with hearing loss: Who benefits? What is it that they gain? When or under what conditions do they obtain the advantages? And why are those benefits observed?

6.1 Preliminary EAS Experiment

Prior to conducting the main experiment, a preliminary experiment was done to see if there was evidence of a benefit when the low-frequency signal component was combined with the vocoded signal. In this experiment, only sentence materials in the diotic configuration were included. The outcomes of this preliminary experiment were used to establish reliability for the dependent measures used in the main experiment.

6.1.1 Participants

A total of sixty listeners participated in this experiment: 20 adults between the ages of 18 and 39, 20 7-year-olds (ranging from 7 years; 0 months to 7 years; 11 months) and 20 5-year-olds (ranging from 5 years; 0 months to 5 years; 11 months). All listeners were native speakers of American English, and all passed hearing screenings at 25 dB hearing level for the frequencies 0.5, 1, 2, 4, and 6 kHz. All listeners had histories of normal speech and language skills.

6.1.2 Equipment

All sentence materials were recorded in a sound booth onto a computer, using an AKG C535 EB microphone, a Shure M268 amplifier, and a Creative Laboratories Soundblaster soundcard. Perceptual testing took place in a sound-isolated booth, with the computer that controlled the experiment in an adjacent room. Stimuli were presented through a Samson headphone amplifier and AKG-K141 headphones. The hearing screening was done with a Welch Allyn TM262 audiometer and TDH-39 headphones.

6.1.3 Stimuli

Fifty-six 4-word sentences (6 for practice, 50 for testing) were created following standards described in previous studies (Boothroyd & Nittrouer, 1988; Nittrouer et al., 2009; Stelmachowicz et al., 2000). These sentences are comprised of monosyllabic content words, are syntactically appropriate, but are semantically anomalous. These sentences provide lexical and syntactic constraints, but no semantic constraints. Using these sentences limited the advantage that older listeners can gain from using linguistic. This allowing for a more sensitive examination of the role of perceptual organization. They were recorded by an adult male speaker of American English at a 44.1-kHz sampling rate with 16-bit digitization.

A MATLAB routine was used to create the vocoded stimuli. All signals were band-pass filtered with a low-frequency cut-off of 250 Hz and a high-frequency cut-off of 8,000 Hz. Cutoff frequencies of the vocoding channels were 800, 1600, and 3200 Hz. All filtering was done with digital filters that had greater than 50-dB attenuation in stop bands, and had 10-Hz transition bands. To extract an amplitude envelope, each channel was half-wave rectified and filtered using a 160-Hz high-frequency cut-off. The temporal envelopes derived for separate channels were used to amplitude modulate white noise, filtered to the same frequency channels as those used to divide the speech signal. The resulting bands of amplitude-modulated noise were combined to create vocoded sentences. Root mean square amplitude was equalized across sentences. These stimuli consisting of only the noise-vocoded signals are described here as the VOC-only stimuli. Fig. 6.1 shows unprocessed and processed versions of a sentence included in this experiment. The top panel shows the waveform and spectrogram of

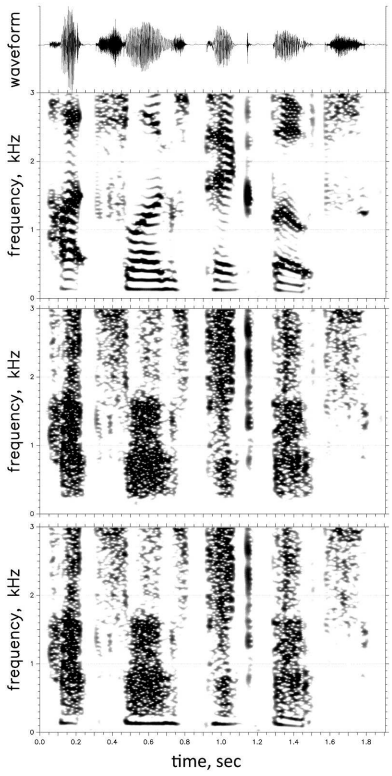

Figure 6.1: Unprocessed and Processed Versions of a Sentence in the EAS Experiment

the unprocessed version. The middle panel shows a spectrogram of the VOC-only version of the sentence shown in the top panel.

Low-frequency signals were created for each sentence using MATLAB. The original sentences were low-pass filtered with a cut-off frequency of 250 Hz, using the same type of digital filter described above. The root mean square amplitude of each low-pass filtered signal was adjusted to equal that of the matching vocoded signal, and those signals were combined to create the stimuli termed LOW-plus in this experiment. The bottom panel of Fig. 6.1 shows a spectrogram of the LOW-plus version of the sentence shown in the top panel.

6.1.4 Procedure

Stimuli for testing were presented under headphones at 68 dB SPL, and all presentation was done in the diotic configuration. Listeners were asked to repeat the sentences they heard. Prior to testing, six practice sentences were presented as training with VOC-only processing. For each training sentence, the unprocessed version was played first, and the listener was asked to repeat it. Then the listener was told that a processed version would be presented. The VOC-only version was presented, and the listener was asked to repeat that version, as well.

During testing, the order of presentation of the sentences was randomized for each listener. Listeners heard half the stimuli in the VOC-only processing condition and half as LOW-plus. The selection of sentences to be presented with each type of processing was randomly made for each listener. Presentation of these two kinds of stimuli was interspersed during testing with the stipulation that no more than two of each processing type could be presented in a row. Each sentence was played once, and the listener repeated it as best as possible. The number of incorrect words for each sentence was recorded by the examiner.

6.1.5 Analyses

The primary dependent measure used was the percentage of words recognized correctly in each processing condition. Data were screened for normal distributions and homogeneity of variance prior to statistical analysis. For inferential tests, arcsine transformations were applied.

6.1.6 Results

Fig. 6.2 shows mean correct word recognition for each group in the top panel, and mean correct sentence recognition for each group in the bottom panel. A two-way, repeated-measures analysis of variance (ANOVA) was performed, with age as the between-subjects factor and processing as the within-subjects factor. The main effect of age was significant, F (2,57) = 111.77, $p < .001$, $\eta^2 = .797$, as were all post hoc contrasts among age groups ($p < .001$ for all contrasts using a Bonferroni adjustment for multiple contrasts). The main effect of processing (VOC-only or LOW-plus) was also significant, F (2,57) = 58.55, $p < .001$, $\eta^2 = .507$. The Age x Processing interaction was not significant. These results indicate that participants were better at recognizing words for the LOW-plus stimuli than for the VOC-only stimuli, and that recognition generally improved with increasing age.

6.1.7 Magnitude of the low-frequency effect across recognition probabilities

Mean recognition probabilities generally differed across listener groups, which complicated the analysis of the low-frequency effect. That situation gives rise to the question of whether a difference between the two processing conditions that is consistent in absolute size across groups represents an effect of adding the low-frequency signal component that is equivalent in magnitude. In order to normalize for recognition probabilities, a metric of effect size was introduced using the formula:

EFFECT = (pLOW-plus pVOC-only)/pVOC-only

where pLOW-plus is the recognition probability for the LOW-plus condition and pVOC-only is the recognition probability for the VOC-only condition.

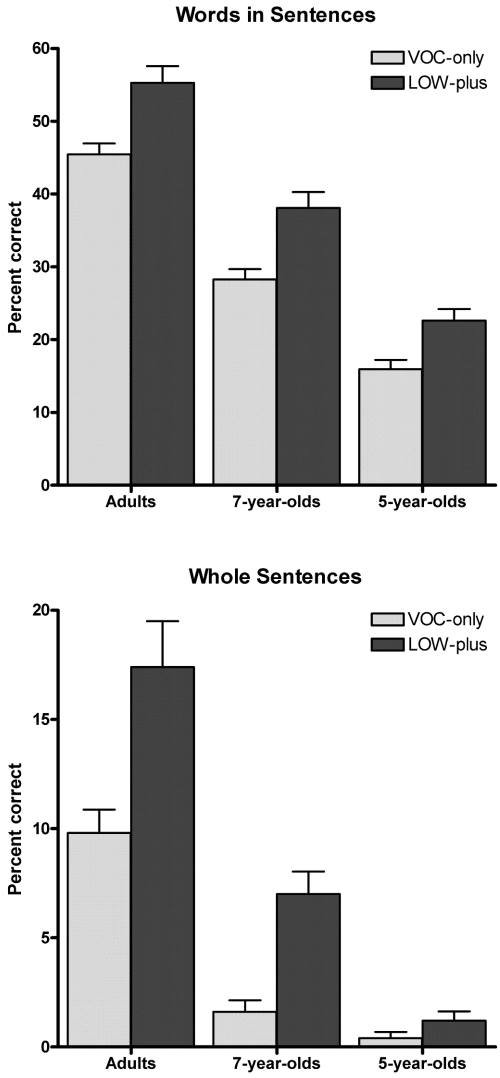

Figure 6.2: Mean Correct Word and Sentence Recognition in the Preliminary Experiment

Mean EFFECT scores (and SDs) were .53 (.52) for 5-year-olds, .39 (.38) for 7-year-olds, and .23 (.23) for adults. However, in spite of the appearance of a developmental trend to smaller EFFECT scores with increasing age, a one-way ANOVA performed on these scores was not significant for age, $F(2,57) = 2.84$, $p = .067$.

6.1.8 Discussion

This preliminary experiment was undertaken to see if a benefit to speech recognition would be observed when a very low-frequency signal is added to a spectrally degraded signal for adults or children, in conditions predicted to strongly facilitate such an effect. Results showed that under these conditions, effects were observed on the order of 20 to 50 percent improvement over what was found for the spectrally degraded signals alone. The results of the preliminary experiment provided the foundation for the main experiment with an expanded design.

6.2 Main EAS Experiment

This experiment was conducted to examine the four hypotheses described in the Introduction. The experimental conditions used in the preliminary experiment were expanded to test each hypothesis. In this main experiment, both a dichotic and diotic configuration were used. Single word materials were included, along with the four-word sentences used in the preliminary experiment. Adults and 7-year-olds participated in this experiment; 5-year-olds were not included because the expanded protocol was too difficult.

6.2.1 Participants

A total of forty listeners were tested in this experiment: 20 adults between the ages of 18 and 38 years; and 20 children between 7 years; 0 months and 7 years; 8 months. All children had fewer than 6 episodes of otitis media before the age of 3 years. All participants (or in the case of children, their parents on their behalf) reported having normal hearing, speech and language. All participants passed hearing screenings of the frequencies of 0.5, 1, 2, 4, and 6 kHz presented at 25 dB hearing level to each ear separately.

subsectionEquipment The same equipment was used as in the preliminary experiment. For this experiment, all test sessions were video recorded so that scoring could be done at a later time. Participants wore Sony FM microphones that transmitted speech signals directly to the line input of the camera in order to ensure good sound quality for all recordings.

6.2.2 Stimuli

Forty-eight of the 50 sentences from the preliminary experiment were used, as well as 16 lists of words from Mackersie, Boothroyd, and Minniear (2001). Each word list consisted of 10 phonetically balanced CVC words. These words were recorded by an adult male speaker of American English at a 44.1-kHz sampling rate with 16 bit digitization, just as the sentences had been. The same signal processing methods as those employed in the preliminary experiment were used to create noise-vocoded and low-frequency signals for all stimuli. These signals were then used to create VOC-only and LOW-plus stimuli that could be presented in both a diotic and dichotic manner. Root mean square amplitude of the vocoded and low-frequency signals was

equalized in all LOW-plus conditions. In the diotic configuration, the same signal was presented to both ears, regardless of whether it was the VOC-only or the LOW-plus stimuli. This configuration simulated bilateral cochlear implants in the diotic VOC-only condition (VOC-only signals to both ears) and bilateral hybrid implants in the diotic LOW-plus condition (combined vocoded and low-frequency signals to both ears). In the dichotic configuration, each ear was presented with a different signal. This configuration simulated a unilateral cochlear implant in the dichotic VOC-only condition (a VOC-only signal to just one ear) and a unilateral cochlear implant with a contralateral hearing aid in the dichotic LOW-plus condition (a VOC-only signal to one ear and the low-frequency signal to the other ear). Half of the listeners in each group heard the VOC-only signals in their right ears in the dichotic condition; the other half heard those signals in their left ears.

Four word lists (40 words total) were selected randomly for presentation in each of the four stimulus conditions (VOC-only diotic, LOW-plus diotic, VOC-only dichotic, LOW-plus dichotic) for each participant. Similarly, twelve sentences were randomly selected for each of the four stimulus conditions, for each participant.

6.2.3 Procedures

There were four sections in the experiment distinguished by configuration and materials: diotic words, dichotic words, diotic sentences, and dichotic sentences. The same number of stimuli of each processing type (VOC-only and LOW-plus) were presented in each section. For the word materials, VOC-only and LOW-plus stimuli alternated between each ten-word list. For the sentences, VOC-only and LOW-plus

stimuli were interspersed, with the only rule being that no more than two VOC-only or LOW-plus stimuli could be presented in a row. Therefore, each word section consisted of 8 word lists (4 VOC-only and 4 LOW-plus), and each sentence section consisted of 24 sentences (12 VOC-only and 12 LOW-plus).

The conditon of stimuli presented in the first section was randomly selected from the four possible types (diotic words, dichotic words, diotic sentences and dichotic sentences), with the stipulation that each starting condition was evenly distributed across subjects in each age group. Section presentation alternated between sentences and words. The second and third sections were the alternative configuration. The fourth section was the same configuration (diotic or dichotic) as the first section.

Each section of testing was preceded by a set of practice stimuli. For the sections consisting of words, the listener first heard and repeated five unprocessed words. Next, 10 processed words (5 VOC-only and 5 LOW-plus) were presented with the software randomly selecting whether VOC-only or LOW-plus stimuli were presented first. For sections consisting of sentences, the listener first heard and repeated two unprocessed sentences. Next, two processed sentences were presented, one VOC-only and one LOW-plus.

6.2.4 Scoring and Analyses

There were four measures used to analyze the results. Dependent measures for the word materials were the percent of phonemes and whole words recognized correctly. Dependent measures for the sentences were the percent of words and whole sentences recognized correctly.

6.2.5 Results

Recognition scores obtained for words, with phoneme scores on the top and whole-word scores on the bottom are shown in Fig. 6.3. Recognition scores obtained for sentences, with word scores on the top and whole-sentence scores on the bottom are shown in Fig. 6.4. To compare the results of this experiment with the preliminary experiment, t tests were performed on scores from comparable conditions between experiments, for adults and children separately. None of the tests resulted in a statistically significant difference. Therefore, it was concluded there was adequate inter-subject reliability.

Another question analyzed was whether differences were found in recognition probabilities dependent upon which ear heard the VOC-only signal in the dichotic configuration. In order to answer this question, t tests were performed for adults and children separately, with groups defined by which ear was presented with the VOC-only stimuli. Tests were done separately on all four measures obtained, for both the VOC-only and LOW-plus processing conditions, making a total of eight t tests per age group. For children, none of these t tests resulted in statistically significant outcomes, so it was concluded that there was no effect of which ear heard the VOC-only signal in the dichotic configuration for children. For adults, none of the tests for sentence materials resulted in a statistically significant ear effect. This was also true for the tests involving LOW-plus stimuli for word materials. However, for the VOC-only stimuli involving isolated words, statistically significant differences in recognition were found based on which ear received the VOC-only signal: for phonemes in words, $t(18) = 3.28$, $p = .004$; for whole words, $t(18) = 2.28$, $p = .035$. Mean recognition scores are shown in Table 6.1 for VOC-only, isolated word stimuli in the dichotic

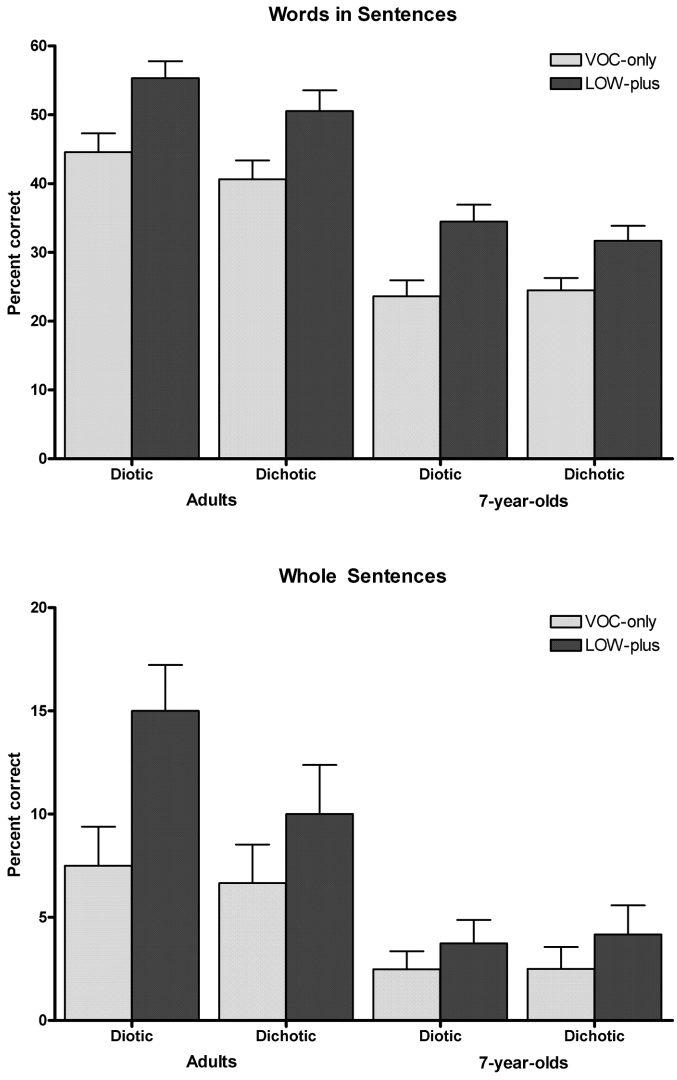

Figure 6.3: Mean Correct Phoneme and Whole-Word Scores for Words Materials in the Main Experiment

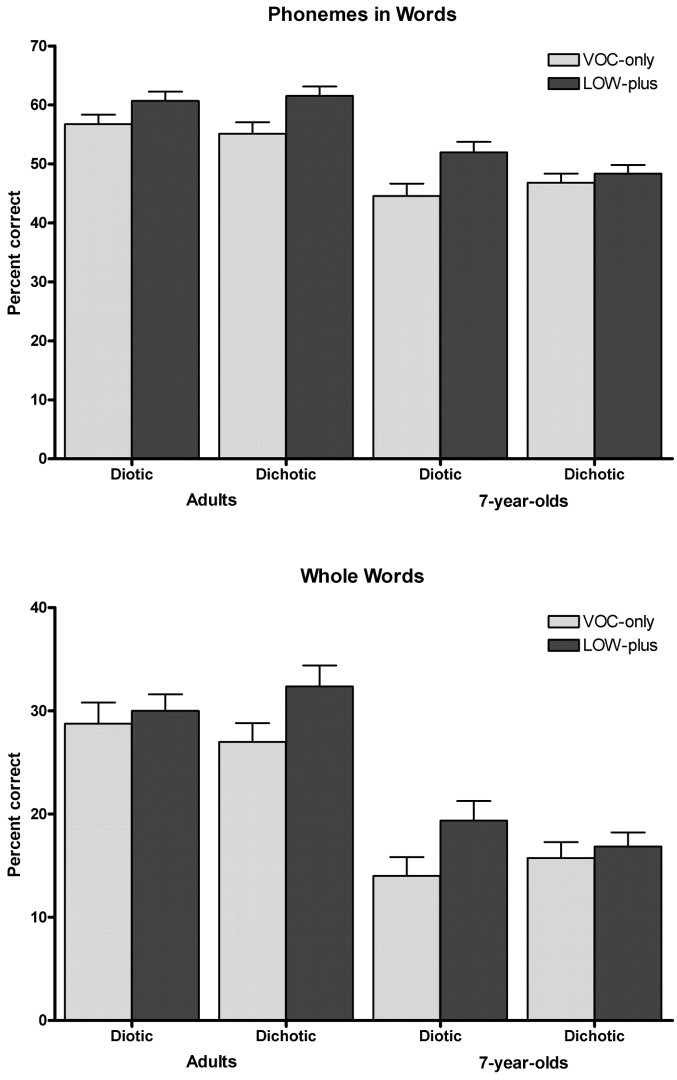

Figure 6.4: Mean Correct Word and Whole-Sentence Scores for Sentence Materials in the Main Experiment

Table 6.1: Means Recognition Probabilties for Adults in the Main Experiment, for VOC-only, Isolated Word Stimuli Presented in the Dichotic Configuration

	Right Ear	Left Ear
Phonemes in words	60.33 (6.28)	49.93 (7.82)
Whole Words	30.75 (6.88	23.25 (7.82)

configuration. Because there was no significant difference for almost all of the t tests, scores in the dichotic configuration were collapsed across listeners, regardless of which ear heard the VOC-only signal.

6.2.6 Separate analyses on dependent measures

Separate analyses were performed on the four dependent measures separately: phonemes in words, whole words, words in sentences, and whole sentences. For each score, a three-way, repeated-measures ANOVA was performed, with age as the between-subjects factor and processing and configuration as within-subjects factors. Results of the analyses for phonemes in word materials are shown in Table 6.2. Similarly, results of the analyses for whole words are shown in Table 6.3. The primary finding for these word materials were that adults performed better than 7-year-olds. Additionally, scores were better in the LOW-plus processing condition than in the VOC-only processing condition. However, the main effect of configuration was not significant. None of the two-way interactions reached statistical significance, but there was a significant three-way interaction for both phoneme and whole word scores. Examination of Fig. 6.3 reveals that it was due to adults showing a larger benefit of the

Table 6.2: Statistical Outcomes of Three-way ANOVAs for Phonemes in Words Materials

Phonemes in Words			
	F	p	η^2
Age	31.05	$< .001$	.450
Processing	44.26	.001	.538
Configuration	0.15	NS	
Age x Processing	0.24	NS	
Age x Configuration	0.01	NS	
Processing x Configuration	1.67	NS	
Age x Proc x Config	9.68	.004	.203

Table 6.3: Statistical Outcomes of Three-way ANOVAs for Whole Words

Whole Words			
	F	p	η^2
Age	54.56	$< .001$	.589
Processing	13.28	.001	.259
Configuration	0.02	NS	
Age x Processing	0.31	NS	
Age x Configuration	0.001	NS	
Processing x Configuration	0.17	NS	
Age x Proc x Config	5.38	.026	.124

low-frequency signal in the dichotic than the diotic configuration, and children showing the opposite pattern: a stronger benefit of the low-frequency signal in the diotic than the dichotic configuration. Effect sizes were small for this three-way interaction for both scores, so it is difficult to conclude very much of this finding.

Table 6.4 shows the outcomes of statistical analyses for words in sentence materials. Table 6.5 shows the outcomes of statistical analyses for whole sentences. The

Table 6.4: Statistical Outcomes of Three-way ANOVAs for Words in Sentence Materials

Words in Sentences			
	F	p	η^2
Age	74.18	$< .001$	.664
Processing	66.02	$< .001$	.635
Configuration	1.28	NS	
Age x Processing	0.03	NS	
Age x Configuration	0.67	NS	
Processing x Configuration	1.12	NS	
Age x Proc x Config	0.56	NS	

Table 6.5: Statistical Outcomes of Three-way ANOVAs for Whole Sentences

Whole Sentences			
	F	p	η^2
Age	24.91	$< .001$	.396
Processing	20.60	$< .001$	.352
Configuration	0.85	NS	
Age x Processing	4.10	.050	.097
Age x Configuration	0.79	NS	
Processing x Configuration	1.10	NS	
Age x Proc x Config	1.79	NS	

only effects that were significant were those of age and processing. A significant Age x Processing interaction for whole sentences. As with the word materials, the effect size is small.

6.2.7 Words versus sentences: The materials' effect

A four-way ANOVA was performed on word scores to obtain a test of the materials' effect: percent correct recognition for whole words (bottom of Fig. 6.3) and words in sentences (top of Fig. 6.4). Results of this analysis revealed significant main effects of age, F (1,38) = 99.42, $p < .001$, $\eta^2 = .723$, and processing, F (1,38) = 67.95, $p < .001$, $\eta^2 = .641$, matching what had been found for the separate analyses. The main effect of configuration was not statistically significant. However, the main effect of materials was significant, F (1,38) = 159.49, $p < .001$, $\eta^2 = .808$, indicating that words were more readily recognized in a sentence context than in isolation. The only two-way interaction that was significant was the Materials x Processing, F (1,38) = 11.77, $p = .001$, $\eta^2 = .236$. Examination of Fig. 6.3 and 6.4 reveals that this outcome was due to the low-frequency effect being greater when words were presented in sentences, rather than in isolation. Finally, the three-way interaction of Age x Processing x Configuration was significant, F (1,38) = 6.27, $p = .017$, $\eta^2 = .142$. That is the same three-way interaction that was significant for both measures obtained for the word lists. In this case, examination of the figures suggests that the interaction was due entirely to the age effect found for word lists.

6.2.8 Magnitude of the low-frequency effect across recognition probabilities

The EFFECT scores were computed as they had been in the preliminary experiment. This was done to see if the magnitude of the low-frequency effect differed for adults and children, once the overall difference in recognition probabilities was taken into account. In this main experiment, two kinds of materials were used, and stimuli

were presented in two configurations. Thus, four EFFECT scores were computed, using recognition scores for whole-word and words in sentences, and scores for the diotic and dichotic configurations. From these four scores, two summary EFFECT scores were computed for whole words and words in sentences by taking the means across the diotic and dichotic configurations.

The EFFECT means (and SDs) scores for whole words were .50 (.73) for 7-year-olds and .19 (.31) for adults. A t test performed on these scores failed to show a significant age effect, t (38) = 1.79, p = .081. For words in sentences, mean EFFECT scores were .57 (.79) and .29 (.21) for 7-year-olds and adults, respectively. This difference was not statistically significant. Thus it was concluded that there was a developmental trend to smaller low-frequency effects, but it was not statistically significant.

6.2.9 Discussion

Cochlear implants have significantly improved the communication capabilities of patients with severe-to-profound hearing loss. However, opportunities for improvement remain. For instance, the listening abilities of listeners with cochlear implants greatly vary. The best-performing patients do quite well with implants, but many other patients continue to struggle with speech recognition. As a consequence, research efforts have continued to investigate improvements for more effective implantation and signal processing strategies for those implants. Another approach that has been taken to try to improve speech recognition is to combine acoustic and electric signal with hybrid implants. These efforts have focused on patients with substantial residual hearing. The potential benefit of the residual hearing is that more detailed

access to linguistically significant information is possible, but also that the residual hearing may help a listener better perceptually organize the signals provided electrically.

The work reported in the current study investigated the potential advantage of combining a low-frequency acoustic signal with a spectrally degraded electric signal. In this work, the possibility was tested that even a very low-frequency signal would have beneficial effects on the recognition of spectrally degraded speech. The primary hypothesis tested in this work was that the low-frequency signal portion would facilitate better perceptual organization of the spectrally degraded speech signal. The low-frequency signal provides minimal linguistic information, mostly about voicing, as well as about word segmentation and sentence prosody.

In total, four specific hypotheses were tested:

1. Adding the very low-frequency component of the speech signal to an implant-simulated signal would improve speech recognition.

2. The advantage would be greater in magnitude when both signal components were presented diotically, rather than dichotically.

3. The advantage would be greater in magnitude for words presented in sentences rather than in isolation.

4. Children would demonstrate a greater advantage than adults.

These hypotheses were tested using conditions of signal processing: noise-vocoded signals (high-pass filtered above 250 Hz), and noise-vocoded plus low-frequency signals. Two types of materials were used: isolated words and four-word sentences. Two

listening configurations were used: diotic and dichotic. Two age groups were used: adults and 7-year-olds.

Results for both sets of speech materials and both configurations showed an advantage for the addition of the very low-frequency signal component. This effect improved recognition by as much as 60 percent over what was obtained for the noise-vocoded signals alone. These outcomes suggest that even if patients with hearing loss can hear only the lowest frequencies in the speech signal, there is likely an advantage to be obtained to amplifying those frequencies with hearing aids or with a hybrid implant.

Another question in the experiment was whether the benefit of acoustic hearing was possible in both diotic and dichotic configurations. It was hypothesized that listeners may not perceptually integrate the different types of signals in the dichotic configuration. However, equivalent effects were found in both configurations suggesting that listeners integrated the signals in both configurations.

It was hypothesized that the magnitude of the effect of adding the low-frequency signal would be greater for word recognition when words were heard in sentences rather than in isolation. This hypothesis was supported by the finding of a significant Materials x Processing interaction for word recognition scores. It is likely the case that longer sequences of materials make it easier to achieve the perceptual organization required to recover speech-like form.

Finally, adults and children were compared as listeners in this experiment to investigate whether there were any age related differences. Absolute differences in recognition scores between VOC-only and LOW-plus conditions were consistent across listener age, in both the preliminary and main experiments. However, when those

differences are transformed to proportions of recognition scores for the VOC-only condition, it appears as if children gain more by the addition of the low-frequency signal component. This is also supported in the fact that significant correlation coefficients were obtained between the EFFECT scores and recognition scores for the VOC-only stimuli, both across and within listener groups. Therefore, it seems in general that the poorer listeners' abilities were to recognize the spectrally degraded speech signals, the more they benefited from the addition of the low-frequency signal component, regardless of age.

6.2.10 Clinical implications and limitations

There may be several clinical implications based on the results in this study. Simulations of signal processing with cochlear implants were created by using noise-vocoded signals. Therefore, the signals were spectrally degraded and lacked the kind of detailed frequency structure that is typically used in speech recognition. The signals in this experiment were in a similar frequency range as what is typically used with cochlear implants, and also for hearing aids used by patients with residual hearing.

The results of this experiment suggest a benefit of combining acoustic stimulation with electric stimulation either with a hyprid implant or with a contralateral hearing aid for unilaterally implanted patients. The results from the experiment are important because they suggest that even patients with profound hearing loss might be candidates for electric-acoustic stimulation. Additionally, it seems patients who have the most difficulty recognizing speech with their cochlear implants might have the most to gain by the addition of low-frequency acoustic amplification. Finally, both

adults and children benefited from perceptually integrating the very low-frequency signal with their cochlear implant signal.

The finding of a strong effect of the materials used has implications for clinical testing. Audiological evaluations typically use lists of words presented individually. That method may not realize the benefit found with sentence materials that individual patients might demonstrate in the clinic when a very low-frequency acoustic signal is presented with the electric signal they get through their cochlear implants. Thus, sentence materials would be included for evaluating whether a patient might benefit from electric-acoustic stimulation.

6.2.11 Conclusions

The experiments reported here were conducted to investigate issues related to the combination of a very low-frequency signal with a spectrally degraded signal, as might be the configuration for a patient with profound hearing loss who uses a cochlear implant and a hearing aid. Results showed that both pediatric and adult patients can benefit from this kind of configuration, and patients with the poorest speech recognition using a cochlear implant could likely benefit the most from the addition of a very low-frequency signal. In this study, it was found that speech recognition can be improved combining acoustic and electric signal types by as much as 60 percent. These benefits can be obtained in both diotic and dichotic configurations. Therefore, both hybrid and bimodal stimulation can support the kind of integration required to achieve the benefit of low-frequency signals. Finally, evidence from this study suggests that the benefit of the very low-frequency signal is a result of the increased ability by listeners to perceptually organize spectrally degraded signals.

CHAPTER 7

Conclusion

7.1 Pitch Synchronous Processing

A pitch synchronous framework for representing acoustic characteristics of speech has been presented. The primary application of this signal processing it to investigate the perceptual importance of the acoustic characteristics of speech signals. There may also be opportunities for enhancing the perception of speech, with one example being processing the spectral envelope of speech to sharpen formant peaks.

Signal analysis and decomposition were motivated by the descriptions of temporal *envelope, fine structure,* and *periodicity* provided in Rosen (1992). The PS representations are alternatives to the HT method derived from an analytic signal of decomposing a signal into an envelope and instantaneous frequency. PS Temporal Amplitude Normalization was presented to recover and remove the GTE from a speech signal. This technique can be used to produce an amplitude envelope and phase signal pair that satisfies $x[t] = \mathbb{R}\{A[t]e^{j\phi[t]}\}$ while estimating aspects of speech production from a signal. HT processing violated physical conditions presented in Loughlin and Tacer (1996) for representing a signal as $A[t]$ and $\phi[t]$. PS processing could be used to decompose a signal while satisfying these physical conditions. PS Spectral Amplitude Normalization was presented to recover and remove the GSE

from a speech signal. This technique can be used to produce a spectral envelope and source signal pair that satisfies $x[t] = A_1[t](h_1[t] * u_1[t])$. Methods were presented to replace each acoustic characteristic for voiced portions of speech.

A technique to process the spectral envelope of speech was presented by changing the relative amplitude of individual harmonics in the signal. Sharpening the resonances of the spectral envelope may improve intelligibility for certain listeners or listening conditions. Listeners with sensorineural hearing loss have been found to have broadened auditory filters compared to listeners with normal hearing. Sharpening the spectral envelope of speech may help compensate for the broadened auditory filters. Additionally, sharpening the spectral envelope of speech in the presence of noise may improve intelligibility by helping differentiating the speech signal from the noise.

7.2 Perceptual Experiments

7.2.1 Temporal Envelope

The perceptual importance of the temporal envelope in speech signals was investigated for listeners with normal hearing and listeners with cochlear implants. It was found that these listening groups use different perceptual weighting strategies when listening to speech signals. Listeners with normal hearing are very sensitive to changes in a signal's temporal envelope and spectral envelope. However, they perceptually weight the spectral envelope more than the temporal envelope when listening to speech signals. Listeners with cochlear implants are less sensitive to changes in a signal's temporal envelope and spectral envelope. The perceptual weighting strategies for listeners with cochlear implants are much less consistent than listeners with

normal hearing. This is likely a result of the diminished sensitivity to these acoustic characteristics.

7.2.2 Spectral Envelope

The perceptual importance of the spectral envelope in speech signals presented in noise was investigated for adults and child listeners. The major outcome of this study was that young children showed greater perceptual coherence than adults, and were less readily perturbed from integrating spectral components in that way than were adults. The effect for children was not restricted to conditions in which all components shared a harmonic relationship, or even to conditions in which all components had harmonic structure. It was more related to listener characteristics than to signal characteristics. These outcomes support the hypothesis that young children are more strongly obliged than adults to fuse spectral components when those components are recognized as being part of a speech signal.

7.2.3 Periodicity

The perceptual importance of periodicity was investigated for cochlear implant simulated speech. The experiments reported here were conducted to investigate issues related to the combination of a very low-frequency signal with a spectrally degraded signal, as might be the configuration for a patient with profound hearing loss who uses a cochlear implant and a hearing aid. Results showed that both children and adult patients can benefit from this kind of configuration, and patients with the poorest speech recognition using a cochlear implant could likely benefit the most from the addition of a very low-frequency signal. In this study, it was found that speech recognition can be improved combining acoustic and electric signal types by

as much as 60 percent. These benefits can be obtained in both diotic and dichotic configurations. Therefore, both hybrid and bimodal stimulation can support the kind of integration required to achieve the benefit of low-frequency signals. Evidence from this study suggests that the benefit of the very low-frequency signal is a result of the increased ability by listeners to perceptually organize spectrally degraded signals.

7.3 Future Work

There are several applications of PS that remain to be considered. The possibility of enhancing speech perception by processing the spectral envelope can be pursued using listening experiments. These experiments should include listeners with normal hearing and listeners with sensorineural hearing loss. Speech signals should be presented with noise in some conditions. It may also be possible that children and adults perform differently when spectral resonances are processed. Lastly, it may be useful to investigate both the detection of changes in the spectral envelope, along with any benefit to intelligibility.

There are also possibilities to consider PS processing methods in a multi-band analysis. The PS temporal envelope processing methods presented previously were based on a single-channel, wide-band analysis of the speech signal. This was compared with auditory chimera processing using the HT on each channel of a filter bank. PS processing could be used in a multi-channel vocoder to continue investigating speech perception by cochlear implant simulations. For instance, there are cochlear implant processing strategies where the stimulation rate is determined by the pitch of the speech signal. A PS vocoder may be one option to simulate this processing strategy for listeners with normal hearing.

The perceptual experiments presented previously have several extensions that could be continued. Regarding coherence masking protection, including listeners with hearing aids and cochlear implants may reveal new conclusions about spectral integration. It may be possible that hearing aids which apply frequency compression disrupt spectral integration for experienced listeners. Separately, listeners with cochlear implants may show different perceptual strategies of integration compared to listeners with normal hearing. Testing listeners with cochlear implants on both the synthetic speech stimuli and sine-wave stimuli may be one way to investigate these differences. There are also other conditions that could be considered for listeners with normal hearing, including: dichotic target and cosignal, vocoded signals, various types of noise other than white noise,

Regarding the electric and acoustic stimulation experiment, it may be useful to further investigate the extent that a low-frequency signal to improve perception for degraded signals. For instance, the cut-off frequency could be varied to stimulate listeners with different degrees of hearing loss. Also, it seemed the time-varying nature of the low-frequency signal contributes to the perceptual organization of the vocoded signal. Rather than only investigating the fundamental frequency as the time-varying component of the signal, a listening experiment using the lowest formant as a time-varying component could be conducted to measure changes in intelligibility. This may lead to new ideas about how an electric and acoustic signals should be presented for listeners with bimodal auditory prostheses.

BIBLIOGRAPHY

[Alku(1992)] P. Alku. "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," Speech Commun. **11**, 109–118 (1992).

[Alwan(2013)] A. Alwan. "UCLA SPAPL Consonant Vowel Database," `http://www.ee.ucla.edu/~spapl/cv.html`, (April 3rd, 2013).

[Bacon and Gleitman(1992)] S. Bacon, R. Gleitman. "Modulation detection in subjects with relatively flat hearing losses," J. Speech. Hear. Res. **35**, 642–653 (1992).

[Barker and Cooke(1999)] J. Barker, M. Cooke. "Is the sine-wave speech cocktail party worth attending?," Speech Commun. **27**, 159–174 (1999).

[Baer, Moore, and Gatehouse (1993)] T. Baer, B. Moore, S. Gatehouse, "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response time," J. of Rehab. Res. and Develop. **30**, 49–72 (1994).

[Bladon and Lindblom(1981)] R. Bladon, B. Lindblom. "Modeling the judgement of vowel quality differences," J. Acout. Soc. Amer. **97**, 585–592 (1981).

[Bregman(1990)] A.S. Bregman. 1990, *Auditory Scene Analysis:The Perceptual Organization of Sound,* Cambridge, Massachusetts: The MIT Press.

[Carney, Widin, and Viemeister(1977)] A. Carney, G. Widin, N. Viemeister. "Noncategorical perception of stop constants differing in VOT," J. Acoust. Soc. Amer. **62**, 961-970 (1977).

[Chen(1970)] M. Chen. "Vowel length variation as a function of the voicing of the consonant environment," Phonetica **22**, 129-159 (1970).

[Childers(1987)] Childers, D. (2000), *Speech Processing and Synthesis Toolboxes*, (Wiley & Sons Inc.)

[Crowther and Mann(1992)] C. Crowther, V. Mann. "Native language factors affecting use of vocalic cues to final consonant voicing in English," J. Acoust. Soc. Amer. **92**, 711-722 (1992).

[Crowther(1994)] C. Crowther, V. Mann. "Use of vocalic cues to consonant voicing and native language background: The influence of experimental design," Perception & Psychophysics, **55**, 513-525 (1994).

[Dorman and Gifford(2010)] M. Dorman and R. Gifford. "Combining acoustic and electric stimulation in the service of speech recognition," International Journal of Audiology **49**, 912-919 (2010).

[Dubno and Dorman(1987)] J.R. Dubno, M.F. Dorman. "Effects of spectral flattening on vowel identification," J. Acoust. Soc. Amer. **82**, 1503–1511 (1987).

[Duquesnoy(1983)] A.J. Duquesnoy. "Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons," J. Acout. Soc. Amer. **74**, 739–743 (1983).

[Dwyer and Nelson(1992)] S. Dwyer, P. Nelson. "The effect of intensity on gap detection in hearing-impaired listeners," J. Acout. Soc. Amer. **92**, 2386A (1992).

[Ercelebi(2003)] E. Ercelebi. "Second generation wavelet transform-based pitch period estimation and voiced/unvoiced decision for speech signals," Appl. Acoust. **64**, 25–41 (2003).

[Feng et al.(2012)] Y. Feng, L. Xu, N. Zhou, G. Yang, S. Yin. "Sine-wave speech recognition in a tonal language," J. Acoust. Soc. Amer. **131**, EL133–EL138 (2012).

[Flanagan and Golden(1966)] J.L. Flanagan, R.M. Golden. "Phase Vocoder," pp. 1493–1509,1966, *The Bell System Technical Journal*.

[Flanagan(1980)] J.L. Flanagan. "Parametric coding of speech spectra," J. Acout. Soc. Amer. **68**, 412–419 (1980).

[Florentine and Buus(1984)] M. Flanagan, S. Buus. "Temporal gap detection in sensorineural and simulated hearing impairments," J. Speech. Hear. Res. **27**, 449–455 (1984).

[Flege and Wang(1989)] J. Flege, C. Wang. "Native-language phonotactic constraints affect how well Chinese subjects perceive the word-final English /t/-/d/ contrast," Journal of Phonetics **17**, 299-315 (1989).

[Gold and Rabiner(1964)] B. Gold, L. Rabiner. "Parallel processing techniques for estimating pitch periods of speech in the time domain," J. Acoust. Soc. Amer. **46**, 442–448 (1964).

[Goldman and Fristoe(2000)] R. Goldman, M. Fristoe, 2000. *Goldman Fristoe 2: Test of Articulation,* Circle Pines, MN: American Guidance Service, Inc.

[Goswami et al.(2002)] U. Goswami, J. Thomson, U. Richardson, R Stainthorp, D. Hughes, S. Rosen, and S. Scott. "Amplitude envelope onsets and developmental dyslexia: A new hypothesis," Proceedings of the National Academy of Sciences of the United States of America **99**, 10911–10916 (2002).

[Goswami et al.(2010)] U. Goswami, T. Fosker, M. Huss, N. Mead, D. Szucs. "Rise time and formant transition duration in the discrimination of speech sounds: the Ba-Wa distinction in developmental dyslexia," Develop. Science **14**, 34–43 (2010).

[Gordon(1997)] P. Gordon. "Coherence masking protection in speech sounds: The role of formant synchrony," Perception & Psychophysics, **59**, 232-242 (1997).

[Gordon(2000)] P. Gordon. "Masking protection in the perception of auditory objects," Speech Communication **30**, 197-206 (2000).

[Grose and Hall(1992)] J. Grose, J. Hall. "Comodulation masking release for speech stimuli," J. Acout. Soc. Amer. **91**, 1042-1050 (1992).

[Hall and Grose(1990)] J. Hall, J. Grose. "Comodulation masking release and auditory grouping," J. Acout. Soc. Amer. **88**, 119-125 (1990).

[Hawks et al. (1997)] J. Hawks, M. Fourakis, M. Skinner, T. Holden, L. Holden. "Effects of formant bandwidth on the identification of synthetic vowels by cochlear implant recipients," Ear Hear. **18**, 479–487 (1997).

[Holt and Lotto (2005)] R. Holt, A. Carney. (2005) "Multiple looks in speech sound discrimination in adults," J. Speech, Lang., and Hear. Res. **48**, 922-943 (2005).

[Hopkins and Moore(2007)] K. Hopkins, B.C.J. Moore. "Moderate cochlear hearing loss leads to a reduced ability to use temporal fine structure information," J. Acout. Soc. Amer. **122**, 1055–1068,(2007).

[Hopkins and Moore(2009)] K. Hopkins, B.C.J. Moore. "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," J. Acout. Soc. Amer. **125**, 442–446 (2009).

[Horii, House, and Hughes(1971)] Y. Horii, A.S. House, G.W. Hughes. "A masking noise with speech-envelope characteristics for studying intelligibility," J. Acoust. Soc. Amer. **49**, 1849–1856 (1971).

[Ihlefeld and Shinn-Cunningham(2008)] A. Ihlefeld, A. and B. Shinn-Cunningham. "Spatial release from energetic and informational masking in a selective speech identification task." J. Acoust. Soc. Am. **123**, 4369-4379 (2008).

[Kadambe(1992)] S. Kadambe. "Application of wavelet transform for pitch detection of speech signals," Information Theory, IEEE Transactions on **38**, 917–924 (1992).

[Kim and Chung(2004)] Y.J. Kim, J.H. Chung. "Pitch synchronous cepstrum for robust speaker recognition over telephone channels," Electronics Letters **40**, 207–209 (2004).

[Kuwabara(1984)] H. Kuwabara. "A pitch-synchronous analysis/synthesis system to independently modify formant frequencies and bandwidths for voiced speech," Speech Commun. **3**, 211–220 (1984).

[Leek, Dorman, and Summerfield(1987)] M.R. Leek, M.F. Dorman, Q. Summerfield. "Minimum spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Amer. **81**, 148–154 (1987).

[Levitt(1971)] H. Levitt. "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Amer., 49, 467-477 (1971).

[Li, Nie, Atlas, and Rubinstein(2010)] X. Li, K. Nie, L. Atlas, J. Rubinstein. "Harmonic coherent demodulation for improving sound coding in cochlear implants," Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 14-19 March 2010, 5462–5465 (2010).

[Liberman, Isenberg, and Rakerd(1981)] A. Liberman, D. Isenberg, B. Rakerd. "Duplex perception of cues for stop consonants: Evidence for a phonetic mode." Perception & Psychophysics **30**, 133-143 (1981).

[Licklider and Pollack(1948)] J.C.R. Licklider, I. Pollack. "Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of Speech," J. Acoust. Soc. Amer. **20**, 42–51 (1948).

[Lorenzi, Gilbert, Carn, Garnier, and Moore(2006)] C. Lorenzi, G. Gilbert, H. Carn, S. Garnier, B.C.J. Moore. "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proceeding of the National Academy of Sciences **103**, 18866–18869 (2006).

[Loughlin and Tacer(1996)] P.J. Loughlin, B. Tacer. "On the amplitude- and frequency-modulation decomposition of signals," J. Acoust. Soc. Amer. **100**, 1594–1601 (1996).

[Mackersie, Boothroyd, and Minniear(2001)] C. Mackersie, A. Boothroyd, D. Minniear. "Evaluation of the Computer-Assisted Speech Perception Assessment Test (CASPA)," J Am. Acad. Audiol. **12**, 390-396 (2001).

[Macmillan and Creelman(2005)] N. Macmillan, C. Creelman, 2005. *Detection Theory: A Users Guide.* Mahwah, NJ: Lawrence Earlbaum Associates.

[Makhoul(1976)] J. Makhoul. "Method for nonlinear spectral distortion of speech signals." Proc. ICASSP, 87-90, (1976).

[Mann, and Liberman(1983)] V. Mann, A. Liberman. "Some differences between phonetic and auditory modes of perception," Cognition **14**, 211-235 (1983).

[Mathews, Miller, and David(1961)] M.V. Mathews, J.E. Miller, E.E. David, Jr. "Pitch synchronous analysis of voiced sounds," J. Acoust. Soc. Amer. **33**, 179–186 (1961).

[McMurray, Tanenhaus, and Aslin(2002)] B. McMurray, M. Tanenhaus, R. Aslin. "Gradient effects of within-category phonetic variation on lexical access," Cognition **86**, B33-B42 (2002).

[Menell, McAnally, Stein(1999)] P. Menell, K.I. McAnally, J.F. Stein. "Psychophysical sensitivity and physiological response to amplitude modulation in adult dyslexic listeners," J. Speech, Lang., and Hear. Res. **42**, 797–803 (1999) .

[Milenkovic(2004)] P. Milenkovic, 2004. TF32 [Computer software]. University of Wisconsin-Madison.

[Miyawaki et al.(1975)] K. Miyawaki, W. Strange, R. Verbrugge, A. Liberman, J. Jenkins, and O. Fujimura. "An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English," Perception & Psychophysics **18**, 331-340 (1975).

[Moore(2008)] B. Moore. "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people," J. Assoc. Res. Oto. **9**, 399–406 (2008).

[Morrongiello et al.(1984)] B. Morrongiello, R. Robson, C. Best, R. Clifton. "Trading relations in the perception of speech by 5-year-old children," Journal of Experimental Child Psychology **37**, 231-250 (1984).

[Moulines and Charpentier(1990)] E. Moulines, F. Charpentier. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**, 453–467 (1990).

[Neely and Peters(1992)] S. Neely, J. Peters, 1992. WavEd User's Guide. (Tech. Memo. 15) Omaha, NE: Boys Town National Research Hospital.

[Nittrouer(1992)] S. Nittrouer. "Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries," Journal of Phonetics **20**, 351-382 (1992).

[Nittrouer(2004)] S. Nittrouer. "The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults," J. Acoust. Soc. Amer. **115**, 1777-1790 (2004).

[Nittrouer and Crowther(2001)] S. Nittrouer, C. Crowther. "Coherence in children's speech perception," J. Acoust. Soc. Amer. **110**, 2129-2140 (2001).

[Nittrouer and Studdert-Kennedy(1986)] S. Nittrouer, M. Studdert-Kennedy. "The stop-glide distiction: Acoustic analysis and perceptual effect of variation in syllable amplitude envelope for initial /b/ and /w/," J. Acoust. Soc. Amer. **80**, 1026–1029 (1986).

[Nittrouer and Lowenstein(2010)] S. Nittrouer, J. Lowenstein. "Learning to perceptually organize speech signals in native fashion," J. Acoust. Soc. Amer. **127**, 1624–1635 (2010).

[Nittrouer and Tarr(2011)] S. Nittrouer, E. Tarr. "Coherence masking protection for speech and non-speech signals in children and adults," Atten. Percept. Psychophys. **73**, 2606–2623 (2011).

[Nittrouer, Lowestein, and Packer(2009)] S. Nittrouer, J. Lowenstein, and R. Packer. "Children Discover the Spectral Skeletons in Their Native Language Before the Amplitude Envelopes." Journal of Experimental Psychology Human Perception and Performance **35**, 1245-1253 (2009).

[Nittrouer, Lowenstein, and Tarr(2012)] S. Nittrouer, J.H. Lowenstein, and E. Tarr. "Amplitude rise time does not cue the /ba/-/wa/ contrast for adults or children," J. Speech Lang. Hear. Res. first published on September 19,2012 as doi:10.1044/1092-4388(2012/12-0075).

[Nittrouer, Manning, and Meyer(1993)] S. Nittrouer, C. Manning, G. Meyer. "The perceptual weighting of acoustic cues changes with linguistic experience," J. Acout. Soc. Amer. **94**, S1865 (1993).

[O'Shaughnessy(1987)] D. O'Shaughnessy. "Linear predictive coding," Potentials, IEEE **7**, 29–32 (1988).

[Paliwal and Rao(1981)] K.K. Paliwal, P.V.S. Rao. "A modified autocorrelation method of linear prediction for pitch-synchronous analysis of voiced speech," Signal Process. **3**, 181–185 (1981).

[Parker and Hall(1979)] S. Parker, , and G. Hall. "Computer modeling of voice signals for adjustable pitch and formant frequencies." 13th Asilomar Conf. on Circuits, Systems, and Computers, 158-161 (1979).

[Peterson and Barney(1952)] G.E. Peterson, H.L. Barney. "Control methods used in a study of the vowels," J. Acout. Soc. Amer. **24**, 175–184 (1952).

[Peterson and Lehiste(1960)] G. Peterson, I. Lehiste. "Duration of syllable nuclei in English," J. Acout. Soc. Amer. **32**, 693-703 (1960).

[Plomp(1967)] R. Plomp. "Pitch of complex tones," J. Acout. Soc. Amer. **41**, 1526–1533 (1967).

[Rabiner(1977)] L.R. Rabiner. "On the use of autocorrelation analysis for pitch detection," Acoustics, Speech, and Signal Processing, IEEE Transactions on **25**, 24–33 (1977).

[Rabiner, Cheng, Rosenberg, and McGonegal(1976)] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, C.A. McGonegal. "A comparative performance study of several pitch detection algorithms," Audio, Speech, and Language Processing, IEEE Transactions on **24**, 399–418 (1976).

[Raphael(1972)] L. Raphael. "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," J. Acout. Soc. Amer. **51**, 1296-1303 (1972).

[Relkin and Turner(1988)] E. Relkin, C. Turner. "A reexamination of forward masking in the auditory nerve," J. Acout. Soc. Amer. **84**, 584–591 (1988).

[Remez et al.(1981)] R.E. Remez, P.E. Rubin, D.B. Pisoni, T.D. Carell. "Speech perception without traditional speech cues," Science **212**, 947–950 (1981).

[Risberg and Agelfors(1982)] A. Risberg, E. Agelfors, 1982. Speech perception based on non-speech signals. In R. Carlson and B. Granstrom (eds.), *The Representation of Speech in the Peripheral Auditory System*. Elsevier Biomedical Press, pp. 209-215.

[Rocheron, Lorenzi, Fullgrabe, and Dumont(2002)] I. Rocheron, C. Lorenzi, C. Fullgrabe, and A. Dumont. "Temporal envelope perception in dyslexic children," Neuroreport **13**, 1683–1687 (2002).

[Rosen(1992)] S. Rosen. "Temporal information in speech: acoustic, auditory, and linguistic aspects," Phil. Trans. R. Soc. Lond. **336**, 367–373 (1992).

[Scharf(1970)] B. Scharf. 1970 "Critical Bands," *Foundations of modern auditory theory,* New York: Academic Press.

[Shannon, Zeng, Kamath, Wygonski, and Ekelid(1995)] R.V. Shannon, F.G. Zeng, V. Kamath, J. Wygonski, M. Ekelid. "Speech recognition with primarily temporal cues," Science **270**, 303–304 (1995).

[Smith, Delgutte, and Oxenham(2002)] Z.M. Smith, B.Delgutte, A.J. Oxenham. "Chimaeric sounds reveal dichotomies in auditory perception," Nature **416**, 87–90 (2002).

[Speeter Beddor and Hawkins(1991)] P. Speeter Beddor, S. Hawkins. "The Influence of Spectral Prominence on Perceived Vowel Quality," Haskins Laboratories Status Report of Speech Research **105**, 187–214 (1991).

[Tallal(1980)] P. Tallal. "Auditory temporal perception, phonics and reading disabilities in children," Brain and Language **9**, 182–198 (1980).

[Tarr(2010)] E. Tarr, "Formant narrowing using linear predictive coding to improve phonetic perception in children," Master's thesis, The Ohio State University, Columbus, OH, 2010.

[Tarr and Nittrouer(2011)] E. Tarr, S. Nittrouer. "Coherence masking protection for mid-frequency formants by adults and children," J. Acoust. Soc. Amer. **130**, EL290–EL296 (2011).

[Tarr and Nittrouer(2013)] E. Tarr, S. Nittrouer. "Explaining coherence in coherence masking protection for adults and children," J. Acoust. Soc. Amer. **133**, (in press, 2013).

[Tice and Carrell(1997)] B. Tice, and T. Carrell, 1997. TONE: Tone-analog waveform synthesizer [Computer software]. Lincoln, NE: University of Nebraska.

[Turner and Nelson(1982)] C.W. Turner, D.A. Nelson. "Frequency discrimination in regions of normal and impaired sensitivity," J. Speech Hear. Res. **25**, 34–41 (1982).

[Turner and Van Tasell(1984)] C.W. Turner, D. Van Tasell. "Sensorineural hearing loss and the discrimination of vowel-like stimuli," J. Acoust. Soc. Amer. **75**, 562–565 (1984).

[Turner, Relkin, and Doucet(1994)] C.W. Turner, E.M. Relkin, J. Doucet. "Psychopyhsical and Physiological Forward Masking Studies: Probe duration and rise-time effects," J. Acout. Soc. Amer. **96**, 795–800 (1994).

[Turner, Souza, and Forget(1995)] C.W. Turner, P.E. Souza, L.N. Forget. "Use of temporal envelope cues in speech recognition by normal and hearing-impaired listeners," J. Acout. Soc. Amer. **97**, 2568–2576 (1995).

[Walsh and Diehl(1991)] M. Walsh, R. Diehl. "Formant transition duration and amplitude rise time as cues to the stop/glide distinction," The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology **43**, 603-620 (1991).

[Whalen and Liberman(1987)] D. Whalen, A. Liberman. "Speech perception takes precedence over nonspeech perception," Science **237**, 169-171 (1987).

[Wilkinson and Robertson(2006)] G. Wilkinson, G. Robertson, 2006. *The Wide Range Achievement Test (WRAT) (4th ed.),* Lutz, FL: Psychological Assessment Resources.

[Wilson and Dorman(2008a)] B. Wilson, M. Dorman. "Cochlear implants: A remarkable past and a brilliant future," Hear. Res. **242**, 3–21 (2008a).

[Wilson and Dorman(2008b)] B. Wilson, M. Dorman. "Cochlear implants: Current designs and future possibilities," J. Rehab. Res. and Develop. **45**, 695–730 (2008b).

[Xu and Pfingst(2003)] L. Xu, B.E. Pfingst. "Relative importance of temporal envelope and fine structure in lexical-tone perception," J. Acoust. Soc. Amer. **114**, 3024–3027 (2003).

[Ying, Jamieson, Michell(1996)] G.S. Ying, L.H. Jamieson, C.D. Michell. "A probabilistic approach to AMDF pitch detection," Proceedings of the International Conference on Spoken Language Processing (ICSLP, 1996) 1201–1204.

[Zahorian and Jalali Jagharghi (1993)] S. Zahorian and A. Jalali Jagharghi. "Spectral-shape features versus formants as acoustic correlates for vowels," J. Acoust. Soc. Am. **94**, 1966-1982 (1993).

[Zeng, Oba, Garde, Sininger, and Starr(1999)] F.G. Zeng, S. Oba, S. Garde, Y. Sininger, A. Starr. "Temporal and speech processing deficits in auditory neuropathy," Auditory and Vestibular Systems **10**, 3429–3435 (1999).

[Zilca, Kingsbury, Navratil, and Ramaswamy(2006)] R.D. Zilca, B. Kingsbury, J. Navratil, G.N. Ramaswamy. "Pseudo pitch synchronous analysis of speech with applications to speaker recognition," Audio, Speech, and Language Processing, IEEE Transactions on **14**, 467–478 (2006).

[Zwicker(1961)] E. Zwicker. "Subdivision of the audible frequency range into critical bands," J. Acoust. Soc. Am. **33**, 248-248 (1961).

[Zwicker and Terhardt(1980)] E. Zwicker and E. Terhardt. "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency." J. Acoust. Soc. Am. **68**, 1523-1525 (1980).