Temporal Fine Structure and Applications to Cochlear Implants

Xing Li

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Les E. Atlas, Chair

Jay T. Rubinstein

Henrique S. Malvar

Kaibao Nie

Program Authorized to Offer Degree:

Electrical Engineering

University of Washington

Abstract

Temporal Fine Structure and Applications to Cochlear Implants

Xing Li

Chair of the Supervisory Committee:
Professor Les Atlas
Electrical Engineering

Complex broadband sounds are decomposed by the auditory filters into a series of relatively narrowband signals, each of which conveys information about the sound by time-varying features. The slow changes in the overall amplitude constitute envelope, while the more rapid events, such as zero crossings, constitute temporal fine structure (TFS). Although envelope cues from a small number of channels can support robust speech recognition in quiet, TFS seems to plays a significant role for speech perception in noise, especially in fluctuating background. Fundamental questions about the relative importance of envelope and TFS have been addressed by many studies. The definition of TFS poses a critical issue.

Due to the coupling between envelope and phase, it is problematic to isolate the TFS from the envelope for any signal which is not extremely narrowband. Conventionally, a Hilbert transform is used to represent each band as the product of the Hilbert envelope and a frequency-modulated (FM) sinusoidal carrier. The FM component is then taken as the TFS of the band. We show in this dissertation that the Hilbert FM is a distorted representation. To address this concern, we proposed a new distortion-free additive view of signal decomposition, the slow envelope and the fast envelope, using half wave rectification followed by filters reflecting engineering interpretation of neural physiology. The slow envelope is a tool for representing temporal cues that can be coded in the average firing rate of auditory nerve fibers, while the fast envelope instead captures the temporal cues conveyed in neural phase locking patterns. Using this new decomposition and the conventional Hilbert

decomposition, we investigated the relative contribution of neural envelope and TFS coding to speech intelligibility in different noise conditions. The neural representation was generated by a simplified peripheral auditory model (Shamma and Lorenzi, 2013). We observed that the distortions in the Hilbert FM likely confounded the importance of TFS and made it seem insignificant. In contrast, the trends observed with fast envelope were in line with previous perception studies, suggesting that TFS plays a significant role in masking release.

Due to the inherently coarse spectral and temporal resolution in electric hearing, conventional cochlear implant (CI) coding strategies only transmit envelope cues in a small number of channels. The lack of TFS potentially contributes to CI users' difficulties in understanding speech in noise and perceiving music. To encode fine structure information for CI users, we proposed a harmonic-single-sideband-encoder (HSSE) strategy that explicitly tracks the harmonics in complex sounds and transforms them into modulators conveying both envelope and TFS cues. A key distinction about HSSE is that it keeps the envelope and TFS cues together during the transformation to avoid distortions. The effectiveness of HSSE to speech and music perception were tested using three approaches, including acoustic simulation in normal hearing listeners, neural response simulation using a population auditory nerve model (Imennov and Rubinstein, 2009), and acute test in CI patients. Significant effects of HSSE on speech perception in noise and music perception were observed, which illustrated the potentially large benefit of providing fine structure information in a cochlear implant.

TABLE OF CONTENTS

Page

LIST OF FIGURES

ACKNOWLEDGMENTS

I wish to thank my advisors, Les Atlas and Jay Rubinstein, who helped me realize a potential I didn't know I had. Their persistence, creativity, and insight have been invaluable resources and an inspiration. I am also grateful to Kaibao Nie and Adrian KC Lee, whose guidance and expertise have been tremendous. My thanks go also to Bishnu Atal, who taught me what it means to seek the essence of a thing. I appreciate the early and continued encouragement from Rico Malvar, whose advice helped guide this work to its current state. Through their humility and openness, my mentors have shown by example that the highest aspiration is perhaps not to solve but to understand.

Over the course of my graduate years, I have had the opportunity to work with several exceptional individuals. I extend my thanks to Shihab Shamma and Christian Lorenzi for technical discussions; to Ivan Tashev and Jens Ahrens for guidance at Microsoft Research. I would like to thank my colleagues over the years: Jong Ho Won, Nikita Imennov, Ward Drennan, Liz Anderson, and Gary Jones for many helpful discussions; Jim Pitton, Pascal Clark, Brian King, Elliot Saba, Scott Wisdom, Greg Okopal, Bill Kooiman, and Renshu Gu for collaboration and friendship; and all of my lab-mates, past and present: Rahul Vanam, Nicole Nichols, Jaehong Chon, Jessica Tran, David Perlmutter, Dan Tidwell, Suresh Chandrasekan, Patrick McVittie, Adam Greenhall, and Kai Wei for musing on life and everything.

Finally, I thank my family and friends for their love and encouragement, and for making it all worthwhile.

DEDICATION

To my Mom and Dad.

Chapter 1

INTRODUCTION

Speech is probably the primary means of communication between human beings. To tell a certain idea to someone, we operate the vocal tract to make an appropriate speech sound that can be perceived by the person's auditory system. It seems a fairly easy task to process the sound and extract the message conveyed in it. Normal hearing listeners have a remarkable ability to focus on a target talker in the presence of noise or other talkers. Cochlear implant listeners, on the other hand, often experience great difficulty understanding speech in noisy environments. There has been widespread interest in understanding the mechanisms of speech communication. In particular, what acoustic cues of a speech signal contribute to its intelligibility? How are these cues processed and represented in the normal auditory system? How can these cues be appropriately represented for cochlear implant users?

Figure 1.1 shows an example of a speech pressure waveform, produced by a male saying "strawberry jam is sweet". We can see bursts of energy that correspond to phonemes. The temporal characteristics of these bursts carry much information, but their dominant modulation frequency is rather low (typically 4-20 Hz) compared to the temporal capabilities of the peripheral auditory system. There are also fast modulations of several hundred Hz, e.g., in segments of voiced speech where they are perceptually associated with voice pitch. The pressure waveform waxes and wanes in amplitude at much faster rates up to several kHz. These rapid pressure variations constitute the temporal fine structure cues in speech. It has been of interest to carefully asses the importance of and relative roles of different temporal cues in auditory perception. Before proceeding to review the efforts and results from previous studies, it is worthwhile to briefly review the basic structure and function of the peripheral auditory system.

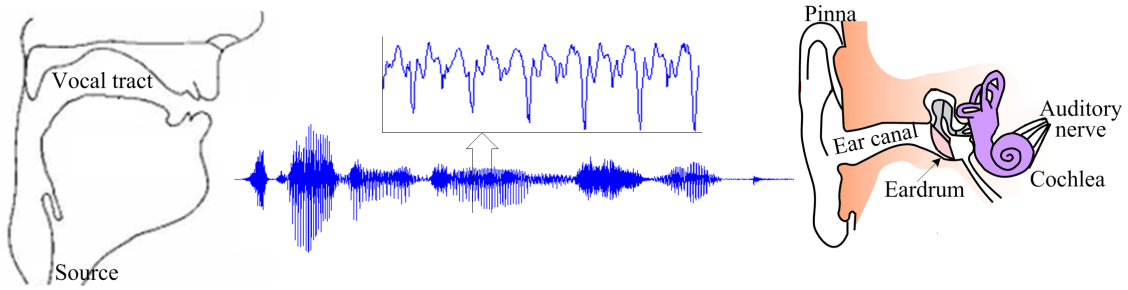

Figure 1.1: From left to right, it shows: cavities of the vocal tract; a speech pressure waveform produced by a male saying "strawberry jam is sweet"; and the anatomy of the human ear. The inset shows the waveform detail over a 50-ms time window.

1.1 Peripheral Auditory System

Figure 1.1 shows the peripheral part of the human auditory system. A sound is first picked up by the pinna. It then travels down the ear canal and causes the eardrum to vibrate. These vibrations are transmitted through the middle ear, by three small bones, to the snail-shaped cochlea of the inner ear. The cochlea is a liquid-filled duct that is divided lengthwise into three chambers [7]. As depicted in Figure 1.2, two of the chambers are separated by the basilar membrane, on which sits the organ of Corti. Sound produces a travelling wave of displacement along the basilar membrane, with different frequencies generating maximal displacement at different places. The movement of the basilar membrane leads to the deflection of the stereocilia (hairs) on top of hair cells that form part of the organ of Corti. As the stereocilia of hair cells bend, the voltage of inner hair cells changes; when the stereocilia are bent sufficiently in one direction (but probably not the other), the voltage changes enough to release neurotransmitter in the synapse between inner hair cells and auditory-nerve fibers. This, in turn, can induce action potentials (nerve spikes) in the auditory-nerve fibers. In this way, the mechanical sound signal is converted into an electrical signal that travels up the auditory nerve to the brain. Inner hair cells are thus the interface between the acoustic world and the perceptual realm.

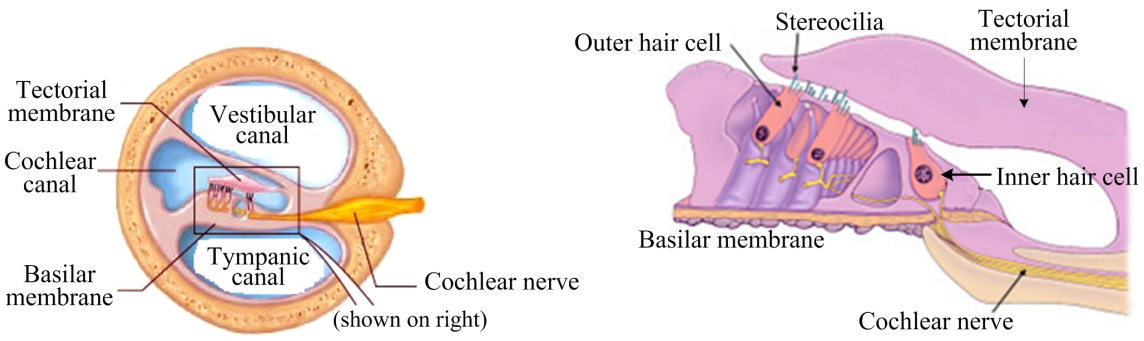

Figure 1.2: The cross section of the cochlea is shown on the left, and the detail of inside the cochlea is shown on the right (www.encyclopedia.com).

1.1.1 Tonotopic Map

The fundamental principle of auditory organization of sound is the tonotopic map. Each place along the basilar membrane responds best to a particular frequency, which is referred to as the characteristic frequency (CF) of the place. High-frequency tones generate maximum displacement near the base, the start of the cochlea; and low-frequency tones produce greatest displacement near the apex, the inner tip of the cochlea. The resulting representation is a tonotopic map of sound frequency along the basilar membrane. The frequency selectivity of the cochlea is often modeled as an array of bandpass filters, called "auditory filters". To describe the bandwidths of auditory filters, Fletcher introduced the concept of critical band in the 1940s [8]. From base to apex, the critical bandwidth decreases in a roughly logarithmic manner, providing an increasingly finer spectral resolution from high to low frequency.

The tonotopic map is maintained at every level of the auditory system through the auditory cortex. Each inner hair cell is most sensitive to a particular frequency, which is approximately the CF of the place it is attached to along the basilar membrane. It follows that each auditory-nerve fiber is also frequency tuned. There are about 30,000 fibers in the cochlea nerve, arranged in a bundle [9]. Those fibers in the center of the bundle originate in the apex of the cochlea and consequently are tuned to low frequencies. Those fibers

near the edge of the bundle originate in the base of the cochlea and are therefore tuned to high frequencies. By virtue of the tonotopic map, the spectral properties of a stimulus are represented in the spatial aspects of the auditory-nerve firing patterns. For instance, the relative response across auditory-nerve fibers, i.e., the magnitude of the response as a function of the CF, encodes the stimulus spectral shape.

Previous studies showed that place cues play a key role in speech perception. For example, the number of frequency channels has a significant effect on speech recognition performance in noise (e.g., [53, 44]). The temporal patterns of neural spikes might also convey important information for speech perception. The role of the detailed patterns of neural spikes is still not well known. In this dissertation, we will propose new engineering signal decomposition approaches for careful study of the temporal cues in complex broadband sounds, such as speech, without the concomitant distortion of current engineering approaches.

1.1.2 Phase Locking

The discharge patterns of single auditory-nerve fibers in response to tonal stimuli have been studied extensively [10, 11, 12, 13]. It is well known that auditory-nerve fibers have the ability to phase lock to low-frequency tones up to several kHz, i.e., neural spikes tend to occur at a particular phase of the cyclical waveform. On each cycle, auditory-nerve fibers may not necessarily fire; but when spikes do occur, they occur at roughly the same phase of the waveform each time. Due to refractoriness, single nerve fibers cannot fire at a rate higher than a few hundred of spikes per second. Also, spike generation is a stochastic process characterized by some jitter. This uncertainty in timing adds to the loss of phase locking above approximately 3-5 kHz (the upper limit is species dependent [14]). Above this frequency, neural spikes cannot run in synchrony with the individual cycles of the stimulus waveform, but they can still phase lock to the modulation envelope of the stimulus [10, 12, 13].

Some of the current knowledge about the coding of envelope and fine structure is based on studies about phase locking to sinusoidally amplitude modulated (AM) tones [1]:

$$s(t) = [1 + m\sin(2\pi f_m t)]\sin(2\pi f_c t) \tag{1.1}$$

A stimulus is specified by three parameters: tone frequency f_c, modulation frequency f_m, and the modulation depth m. One example is shown in Figure 1.3 A. The stimulus is a 1-kHz tone modulated at 100 Hz with a modulation depth of 100%. This signal does not contain energy at f_m, as can be seen from its spectrum in Figure 1.3 B. The modulation in the waveform is due to the interference of the signal components that are separated by a difference frequency f_m:

$$s(t) = \sin(2\pi f_c t) + \frac{m}{2}\left[\sin\left(2\pi(f_c + f_m)t\right) + \sin\left(2\pi(f_c - f_m)t\right)\right] \tag{1.2}$$

Figure 1.3 C shows the average response of an auditory-nerve fiber to the AM stimulus in panel A (stimulus duration, 50 ms). It is satisfied that the tone frequency falls within the tuning curve of the fiber under study. The poststimulus time (PST) histogram is obtained by presenting the stimulus multiple times, recording the evoked spike train each time, counting the firing rate in each time bin (binwidth, 0.1 ms), and averaging the count across repetitions. The fine spacing of peaks at 1-ms intervals indicates phase locking to the 1-kHz fine structure; the grouping into broader peaks spaced by 10 ms indicates phase locking to the 100-Hz envelope.

The Fourier spectrum of the PST histogram is shown in Figure 1.3 D. In contrast to the stimulus spectrum, the response spectrum shows energy at f_m, i.e., the AM signal is demodulated. This demodulation can be attributed to several cochlear nonlinearities with asymmetry between the positive and negative part of the transduction function, the most important being half-wave rectification by inner hair cells. The response spectrum also shows a peak at 0 Hz, which is related to the average firing rate. Manipulations of the AM signal parameters (f_c, f_m, m, and sound level) can result in systematic changes in the phase locking pattern and the average firing rate (see Figure 3 in [1]), indicating that these aspects of neural responses are accessible for sound coding. Next we will describe some of the metrics that have been developed to quantify the different aspects of neural spike trains.

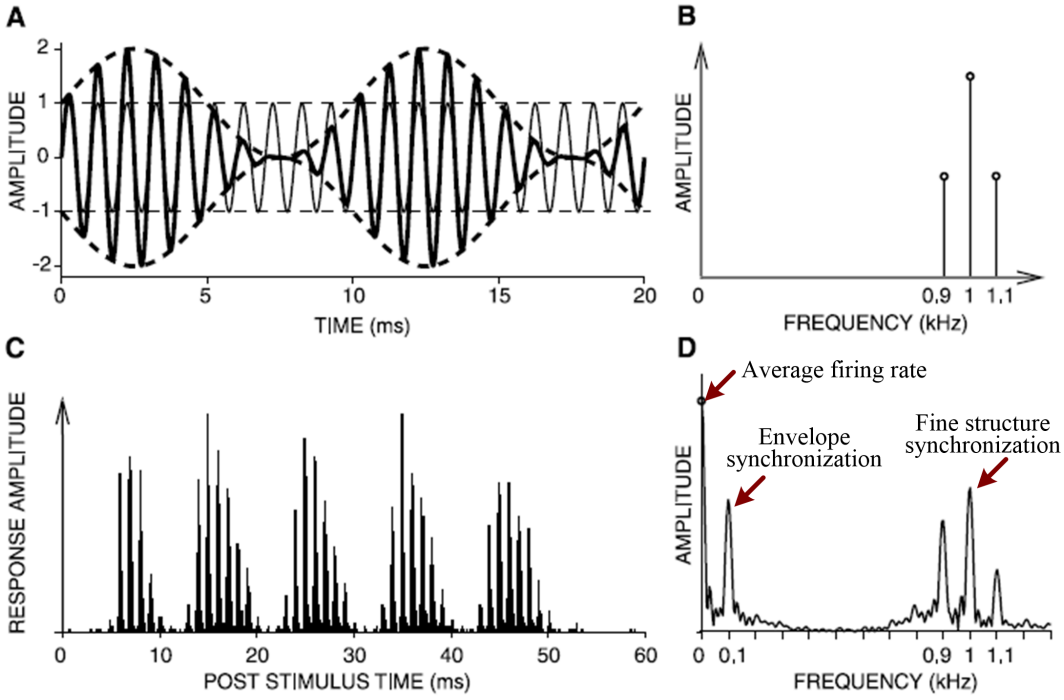

Figure 1.3: (A) superimposed waveforms of a 1 kHz tone (thin line) and the same tone amplitude modulated at 100 Hz (thick line). Dashed lines indicate the envelope. (B) idealized spectrum of the AM tone in panel A. At 100% modulation, the amplitude of the sidebands is half that of the carrier, i.e., a difference of 6 dB. (C) poststimulus time (PST) histogram. (D) spectrum of the PST histogram in C. Figure adopted from [1] with minor modifications.

1.1.3 Neural Response Measures

To quantify phase locking, one popular metric is "vector strength" R, also called synchronization index, which takes values between 0 and 1 [15]. As an illustration, Figure 1.4 A shows the spike trains evoked by a tone stimulus across multiple repetitions. Each spike is treated as a vector of unit length and with phase θ_i, which is measured as the spike time modulo the stimulus period. Arranging all the spikes along the phase axis generates a period histogram, as shown in panel B. The index R is calculated by combining all the spikes through vector addition, and normalizing the resultant vector by the total number of spikes n:

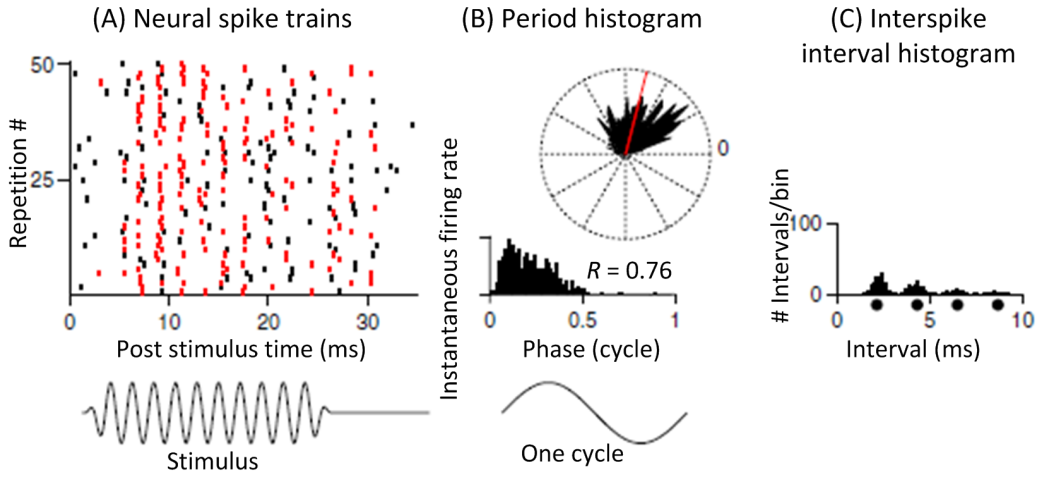

Figure 1.4: (A) rasterplot of neural spike trains evoked by a 460-Hz tone (stimulus duration, 25 ms) over 50 (out of 200) repetitions. Each dot represents one spike. The red dots indicate spikes that are coincident (within a 50-μs bin) across repetitions. The vertical patterning of the dots indicates that spikes occur at a preferred phase of the stimulus. (B) period histogram in which the instantaneous firing rate is measured as the number of spikes occurring at a given phase of the stimulus. The inset shows the same data plotted in polar form, with each line vector representing one bin of the period histogram. (C) interspike interval histogram shows the distribution of intervals between successive spikes within each spike train. The dots below the abscissa indicate integer multiples of the stimulus period. Figure adopted from [2] with a few modifications

$$R = \frac{\sqrt{\left(\sum_{i=1}^{n} \cos\theta_i\right)^2 + \left(\sum_{i=1}^{n} \sin\theta_i\right)^2}}{n} \tag{1.3}$$

It should be noted that the calculation of R requires knowledge of the stimulus period. There is an alternative interpretation of R in frequency domain. By applying the Euler's formula, $\cos\theta_i + j\sin\theta_i = \exp(j\theta_i)$, it becomes apparent that R can also be obtained from the Fourier spectrum of the PST histogram or the period histogram:

$$R = \frac{|\sum_i \exp(j\theta_i)|}{n} = \left|\frac{1}{n}\sum_i \exp(j2\pi\frac{t_i}{T})\right| \tag{1.4}$$

where t_i represents the spike occurence time, $j = \sqrt{-1}$, and T = stimulus period. Specifically, R equals the magnitude value at the frequency bin of interest normalized by the DC

component (i.e., the average firing rate). The metric R describes the degree to which the response is synchronized to a particular frequency at which R is calculated. As a reference, a PST histogram that closely resembles a half-wave rectified AM tone ($m = 1$) gives a synchronization index of 0.5. Higher R values are obtained when the period histograms are more "peaked", i.e., the majority of the spikes occur at roughly the same phase of the stimulus waveform. While most studies have used tonal stimuli with periodic envelopes and have applied the R metric, it is important to keep in mind that most natural sounds are not strictly periodic and that the neural operations by which the central processor extracts envelope information likely differ fundamentally from the analytic ways described above [1].

Alternatively, to bring out the time structure in neural spike trains, it is very common to use interspike interval histograms to reveal the temporal patterns between successive spikes, as depicted in Figure 1.4 C. The all-order interval histograms, also referred to as autocorrelograms, are computed by measuring the intervals not only between successive spikes but for any two spikes within a spike train. Generally the interval histograms are regarded as physiologically plausible mechanisms for pitch perception [16, 17]. However, due to the effect of refractoriness, the interval histograms suffer from a lack of detailed timing information; thus, they are not suitable for evaluating the precision or consistency with which spikes are generated at a particular stimulus phase. To avoid the effect of refractoriness, Joris and colleagues introduced the shuffled correlogram analyses, which are based on counting the spike coincidences across, rather than within, spike trains elicited by a repeated stimulus [18, 2]. These shuffled correlograms can be applied to speech stimuli to examine the neural coding of temporal fine structure information in speech perception [19].

1.1.4 Temporal Modulation Transfer Functions

The ability of auditory-nerve fibers to encode amplitude modulation is normally characterized by temporal modulation transfer functions (TMTFs). The TMTF expresses, as a function of modulation frequency, the ratio of the modulation in the neural responses to the modulation in the acoustic stimulus [20]. Physiologically, TMTFs are often measured as the synchrony to envelopes of AM tones, with f_c and m fixed but f_m altered in a systematic

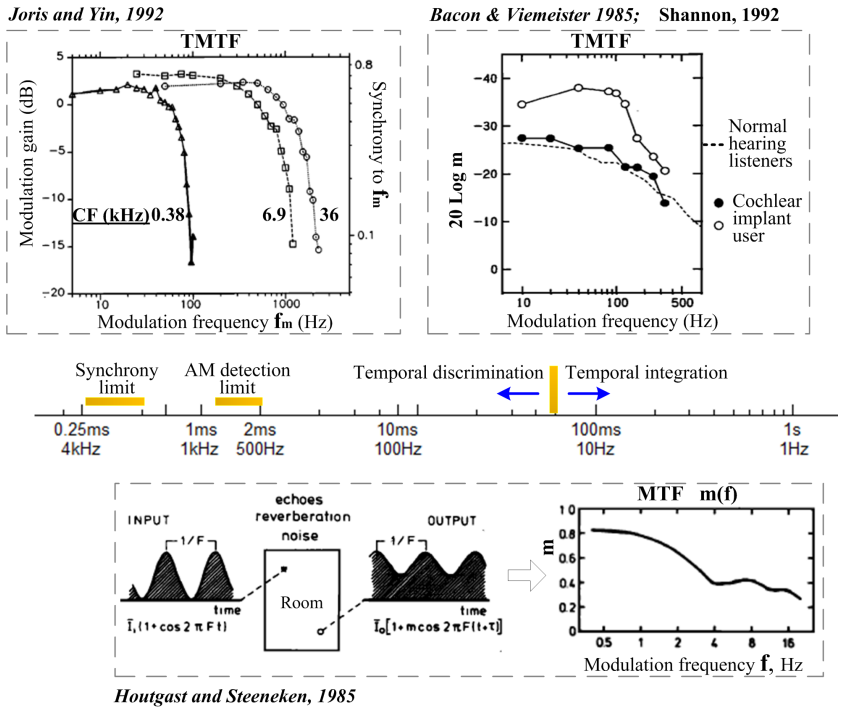

Figure 1.5: Temporal modulation transfer functions (TMTFs) measured in physiology, psychophysics, and room acoustics. In auditory processing, temporal integration is a plausible mechanism to perceive the modulation rates below 16 Hz. The perception of higher frequency modulation depends on the temporal discrimination ability. The AM detection limit affects envelope processing, while the synchrony limit affects temporal fine structure processing. For the upper right panel, the open circles represent TMTF measures from the most basally located electrode pair in one Nucleus cochlear implant user; the solid black circles, from the most apically located electrode pair [24].

manner. Joris and Yin [13] measured the TMTFs of auditory-nerve fibers in cats (Figure 1.5). All the TMTFs have a lowpass shape, presumably because cochlear narrowband filtering limits the range of modulation transmitted by an auditory-nerve fiber. The cutoff frequencies seem to increase with the fiber CF and are typically around a few hundred Hz.

Because modulation typically contains important information, also because highly nonlinear transmission systems often exhibit a quasi-linear response to modulation, the TMTF turns out to be a very useful measure [21]. In Psychophysics, the TMTF provides a "systems analysis" approach to measure the auditory system's temporal discrimination ability,

i.e., temporal resolution, with regard to envelope processing [22]. Bacon and Viemeister [23] measured in normal-hearing listeners the threshold for detecting amplitude modulation as a function of the modulation rate. As shown in Figure 1.5, for modulation rates below 16 Hz, the detection is limited by the amplitude resolution of the ear, rather than by temporal resolution [7]. As the rate increases beyond 16 Hz, temporal resolution starts to have an effect, and the threshold increases gradually. The upper limit for AM detection is typically between 500 and 1000 Hz in normal-hearing listeners; in cochlear implant users, the upper limit is much lower, typically below 300 Hz [24].

In room acoustics, due to reverberation and noise, a transmission path often causes attenuation or smearing effect to the modulation spectrum of speech. Because the modulation spectrum of speech is typically concentrated below 16 Hz and peaked around 4 Hz, Houtgast and Steeneken [25] posited a modulation transfer function (MTF) to describe the "filter characteristic" of a room acting upon the slow envelopes of speech (see Figure 1.5). Based on the MTF, they developed the measure of speech transmission index to predict speech intelligibility in auditoria.

1.2 Spectral and Temporal Cues in Speech

The basic nature of speech has long been thought of as audible sound streams on which the intelligence content is impressed by modulation process [26]. This modulation view directs our attention to the temporal characteristics of speech. One ongoing theme in auditory research is to understand the added perceptual relevance of different temporal cues on various listening tasks [27, 28]. Spectral features, such as formants and their transitions, are known to be important for speech perception. For example, the source filter model, which is at the heart of many speech applications (low-bit-rate speech coding, speech recognition, and speech synthesis), intentionally models the spectral aspects of speech.

1.2.1 Speech-Evoked Neural Response Patterns

Neural firing patterns of large populations of auditory-nerve fibers in response to speech-like stimuli have been investigated extensively in the past [29, 30]. Typically, given the single-unit recordings of large populations of auditory-nerve fibers, the PST histogram for each

unit is first computed. By arranging all the PST histograms in a single figure according to fiber CF, the spatio-temporal patterns (neurograms) can be revealed [3]. An example of a neurogram is depicted in Figure 1.6, where the stimulus is a synthetic consonant-vowel /da/. The fundamental frequency is about 120 Hz. Linear formant transitions occur in the first 50 ms, with the first three formant frequencies $F1$, $F2$, and $F3$ evolving from 0.5, 1.6, and 2.8 kHz to 0.7, 1.2, and 2.4 kHz, respectively. We can make several observations from this neurogram.

First, all the fibers exhibit a response delay, which is about 2 ms in the basal region and increases to 4 ms in the apex as the lower-frequency components travel the length of the cochlea. Second, there is a noticeable drop in the overall firing rate 1-2 ms following the onset of the response because of the adaptation effects. Third, the stronger harmonics, particularly those near the formants, are reflected in the phase-locking patterns at corresponding CFs. For example, as $F2$ decreases from 1.6 to 1.2 kHz, the dominant synchronization pattern at the corresponding region exhibits a gradual shift from the 13^{th} to the 10^{th} harmonic. Similar evidence can be observed around $F1$ region. Forth, phase locking diminishes at higher frequencies, but fibers may still lock to the envelope modulation due to the beating of harmonics within the bandwidth of the fiber

Based on the above observations, one may ask which aspects of the neural responses code speech features that are perceptually significant. Regarding pitch perception, the place and temporal codes elicited by harmonics are potentially important. However, for the complicated task of speech recognition in various acoustic environments, it is not immediately obvious which aspects of the neural spike trains are the most relevant. We refrain from a more substantial discussion of the relationship between physiological mechanisms and perceptual outcomes. Rather, we are concerned with a simpler and more basic question; namely, how can engineering signal representations best be used to study encoding of speech parameters and to what accuracy? We will review the related studies next.

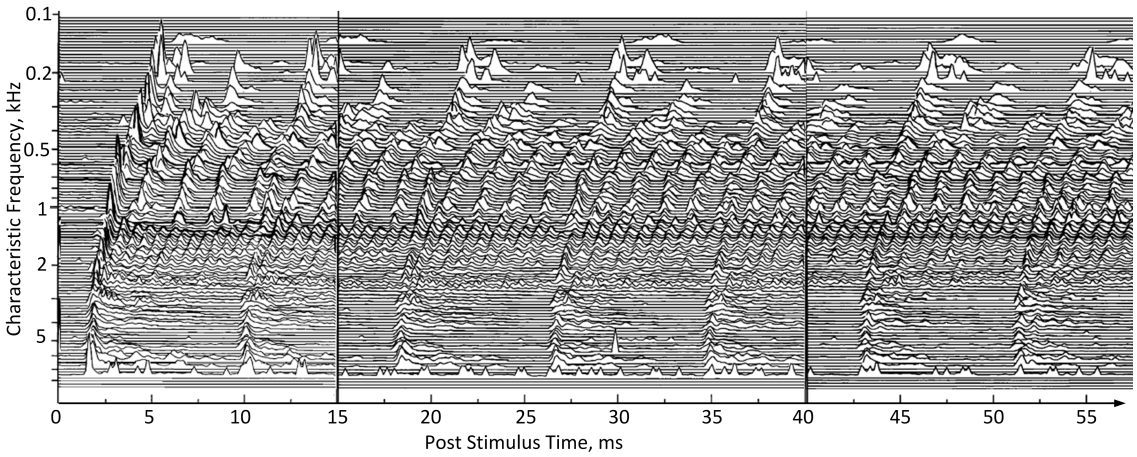

Figure 1.6: Neural firing patterns evoked by a synthetic consonant-vowel /da/. Spike data were recorded from single units of large populations of auditory-nerve fibers (275 units) in the cat over a sufficient number of repetitions (Miller and Sachs, 1983). For each unit, a PST histogram is first constructed (binwidth, 0.1 ms); it is then smoothed by a 3 point triangle window to reduce the noisy character of neural responses (0.25, 0.5, 0.25). Consequently, response components that vary faster than 2.5 kHz become apparently attenuated (> 6 dB); thus, phase locking to $F3$ near the onset of the stimulus appears somewhat smaller than it really is. Figure adopted from [3] with minor modifications.

1.2.2 Neural Coding of Spectral Features

Unlike the AM tones that can be specified by a small number of parameters, speech signals have many different aspects, with temporal and spectral cues being the two collective terminologies. It follows that the subjective perception of speech is also multidimensional, involving pitch, timbre, loudness, localization, and other percepts. The multidimensionality of these percepts must, in general, involve a large number of perceptually relevant parameters. However, for the perception of some voiced speech, e.g., vowel-like sounds, a minimal sufficient set of descriptors have been established [31, 32], including the formats and their transitions, and the low-frequency harmonics which are important for pitch perception [33]. The usefulness of these parameters directs our attention to the spectral aspects of speech.

Sachs and Young [34] showed that over a restricted dynamic range, the stimulus spectrum can be retrieved from the evoked neural responses. Specifically, they measured fibers average

firing rate profile as a function of fiber CF. The measure of average firing rate reflects the slow variations in neural firing; it is related to the DC component of the instantaneous firing rate. Differently, several studies took advantage of the timing information to estimate stimulus spectrum. Young and Sachs [35] reconstructed stimulus spectrum from the evoked neural period histograms using a spectral averaging technique, essentially a bandpass filter centered at the fiber CF. They then picked peaks from this spectrum to estimate vowel formants. Miller and Sachs [36] applied this same technique to neural PST histograms that are generated by synthetic stop-consonant syllables. These processing schemes depend critically on a precise tonotopic map to retrieve stimulus spectrum. Delgutte [37] proposed a different scheme, in which a dominant response component was defined for each fiber. The formants were then identified as the dominant components that occur most frequently in the ensemble. These schemes were all based on the short-time spectra of the firing-rate waveforms to estimate formants. As an alternative to these spectral techniques, Secker-Walker and Searle [38] used interval histograms to examine the temporal structure of neural spike trains and obtained more precise estimates of formants.

These studies have demonstrated that the spectral features of speech can be retrieved from the evoked neural responses. In general, the results compare favorably with direct spectral analyses of the stimuli.

1.2.3 Recent Studies on Temporal Cues

Recently, more understanding has been gained about the perceptual relevance of temporal cues. Shannon et al. [39] showed that under conditions of greatly reduced spectral information, high speech recognition performance can be achieved by presenting temporal envelope cues in as few as four frequency channels. Their signal processing scheme for the 4-channel noise vocoder is shown in Figure 1.7. A speech signal is first split into four channels. Within each channel, the envelope was extracted by half-wave rectification and lowpass filtering. Spectral information was heavily reduced by replacing the frequency-specific information in each channel with band-limited noise. Consequently, only crude spectral cues were available in the resulting synthetic speech. Nevertheless, speech recognition was achieved with these

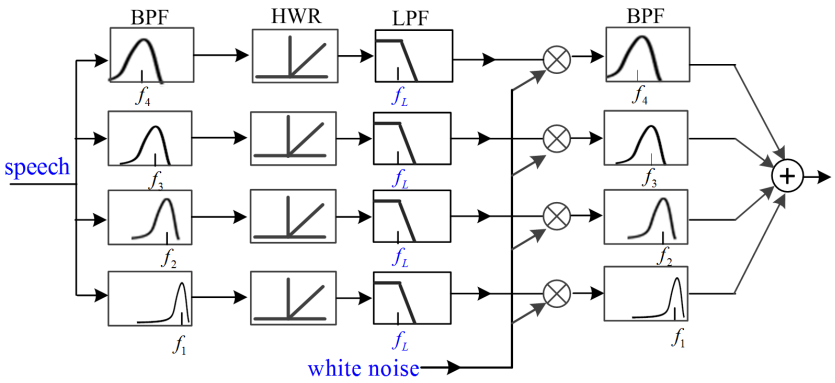

Figure 1.7: Processing scheme of a 4-channel noise vocoder. The crossover frequencies between adjacent bands are 800, 1500, and 2500 Hz, sequentially. The cutoff frequency of the extracted envelope signal can be as low as 50 Hz.

degraded spectral cues and preserved envelope cues.

Because degraded spectral resolution often is a consequence of hearing impairment, the finding that speech recognition can be achieved with primarily temporal cues suggests alternative signal-processing strategies for auditory prostheses. In the late 1980s and early 1990s, a paradigm shift occurred in cochlear implant signal processing, from explicit encoding of spectral features (e.g., [40, 41]) to explicit encoding of temporal envelope cues (e.g., [42, 43]). This paradigm shift does not imply that spectral cues are unimportant for speech perception. Rather, given the coarse spectral resolution in electric hearing, the benefits of temporal cues should be maintained.

Although these past studies indicate that temporal envelope cues seem sufficient for speech recognition in quiet, spectral or, perhaps, other finer (high rate) temporal information could become necessary and important in common adverse listening situations. For example, when background noise or competing talkers are present, the lack of this fine structure information leads to poor speech recognition performance [44, 45]. Numerous perceptual studies have been conducted to address fundamental questions about the relative roles of envelope and fine structure information in auditory perception [46, 47, 48, 49]. The choice of engineering signal decompositions used to manipulate envelope and fine structure information have important consequences for the perceptual outcomes, which we will discuss

in the rest of this dissertation.

1.3 Cochlear Implants

The hair cells in the cochlea are very sensitive. They are susceptible to different kinds of damages, e.g., excessively loud sound and disease. Abnormalities in the hair cells often lead to sensorineural hearing loss, which can be mild, moderate, or total deafness. A cochlear implant (CI) is a hearing device that can bypass absent or damaged hair cells and use electric current to directly stimulate the auditory nerve in a totally deafened person. The implant system consists of a number of components. As depicted in Figure 1.8 A, a battery case (2) supplies power to the speech processor (1), which uses a microphone to pick up sound, converts the sound into a digital signal, processes the signal into stimulation patterns, and encodes them into a radio frequency (RF) signal. The RF signal is then sent by the antenna inside a headpiece (3) to an internal receiver (4) placed under the skin behind the ear. The headpiece is held in place by a magnet attracted to the internal receiver. A hermetically sealed stimulator (5) contains active electronic circuits that can derive power from the RF signal, decode the signal, convert it into electric currents, and send them along wires (6) threaded into the cochlea. The electrodes (7) at the end of the wire stimulate the auditory nerve (8) connected to the central nervous system, where the electric impulses are interpreted as sound [50].

1.3.1 Envelope Based Coding Strategies

The continuous-interleaved-sampling (CIS, [42]) strategy is a representative coding scheme that has been implemented by all major device manufacturers. As shown in Figure 1.8 B, sound is first subject to a number of bandpass filters. Within each channel, the temporal envelope is extracted by either half-wave (shown in the figure) or full-wave rectification followed by a lowpass filter. The envelope is then logarithmically compressed to match the wide range of acoustic amplitudes to the narrow range of usable current levels. Next, the compressed envelope is used to amplitude-modulate a pulse carrier, whose rate is fixed and can vary from several hundred to several thousand Hz. The modulated pulse train is then delivered to the respective electrode. To suppress channel interference, the pulse trains are

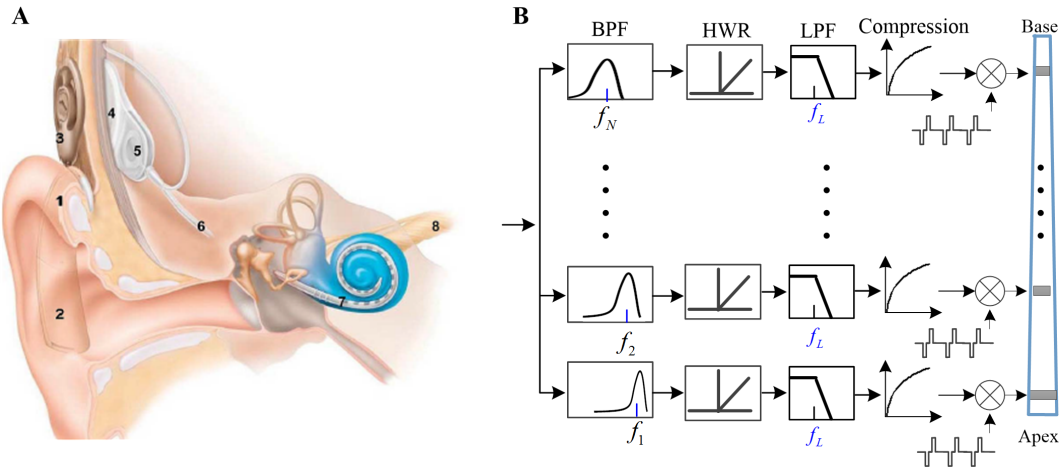

Figure 1.8: (A) a typical modern cochlear implant system (www.cochlear.com). (B) functional block diagram of the continuous-interleaved-sampling strategy. The cutoff frequency of the lowpass filter (f_L) is typically set around 400 Hz or slightly lower.

time-interleaved among all the electrodes, such that no simultaneous stimulation occurs at any time.

A number of strategies differ from CIS in terms of how envelopes are extracted, e.g., by using the Short-Time-Fourier-Transform [51] or the Hilbert transform [52] to extract amplitude modulation. The "n-of-m" strategy family differ from CIS in that only a subset of the electrodes with largest amplitudes, instead of all, are selected for stimulation in a certain analysis window [43, 51]. Despite these variations, CIS is still a representative strategy for envelope based coding schemes. Typically CI users' perception performance can be gauged by a four to eight channel CIS simulation in normal-hearing listeners [53].

1.3.2 Performance Limitations and Fine Structure Coding

To evaluate the effectiveness of CI systems, the sentence recognition task is often chosen because it can best measure the user's ability to communicate in daily life, e.g., a 70% sentence recognition score would support a telephone conversation [54]. In quiet settings, many CI users can recognize speech remarkably well (Figure 3, [50]), which is in line with the finding that envelope cues can support robust speech recognition in quiet [39]. In the

presence of noise or competing talkers, however, most CI users experience great difficulty understanding speech at typical signal-to-noise ratios, e.g., +5 dB [45, 55]. Additionally, CI users perform poorly on lexical tone discrimination, which is an important aspect of tonal language understanding [56]. On music perception, there is also a significant performance gap between CI users and normal hearing listeners [57, 58].

A number of studies have suggested that temporal fine structure information is important to the above mentioned tasks (review by [28]). The lack of fine structure information in CI coding strategies possibly contributes to the perception difficulties experienced by CI users. It becomes a critical issue how to encode fine structure information for CI users. In solving this issue, we shall at least take into account of two factors: the engineering signal processing methods for manipulating envelope and fine structure cues, and CI users' sensitivity to fine structure information. The second factor depends primarily on CI users' spectral and temporal processing abilities. In general, present-day CI systems can only provide a small number of effective channels, e.g., 4-8, although the number of implanted electrodes can be as many as 22 [53]. The TMTFs in CI users exhibit a lowpass shape, with a cutoff frequency below 300 Hz [24]. The significantly reduced processing sensitivity in CI users must be carefully considered in designing fine structure coding strategies for them.

1.4 Organization

This dissertation is aimed to investigate engineering tools for representations of fine structure information in three topical areas. The overall structure is shown in Figure 1.9. Acoustic cues are the basis of auditory perception. The manipulation of acoustic cues has importance consequences for perceptual outcomes. Thus we start by reviewing acoustic signal processing methods. Then we proceed to discuss how fine structure cues can be coded for CI users and evaluate their effects on speech and music perception with CIs. To understand how information is transformed along the auditory pathway, we use others' modeling techniques to examine the intermediate neural representation of relevant cues.

In Chapter 2, we review a number of representative speech analysis techniques. To study the perceptual relevance of envelope and fine structure information, and to improve fine structure coding for CI users, we need signal processing tools to appropriately manipulate

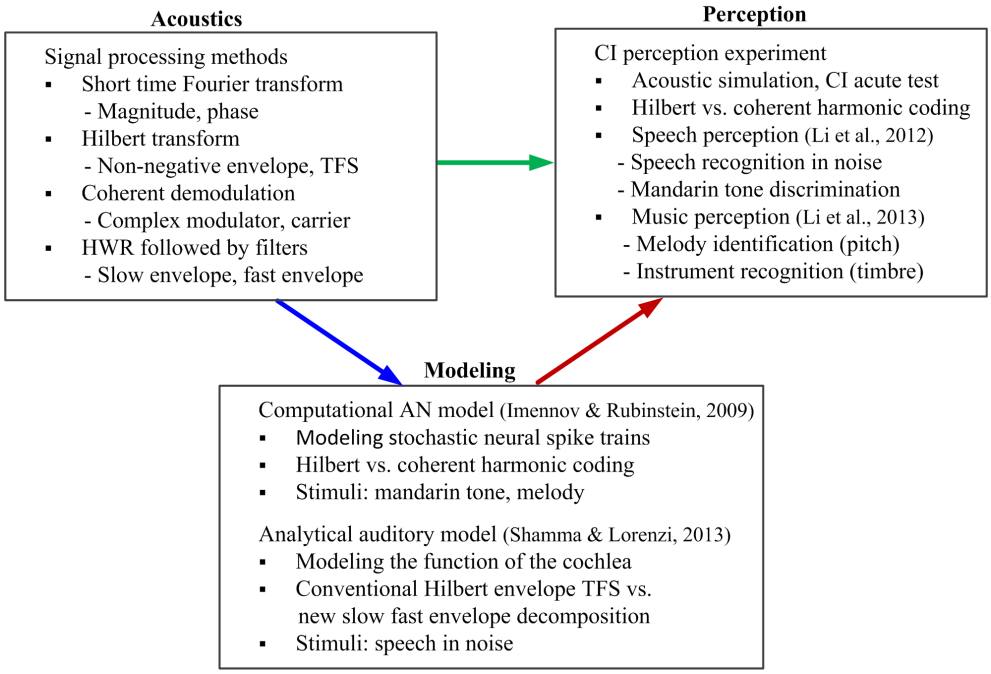

Figure 1.9: Summary of dissertation contents. TFS is short for temporal fine structure. HWR represents half wave rectification. CI stands for cochlear implant. AN represents auditory-nerve.

the envelope and fine structure cues in speech and music. In describing each processing scheme, our discussion will be centered on how envelope and fine structure information are extracted. We also review the related perception studies, and make connections between acoustic processing and perceptual outcomes.

In Chapter 3 and 4, we introduce a harmonic based coding strategy for improving fine structure representation in CIs [59]. The effectiveness of the proposed strategy on speech perception [60] and music perception [61] have been evaluated using three approaches, including acoustic simulation in normal-hearing listeners, computational modeling of the evoked responses [62], and acute test in CI users. We will describe the experimental setup, report the results, and discuss the implications to CI signal processing.

In Chapter 5, we first review the peripheral auditory processing and analyze the information available after the hair cell transduction. We then propose a new engineering

signal decomposition of temporal cues, called the slow envleope and the fast envelope. We will compare this new decomposition with the conventional Hilbert AM and FM vocoder decomposition using the auditory model by Shamma and Lorenzi [6].

Finally, in Chapter 6, we conclude the dissertation with a summary of main points and directions for future work.

Chapter 2

BACKGROUND

In 1939, Dudley [26] introduced an apparatus, called a "vocoder", which operates on the principle of coding a voice and then reconstructing the voice in accordance with this code. The vocoder analyzes a speech signal by first splitting it into multiple channels, and then extracting the amplitude modulation information within each channel. For synthesis, each extracted amplitude signal is used to modulate an appropriate carrier, which can be either a synthetic tone or band-limited noise depending on whether or not the underlying speech segment is voiced. All the modulated carriers are then combined to generate the synthetic speech, which has been shown to successfully convey both intelligibility and pitch information. Nowadays, channel vocoder is still a very useful technique for manipulating sound stimuli in various listening tests [63, 46, 48]. Additionally, many mathematically sophisticated analysis-synthesis techniques have been developed over the past few decades. For example, the phase vocoder [64, 65] is so named to distinguish it from the earlier channel vocoder. In this chapter we will review some of the representative speech analysis techniques, and discuss the definitions of envelope and fine structure information in each processing scheme.

2.1 Short Time Fourier Transform: Magnitude and Phase

The short time Fourier transform (STFT) and the associated spectrogram are widely used techniques to analyze a signal jointly in time and frequency [66]. It consists of a succession of Fourier transforms that are taken over finite-duration analysis windows. Given an analysis window $h(t)$, the STFT of a signal $x(t)$ can be defined as:

$$\begin{aligned} X(t,\omega) &= \int_{-\infty}^{\infty} x(u)h(t-u)\exp\left(-j\omega(u-t)\right)du \\ &= |X(t,\omega)|\exp\left(j\angle X(t,\omega)\right) \end{aligned} \tag{2.1}$$

where $|X(t,\omega)|$ and $\angle X(t,\omega)$ respectively stand for the magnitude and the phase of the complex valued Fourier spectrum. The choice of the window function turns out to be important, because it determines the resolution in time and in frequency. For a rapidly varying signal, it is desirable to have a shorter window to effectively respond to the temporal changes in the signal. On the other hand, windowing operation smears the signal's spectrum. The amount of smearing increases as the window duration gets shorter. The tradeoff between time resolution and frequency resolution is reflected in the comparison between wideband spectrograms and narrowband spectrograms. The discrete STFT is often interpreted as a filterbank consisting of a set of uniformly-spaced bandpass filters:

$$X(t,\omega_n) = x(t) * \left[h(t)\exp\left(j\omega_n t\right)\right], \quad 1{\leq}n{\leq}N \tag{2.2}$$

where ω_n is the center (radian) frequency of the n^{th} passband. All the bandpass filters are based on the same lowpass window function, thus they all have the same bandwidth. Flanagan [67] showed that under normal conditions, $h(t)\exp\left(j\omega_n t\right)$ has a single-sided spectrum and thus can be treated as an analytic filter, which means that the bandpass signal $X(t,\omega_n)$ is also analytic. It becomes apparent that the Hilbert envelope and phase of the bandpass signal are, respectively, related to the magnitude and phase of the STFT at the corresponding frequency bin.

In general, STFT is not a good model of the human auditory system, because we are restricted to a linear frequency scale and only one type of filters. However, STFT has been well studied mathematically. It can help us gain some insight about the relationship between magnitude and phase.

2.1.1 Relationship between Magnitude and Phase

The STFT with a Gaussian window is known to have some special properties [68]. If the analysis window is chosen as a unit-energy Gaussian function with a given time width λ:

$$h(t) = \lambda^{-1/2}\pi^{-1/4}\exp\left(-t^2/(2\lambda^2)\right) \tag{2.3}$$

there is a direct and explicit relationship connecting the magnitude and the phase of the

corresponding STFTs, as shown by Chassande-Mottin et al. [69, 70]:

$$\frac{\partial}{\partial t}\angle X(t,\omega) = \lambda^{-2}\frac{\partial}{\partial \omega}\log|X(t,\omega)| + \omega \tag{2.4}$$

$$\frac{\partial}{\partial \omega}\angle X(t,\omega) = -\lambda^{2}\frac{\partial}{\partial t}\log|X(t,\omega)| - t \tag{2.5}$$

The terms on the left hand side are the partial derivatives of the phase function with respect to time and frequency, respectively. The time derivative of the phase function is commonly known as the *instantaneous frequency*. In the original phase vocoder by Flanagan and Golden [65], the instantaneous frequency was shown to be a suitable representation for manipulating signals in various ways. The frequency derivative of the phase function is known as the *local group delay*. Previous studies showed that the group delay function can be useful to formant frequency estimation [71].

For a Gaussian window, the instantaneous frequency can be obtained by computing the frequency derivative of the log-magnitude of the STFT. From purely mathematical reasoning, it should be possible to recover the phase from the magnitude of the STFT, or vice versa. Since the above equations are only concerned with the changes in magnitude and phase, obtaining an absolute value of the magnitude or phase from these equations requires an integration process to some known point. In actual computer processing, significant recovery errors are often observed, due to numerical limitations and the finite running time of reconstruction algorithms [72]. Nevertheless, in time-frequency regions with strong energies, the error signal tends to be a constant phase shift as expected.

If a different short-time window, rather than a Gaussian function, is chosen, the relationship between magnitude and phase is no longer explicit. However, this does not mean that the coupling between magnitude and phase does not exist. In fact, it becomes manifest when one tries to modify the magnitude or phase spectrum separately. In a number of applications, it is desirable to modify the STFT and then estimate the processed signal from the modified STFT (MSTFT). For example, in speech enhancement by spectral subtraction [73], the magnitude spectrum of the STFT is modified (enhanced), but the phase spectrum is left changed. By combining the modified magnitude spectrum with the original phase

spectrum, a MSTFT is obtained, from which the enhanced speech signal is reconstructed. As another example, a phase vocoder performs time-scaling or pitch-shifting operations to a sound by modifying its magnitude or phase spectrum, and then resynthesizing the sound from the MSTFT [74, 75]. In most applications, including the two cited above, the MSTFT is not valid in the sense that no signal actually corresponds to it. Presumably, this STFT validity issue arises because the coupling between magnitude and phase is disturbed as a consequence of the modification [70].

To reconstruct a signal from its MSTFT, the simple overlap-add method is generally not sufficient. Griffin and Lim [76] proposed a least squared error estimation (LSEE) approach to optimally reconstruct a signal from its MSTFT through iterative procedures. A number of studies have used the STFT based analysis-modification-synthesis technique to investigate the relative importance of magnitude and phase in speech recognition [77, 78, 79, 80, 81]. The coupling between magnitude and phase has important implications for interpreting the results from these studies.

2.1.2 The Relative Importance of Magnitude and Phase in Speech Recognition

The general experimental setup is depicted in Figure 2.1. In speech analysis, the two important parameters are window type (e.g., hamming vs. rectangular) and duration (e.g., 4ms vs. 2s). During modification, magnitude-only stimuli are created by retaining the original magnitude, but replacing each frame's phase spectra with a zero phase or random phase values. In the case of phase-only stimuli, the magnitude spectra of each frame are set to unity or random values, while the phase spectra are retained. For synthesis, the basic overlap-add method or the LSEE approach by Griffin and Lim [76] is employed. The experimental parameters, especially the window function and the reconstruction method, have important consequences for the quality [82] as well as the intelligibility [79, 80] of the synthesized speech.

The effect of window length on the intelligibility of the reconstructed speech has been studied extensively across all studies. Kazamal et al. [81] found that for medium range frames (4–64 ms), speech intelligibility can be perfectly preserved by the magnitude spec-

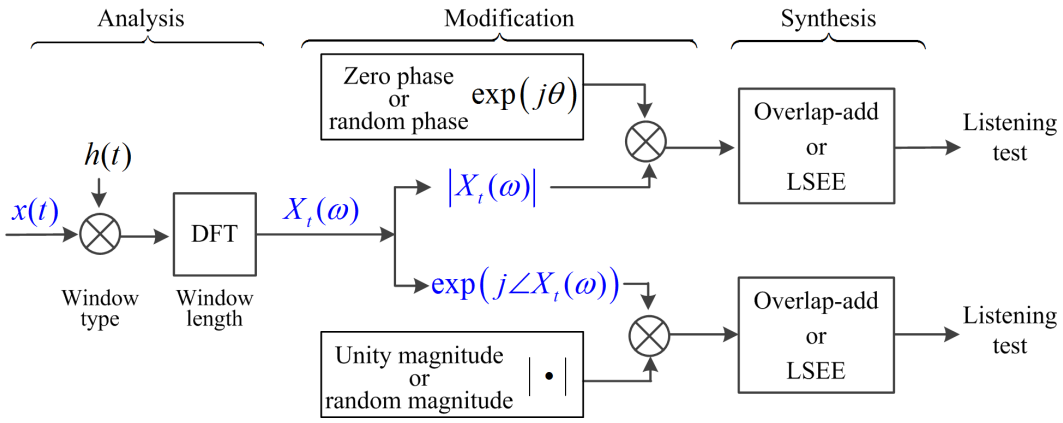

Figure 2.1: The STFT based analysis-modification-synthesis framework for generating magnitude-only and phase-only stimuli.

trum. However, for very long (256–2048 ms) and very short (1/16–1 ms) frames, the magnitude-only stimuli became completely unintelligible. The phase-only stimuli showed the opposite behavior, though less extreme. These results confirmed that for the commonly observed approach in speech processing, i.e., a spectral analysis with a window of a few tens of ms, the use of magnitude spectrum does maintain the essential cues for speech intelligibility. What remains controversial is the relative importance of phase spectrum. It is unclear why better intelligibility is preserved with phase-only stimuli when a longer or shorter frame is used.

For the longer frames, the coarse temporal resolution caused smearing to the temporal characteristics of the magnitude spectrum, consequently, the intelligibility of magnitude-only stimuli degraded significantly. On the other hand, a longer frame means a finer spectral resolution, with which the phase spectra would probably allow a partial recovery of the magnitude, because the temporal dynamic of the magnitude can be conveyed in the very local characteristics of the phase spectrum, such as the group delay (see equation (2.5)). In this respect, the intelligibility of phase-only stimuli might be attributed to magnitude recovery [81]. For the shorter frames, a similar reasoning can be made by replacing the temporal argument with spectral considerations.

It is interesting to note that, where the magnitude spectrum fails in reproducing intelligible speech, the phase spectrum partly takes over this role. Similar observations have been reported by Oppenheim and Lim [77] and Liu et al. [78]. It is unclear whether phase spectrum contributes to speech intelligibility independently, or it is complementary to the magnitude and thus is of secondary importance. This question is ill-posed in the sense that the coupling between magnitude and phase makes it impossible to completely isolate the two. Recently, the perceptual relevance of temporal cues has been widely acknowledged. It is probably more fruitful to directly examine the temporal cues available at the output of auditory filters. To this end, Hilbert transform is often used to mathematically break a signal into envelope (magnitude) and temporal fine structure (phase). Next we will review the Hilbert transform based speech analysis-synthesis technique and the related perception studies.

2.2 Filterbank with Hilbert Transform: Envelope and Temporal Fine Structure

To analyze complex sounds, such as speech and music, the most common method is to split them into multiple channels using a filterbank with bandpass filters, and then find the envelope and temporal fine structure for each channel (Figure 2.2). Compared with the STFT technique, there is much more flexibility in the design of the filterbank. For example, the gammatone filterbank and the octave band filters are both widely used in listening experiments. Generally, from low to high frequency, the bandwidth of each channel increases in a roughly logarithmic fashion. Within each channel, Hilbert transform is used to mathematically break the subband signal $x_n(t)$ into two parts, the envelope and the temporal fine structure (TFS). Specifically, we define a complex-valued signal $s_n(t)$, corresponding to the real-valued, band-limited $x_n(t)$ as:

$$s_n(t) = x_n(t) + j\tilde{x}_n(t) \tag{2.6}$$

where the quadrature signal $\tilde{x}_n(t)$ is the Hilbert transform of $x_n(t)$:

$$\tilde{x}_n(t) = x_n(t) * \left(\frac{1}{\pi t}\right) = \int_{-\infty}^{\infty} x_n(u) \frac{du}{\pi(t-u)} \tag{2.7}$$

It follows that the subband signal $x_n(t)$ can be decomposed into a non-negative time envelope $a_n(t)$ and an instantaneous phase function $\theta_n(t)$. The cosine value of $\theta_n(t)$ is known as the Hilbert TFS.

$$\begin{aligned} s_n(t) &= a_n(t) \exp\left[j\theta_n(t)\right] && (2.8) \\ \text{where } a_n(t) &= |s_n(t)| = \sqrt{x_n^2(t) + \tilde{x}_n^2(t)} && (2.9) \\ \theta_n(t) &= \angle s_n(t) = \tan^{-1}\left[\tilde{x}_n(t)/x_n(t)\right] && (2.10) \\ x_n(t) &= Re\left\{s_n(t)\right\} = a_n(t)\cos\left(\theta_n(t)\right) && (2.11) \end{aligned}$$

It is worth mentioning that although $x_n(t)$ is perfectly band-limited, the extracted envelope $a_n(t)$ and phase function $\theta_n(t)$ are both band-unlimited [67]. To investigate the relative roles of envelope and TFS information in auditory perception, they are often separated using the vocoder analysis-synthesis technique, to generate two types of stimuli, envelope speech versus TFS speech, for listening tests.

As depicted in Figure 2.2, to generate Hilbert envelope speech, the envelope from each channel is first lowpass filtered. The smearing effect of this lowpass operation was studied by Drullman et al. [63]. The filtered envelope then modulates a carrier, which can be either a tone at the respective band center or a band-limited noise. The modulated carriers are then combined across channels to generate a synthetic speech, called the envelope vocoded speech. Similarly, combining all the TFS together generates the Hilbert TFS speech. Lorenzi et al. [48] found that hearing-impaired listeners speech recognition performance in noise was correlated with their ability to perceive Hilbert TFS speech. It was suggested that TFS plays an important role in speech perception, especially in noise. Before diving into previous perception studies, we shall review earlier mathematical analyses of Hilbert envelope and TFS.

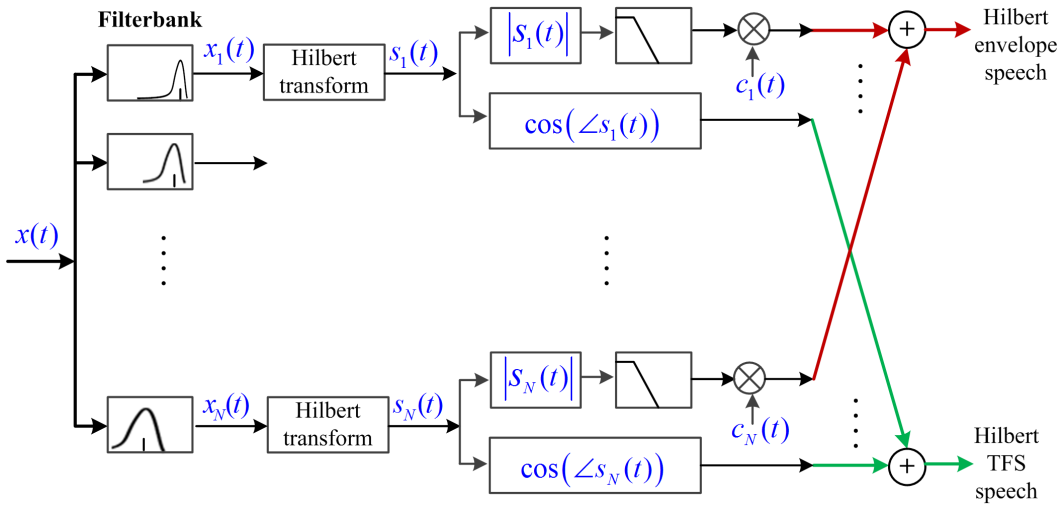

Figure 2.2: Envelope and temporal fine structure decomposition using a filterbank and Hilbert transform. Typically, the spectrum range covers at least 100–6000 Hz. The channel number parameter, N, is usually chosen between 4 and 32. For example, a total of 4–8 channels are used for cochlear implant simulation, while at least 16–32 channels are required for normal auditory simulation. The bandwidth of each channel can then be specified by dividing the whole spectrum range into N channels following the Greenwood map [4]. Thus, the more channels there are, the narrower each channel bandwidth will be.

2.2.1 Relationship between Hilbert Envelope and TFS

The auditory system appears to encode low and medium frequency sound that are critical for speech and music perception using precise timing code of neural spike trains. This prompts a question: what information is conveyed in these timing codes? In signal processing, a related question is raised: is it possible to encode a bandpass signal using purely timing information, such as its zero crossings, or more generally, its crossings of a prespecified constant level or a sinusoid function? In particular, for a bandpass signal, can we recover its envelope from the TFS, or vice versa?

Extensive work on this topic has been done by Voelcker [83], who shows how modulation processes can be viewed as methods of manipulating and extracting the zeros of a signal, and that different systems can be analyzed as to how they affect the zeros of a signal. Voelcker and Requicha [84] showed, through experimental evidence, that certain classes of band-

limited signals are determined uniquely (to within a scale factor) by their zero crossings and/or phase characteristics. Logan [85], following up on their work, discussed in detail the conditions under which a bandpass signal is represented by its zero crossings. Logan's results show that zero crossings can represent bandpass signals only in very special cases. Next let us see two cases, AM signals and narrowband noise, both of which are relevant to speech perception.

Figure 2.3 shows three AM signals. They are all generated using the same model (equation (2.12)) with varying weights for each model component. Specifically, the model consists of the 9th, 10th, and 11th harmonic of a fundamental frequency. The beating between harmonics produces the AM fluctuations that are in synchrony with the fundamental frequency.

$$x(t) = Re\left\{a_1 \exp(j2\pi 9 f_0 t) + a_2 \exp(j2\pi 10 f_0 t) + a_3 \exp(j2\pi 11 f_0 t)\right\} \tag{2.12}$$

Two spectrally asymmetric signals are displayed in the top and bottom rows of Figure 2.3. One is the spectral reversal of the other. In the top row, the weights for each harmonic are respectively set as 1.0, 0.5, and 0.25. The weights are reversed in the bottom row. The middle row shows a spectrally symmetric signal, for which the model can be simplified as follows:

$$\begin{aligned} &\text{Let} \quad a_1 = 0.5, \quad a_2 = 1, \quad a_3 = 0.5 \\ x(t) &= Re\left\{[0.5\exp(-j2\pi f_0 t) + 1 + 0.5\exp(j2\pi f_0 t)]\exp(j2\pi 10 f_0 t)\right\} \\ &= Re\left\{[1 + \cos(2\pi f_0 t)]\exp(j2\pi 10 f_0 t)\right\} \end{aligned} \tag{2.13}$$

The Hilbert envelope and TFS of each signal are overlaid on the respective signal waveform. We can make a number of observations. First, the two signals in the top and bottom rows cannot be differentiated in terms of envelope. Their difference only lies in TFS. In one fundamental period, the top signal exhibits 9 cycles, while the bottom signal exhibits 11 cycles. Apparently the TFS is dominated by the strongest component. Second, in the middle row, the TFS shows discontinuities whenever the envelope approaches zero. If we decrease the value of a_2, the resulting modulator will oscillate between positive and negative

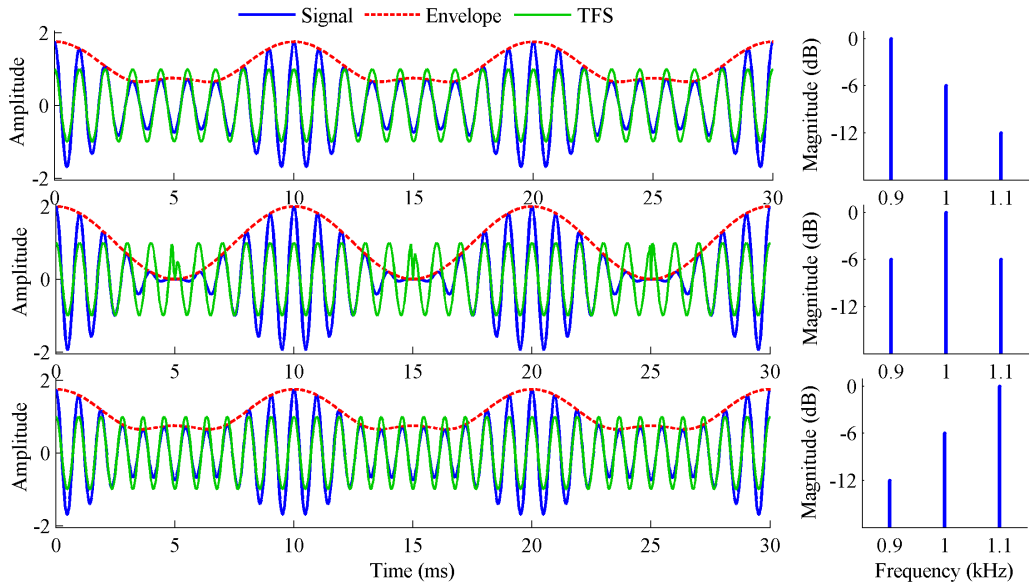

Figure 2.3: Waveforms (left) and spectra (right) of three amplitude modulation (AM) signals. Each AM signal is generated by combining three sinusoidal components with varying weights. These sinusoids are harmonically related. The fundamental frequency is 100 Hz. The Hilbert envelope and TFS of each signal are overlaid on the respective signal waveform.

regularly. Forcing a non-negative envelope causes a jump of in the phase function whenever the modulator approaches zero. Consequently, the TFS appears discontinuous, i.e., it is band-unlimited [67].

The above examples indicate that for some AM signals, it is not possible to uniquely determine the TFS from the envelope. Similarly, for these AM signals, it is not possible to recover the envelope from their zero crossings, for which the mathematical proof can be found in [83] and [85]. In fact, envelope and TFS convey different information about the signal. Thus, they likely play different roles in auditory perception. A key issue is to understand their relative importance to speech recognition in noise, which will be discussed later. It should be noted that Hilbert TFS is a flawed representation of the fine structure information. In auditory perception, the discontinuities in Hilbert TFS cause noticeable distortions or artifacts, which will confound the investigation of the effects of fine structure information. This is a critical issue in auditory perception study. We need to explore better

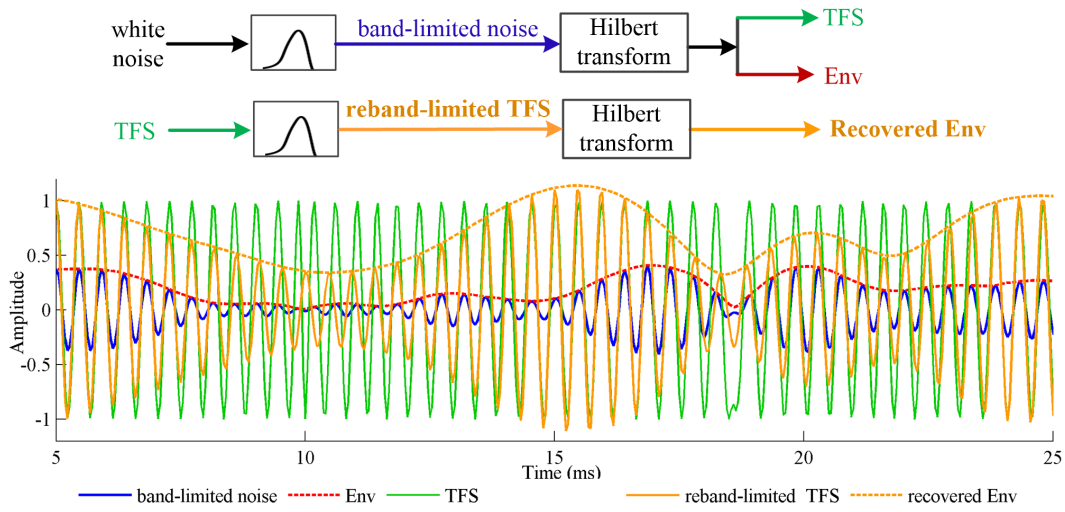

Figure 2.4: Illustration about partial envelope recovery from Hilbert temporal fine structure (TFS). The underlying signal is a band-limited noise, generated by passing whit noise through a bandpass filter (2.0–2.3 kHz). Its envelope and TFS are overlaid on the noise waveform. The recovered envelope is also plotted.

signal processing tools that can facilitate auditory research on fine structure information.

As another example, Figure 2.4 shows a band-limited noise, which is generated by passing white noise through a bandpass filter (2.0 2.3 kHz). The Hilbert envelope and TFS of this bandpass noise are overlaid on the noise waveform. Similar to what we have observed above, the TFS shows discontinuities whenever the envelope approaches zero, e.g., around the 10th ms and the 18th ms. To resume the original bandwidth, we reband-limit the TFS by passing it through the original bandpass filter. The forward-backward filtering is used to achieve zero phase delay. In Figure 2.4, the reband-limited TFS is overlaid on the noise waveform. As we can see, the reband-limited TFS exhibits envelope fluctuations, instead of flat amplitude. Visual observation suggests that the recovered envelope is correlated with the original envelope. This is informally verified by the measure of Pearson's correlation coefficient, which is about 0.63 for this particular noise example.

It is not surprising that envelope cues can be partially recovered from TFS information. Logan [85] has developed a class of bandpass signals which are uniquely specified by their real zero crossings, although he concludes that, in general, "recovering a signal from its

sign changes appears to be very difficult and impractical". For a minimum phase signal, the phase and the log-envelope are Hilbert transform pairs [83], in which case one can fully recover the envelope from the phase function or vice versa. In the above example, partial envelope recovery was achieved by simply reband-limiting the TFS information. One might be able to generate much better recovery results by applying more sophisticated optimization algorithms, such as those described in [72]. Our choice of a simple reband-limiting operation is due to its close connection to cochlear filtering.

The main difference between the previous AM signals and this bandpass noise signal is that the former all have a sparse (concentrated) spectrum, while the latter has a scattered spectrum. Voelcker and Requicha [84] found that the spectral characteristics of a bandpass signal play an important role in determining whether it is possible to recover the signal from its zero crossings. The spectra of speech often exhibit both concentrated and scattered characteristics. Thus, the two cases analyzed above have important implications to speech processing methods used in listening tests.

2.2.2 The Relative Roles of Envelope and TFS in Auditory Perception

A great amount of efforts have been made to investigate the relative roles of envelope and TFS in auditory perception. Using the Hilbert transform based vocoder, Smith et al. [46] created specialized acoustic stimuli called auditory chimaeras, which have the envelope of one sound and the TFS of another. With a large number of vocoder channels ($N \geq 8$), chimaeric speech constructed from two sentences is generally recognized as the sentence that provided envelope, whereas chimaeric music is identified as the melody that contributed TFS. It is therefore concluded that envelope is important to speech recognition and TFS is important to pitch perception. The importance of TFS information to speech recognition is appreciated when background noise is present. Qin and Oxenham [44] found that in the presence of noise or competing talkers, normal-hearing listeners' recognition performance with envelope vocoded speech ($N = 24$) is substantially worse than that with the intact speech. They observed that for intact speech, steady noise produces more masking than a single competing talker at the same SNR condition; whereas for envelope vocoded speech,

the competing-talker scenario turns out to be more difficult. It is indicated that TFS is important for "listening in the dips", where the auditory system takes advantage of the temporal dips in the masker to recognize a target speech [28]. Lorenzi et al. [48] found that listeners with a reduced ability to perceive TFS information tend to perform poorly on speech recognition in a fluctuating background.

Interpretation of perceptual study results are based on the assumptions that Hilbert TFS is a faithful representation of the timing information available to the auditory system, and that envelope and TFS can be isolated. However, the foregoing analyses indicate that these assumptions are invalid. The distortions (discontinuities) associated with Hilbert TFS adversely affect auditory perception, which potentially confounds the results about the relevance of TFS information. However, this issue is largely ignored in many studies. Figure 2.5 shows, side by side, the waveforms of a speech subband signal and its TFS. Their spectrograms are displayed in the bottom. This subband signal looks a lot like the AM example described above. The discontinuities in TFS are very obvious in the zoom-in plot. The spectrogram confirms that Hilbert TFS is band-unlimited. As raised by Moore [28], the fact that subjects need a training session to recognize TFS speech indicates that the extracted TFS cues are distorted. Indeed, by a close examination of the waveforms, one can see that the extracted TFS is different from the fine structure exhibited in the original subband signal.

The interdependence between Hilbert envelope and TFS has been investigated by several studies, most of which are concerned with the issue of envelope recovery form TFS cues. Ghitza [86] reproduced the modulator filtering experiment from Drullman et al. (1994). He observed that if the envelope of a critical-band signal is temporally smoothed while the TFS remains untouched, the resulting synthetic speech signal evokes cochlear envelope cues that are not necessarily smoothed. The narrowband cochlear filtering imposes constraints on the feasibility to isolate sounds TFS from its envelope. Zeng et al. [87] confirmed that with a small number of channels ($N < 4$), envelope cues can be reliably recovered from TFS at the output of auditory filters. Gilbert and Lorenzi [88] demonstrated that when the analysis bandwidth is narrower than four times the bandwidth of a normal auditory filter (i.e., $N \geq 8$), the recovered envelope cues can only play a minor role in consonant

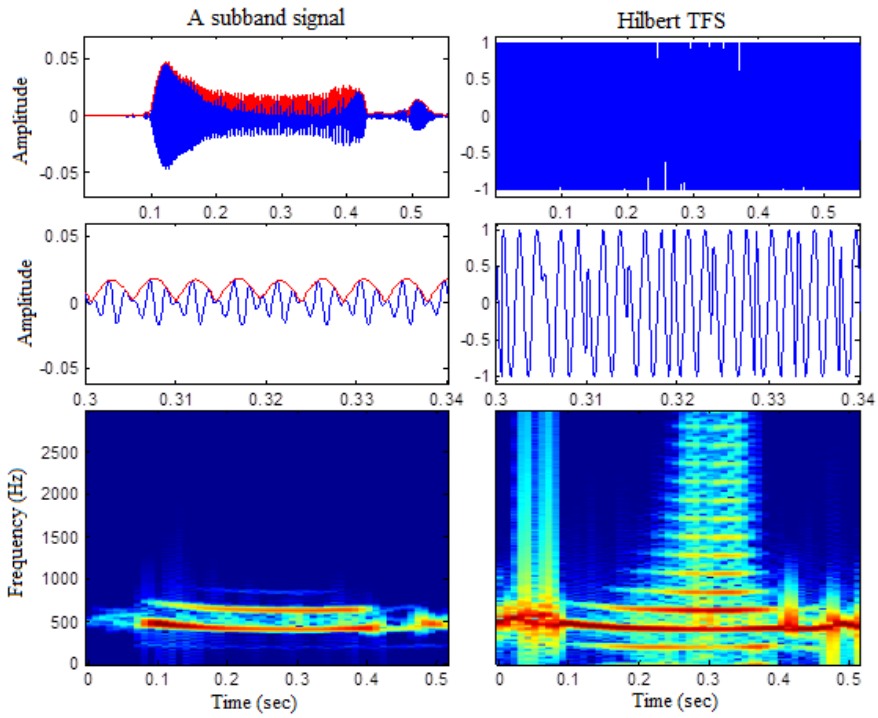

Figure 2.5: illustration about the discontinuities in Hilbert TFS. The left column displays the waveform of a suband signal and its spectrogram. The Hilbert envelope (red) was overlaid on the signal waveform. The right column shows the TFS signal and its spectrogram. The subband signal was extracted from a female utterance saying "head".

identification. They therefore concluded that TFS alone can support speech intelligibility in quiet [88, 89].

Recently, a number of modeling studies reported that neural TFS cues are only marginally significant. The perceptual salience of TFS, as demonstrated in psychoacoustic studies, is possibly due to the recovery of envelope cues [19, 90, 6]. There is apparently a contradiction between the observation that without TFS speech recognition in noise becomes much difficult, and the findings that TFS contributes to speech perception by recovering envelope cues. It should be noted that these studies all use Hilbert transform to extract TFS information, which is a flawed representation. The distortions associated with Hilbert TFS likely confounds their results about the effect of fine structure information. We will return back to this point and continue the discussion in Chapter 5.

2.3 Coherent Demodulation: Complex Modulator and Carrier

Based on the observation that forcing a non-negative envelope causes discontinuities in the instantaneous phase function, Atlas and his students have introduced complex-valued envelope and developed coherent demodulation techniques for more effective modulation filtering of speech and music [91, 92, 5]. In their work, the filterbank based processing is generalized as a sum-of-products model. The key issue is to define a product representation for each subband signal:

$$x(t) = \sum_k x_k(t) = Re\left\{\sum_k m_k(t)c_k(t)\right\} \tag{2.14}$$

where $m_k(t)$ stands for the modulator of the k^{th} subband and $c_k(t)$ is its carrier. It is a fundamentally underdetermined problem to factor a signal into two components without knowing any prior information about the signal. It follows that multiple solutions can be proposed by imposing different constraints.

With the conventional Hilbert decomposition, $m_k(t)$ is specified as a non-negative envelope signal and $c_k(t)$ is essentially the TFS information. This decomposition is problematic in that any modification of the envelope (e.g., modulation filtering) necessarily causes bandwidth expansion in the resynthesized signal. To investigate this issue, Clark and Atlas [5] proposed a modulation filtering projection test. As depicted in Figure 2.6, the modulator is subject to certain modification (e.g., lowpass filtering). It then multiplies the carrier to resynthesize the signal. The test is to see whether the recovered modulator is the same as the intentionally-modified modulator, a question initially raised by Ghitza [86]. The example described in Figure 2.4 can be viewed a specific test case, where the modulator was set as a constant value (i.e., discarded), but the recovered modulator was not flat at all. Through formal operator algebra, Clark and Atlas [5] proved why Hilbert decomposition necessarily fails the modulation filtering projection test. On the other hand, they proved that coherent modulation filtering is a more effective approach.

The concept of coherent demodulation was adopted form communication literature. With regard to speech analysis, the modulator represents the intelligibility related modulation information, while the carrier corresponds to the pitch related acoustic frequency.

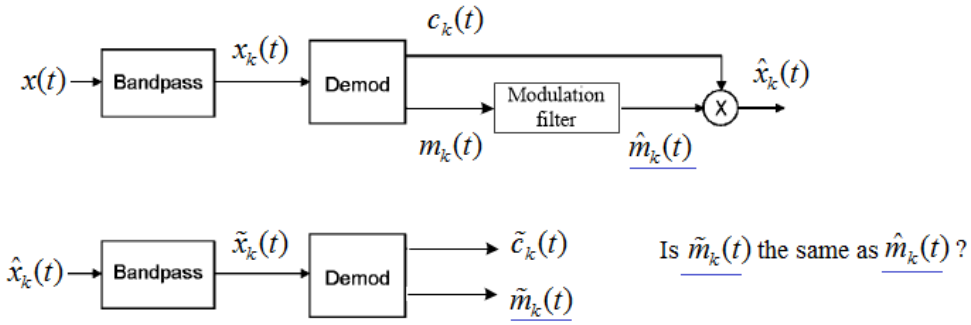

Figure 2.6: system block diagram for the modulation filtering projection test. The "Demod" block can be implemented as Hilbert envelope and TFS decomposition, or coherent demodulation. Figure adopted from [5] with minor modifications.

It should be kept in mind that for an arbitrary sound, there is not necessarily a definite pitch frequency. Consequently, coherent demodulation is often subject to adjustment for different sound stimuli

Given a bandpass signal, it is coherently demodulated as follows (Figure 2.7). First, the signal is converted to an analytic signal using Hilbert transform. Meanwhile, the signal is analyzed to detect the carrier frequency, which is often empirically defined, e.g., as the spectral center-of-gravity of the subband signal [5]. Since speech is non-stationary and time-varying, it follows that the carrier frequency is also a function of time, denoted as $f_k(t)$. Integrating the carrier frequency over time produces a phase function, which is used to specify the carrier:

$$c_k(t) = \exp\left(j\phi(t)\right) = \exp\left(j\int_0^t 2\pi f(\tau)d\tau\right) \tag{2.15}$$

Next, to generate the modulator, the carrier is factored out of the bandpass signal using a conjugate multiplication operation, which corresponds to a frequency shift operation in spectral domain. The resulting modulator is often a lowpass complex-valued signal. Mathematically, this is not surprising because the Fourier spectrum of the modulator is often not symmetric about the frequency origin.

In terms of carrier frequency detection, there are a number of considerations worth

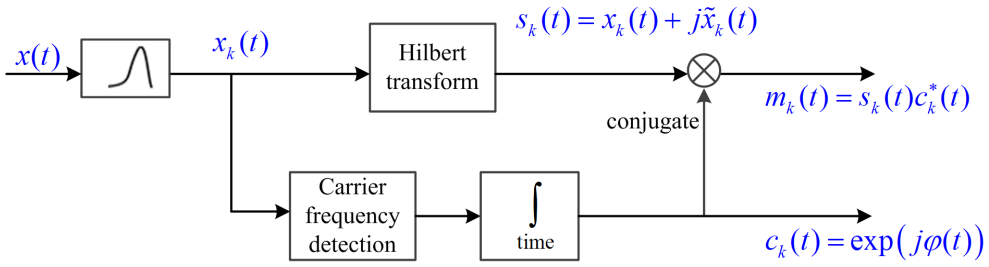

Figure 2.7: functional block diagram of coherent demodulation.

mentioning. First, because harmonic structure is an important feature of complex sounds, it is often appealing to adopt the harmonic frequency as the carrier frequency [59]. We will discuss more about this idea and describe the related experiments in Chapter 3 and 4. Second, because speech is time-varying, so is the carrier frequency, it is practically necessary to segment speech into short-time blocks for carrier frequency detection. Time smoothing is needed in block based speech processing. In this dissertation, we follow the typical engineering practice on choosing frame size and other related parameters.

We can make several comparisons between Hilbert decomposition and coherent demodulation. In both processing schemes, the real-valued bandpass signal is converted into a complex-valued analytic signal. With Hilbert decomposition, the modulator is specified as the non-negative magnitude function, while the carrier is defined using the instantaneous phase function. In contrast, with coherent demodulation, the carrier function needs to be specified first. The modulator is then generated by factoring the carrier out of the subband signal. It should be noted that the key difference lies in how the carrier is defined. If the instantaneous frequency, calculated as the time derivative of the instantaneous phase function, is adopted as the carrier frequency, then coherent demodulation becomes equivalent to Hilbert decomposition. However, provided that the carrier frequency is carefully defined, coherent demodulation can be considered as linear processing, whereas Hilbert decomposition is not. This difference has important consequences for auditory perception, which we will discuss in Chapter 3 and 4.

2.4 Linear Prediction: Minimal Phase Filter and Prediction Residual

Linear predictive coding (LPC) is one of the most useful speech analysis techniques, first developed by Atal and his colleagues [93, 94, 95] It exploits the redundancy in signal sequences by finding the optimal linear combination of a fixed-length history to predict the current sample. The prediction model describes speech in a few physically meaningful parameters. The coder specifies model parameters using an analysis-by-synthesis procedure with a perceptual distortion measure. Efficient coding is obtained because the entropy of speech signal is maximized after the correlation between successive samples is removed by predictive modeling [96].

As illustrated in Figure 2.8, a signal is analyzed into two parts by predictive modeling, a prediction filter and an error signal (prediction residual). The prediction filter is typically an all-pole filter, which must be minimum-phase to minimize the prediction error. The magnitude response of the prediction filter characterizes the spectral envelope of the signal, while the spectral fine structure information is conveyed in the prediction residual. The duality between time and frequency implies that predictive modeling can be applied equally well to sequences of Fourier transform coefficients. In this case the magnitude response of the prediction filter describes the time-domain envelope, and the prediction residual is related to the temporal fine structure information. To estimate the temporal cues in a particular frequency channel, FDLP shall be applied only to the corresponding frequency bins (not indicated in the figure).

The idea of FDLP was first applied in audio coding by Herre and Johnston [97], who dubbed it temporal noise shaping (TNS). Their goal was to eliminate pre-echo artifacts associated with transients in perceptual audio coders by factoring out the parameterized time envelope prior to quantization, then reintroducing it during reconstruction [98]. Kumaresan [99, 100] used linear prediction in the spectral domain to fit the Hilbert envelope without calculating the corresponding analytic signal. He also showed that the instantaneous frequency estimation can be improved. His results suggest that the time-domain product representation (see equation (2.14)) can be better solved using linear prediction in the spectral domain. Athineos and Ellis [101] presented a linear-algebra derivation about

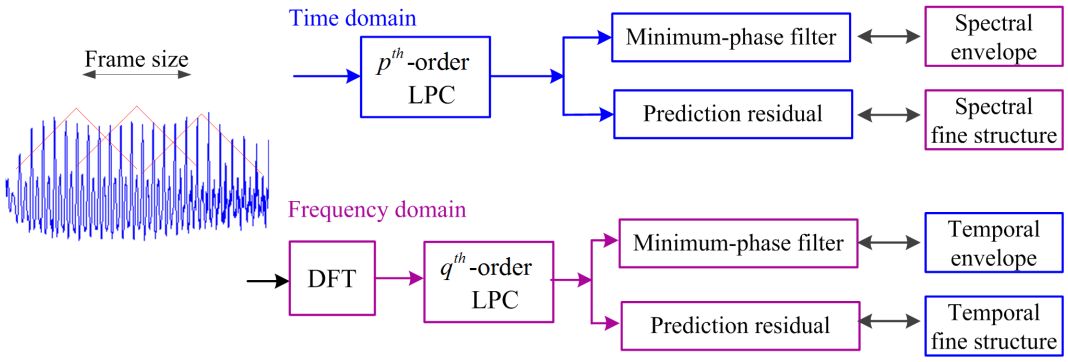

Figure 2.8: illustration of the duality between time-domain and frequency-domain linear prediction.

auto-regressive (AR) modeling of temporal envelopes. They formally proved that the temporal envelope estimated by FDLP is equivalent to the squared Hilbert envelope. Their work was later extended by Hsu and Liu [102], who systematically analyzed the AR modeling of temporal envelopes using different types of finite-length discrete trigonometric transforms and discussed their usefulness in audio coding.

Just as conventional AR models are used most effectively on signals with sharp spectral peaks that can be modeled with complex pole pairs, FDLP models are most appropriate for "peaky" temporal envelopes. Individual poles of the model may be directly associated with specific temporal maxima in the signal waveform. In general, the order of FDLP models depends on the anticipated density of temporal peaks within an analysis window. A higher order model typically results in smaller estimation error, but increased computing complexity and delay, which is a concern for real-time applications. One major disadvantage with AR modeling of temporal envelopes is that the valleys (zeros) of envelopes are usually not well fitted. However, in auditory perception, the temporal valleys are likely as important as the peaks, e.g., the perception of stop consonants rely on both. Due to these concerns, the FDLP technique is not explored in this dissertation.

2.5 Summary

We have reviewed a number of representative speech analysis techniques in this Chapter. Recently, there are some new perspectives on speech decomposition. Sell and Slaney [103, 104] formulated demodulation as a convex optimization problem, in which the envelope is constrained to be non-negative and band-limited. Turner and Sahani [105] used probabilistic inference to solve jointly for stochastic modulators and carriers. These techniques are not explored in this dissertation because they will add variability to the perception experiments.

Between STFT and Hilbert transform, they are similar in the sense that both techniques force the modulators to be non-negative and thus cause phase discontinuity issues. In both processing scheme, the interdependence between envelope (magnitude) and phase complicates the modification of signal. In contrast, coherent demodulation typically can be regarded as linear processing. Thus it is preferred in a number of applications, e.g., modulation filtering and fine structure coding for cochlear implants, which we will discuss next.

Chapter 3

IMPROVED PERCEPTION OF SPEECH IN NOISE AND MANDARIN TONES WITH ACOUSTIC SIMULATIONS OF HARMONIC CODING FOR COCHLEAR IMPLANTS

In Chapter 2, we mentioned that due to the coupling between magnitude and phase, it is an ill-posed problem to separate the two components. The general indication is that, to code temporal fine structure (TFS) information for cochlear implant (CI) users, envelope and TFS should be processed as a whole, rather than separately. In coherent demodulation, the extracted modulators are complex signals because they contain both magnitude and phase information. This distinction about coherent demodulation makes it suitable for coding TFS information for CI users. In this chapter, we will explore how coherent demodulation can be customized to code TFS cues for CI users [1].

3.1 Introduction

Despite the great advances in efficacy of CIs during the past two decades, most CI users still have trouble understanding speech in the presence of noise or multiple talkers (e.g., [45]). Additionally, the fundamental frequency (F_0) cues transmitted by contemporary CIs are unlikely to be sufficient for tonal pattern identification, an important aspect of speech in tonal languages (e.g., [56]). A number of studies have suggested that TFS information is important to the above tasks (for review, see [28]). However, fine structure information is largely discarded in conventional CI signal processing strategies, e.g., the continuous-interleaved-sampling (CIS) strategy [42]. This indicates a need for a new approach to deliver TFS information to CI users.

The importance of TFS has been demonstrated in a variety of listening tasks. Pitch perception [106], lexical tone identification [107], and speech recognition in noise [48, 49] all

[1] The work described in this chapter was published as Li, Nie, Imennov, Won, Drennan, Rubinstein, and Atlas (2012). J Acoust Soc Am, 132(5):3387-98

appear to be linked to TFS perception. The lack of TFS in CI encoding reduces temporal pitch cues and might partly account for CI users' speech perception difficulties in adverse situations. To deliver TFS by pulsatile stimulation, an important caveat to note is the reduced sensitivity to temporal modulation in electric hearing. Most CI users cannot discriminate changes in the repetition rate of the electric waveform above approximately 300 Hz [24, 108], whereas TFS typically oscillates at a much higher rate.

Several approaches have been attempted to encode fine structure information for CI users. The HiRes strategy uses a relatively high envelope cutoff frequency and pulse rate to improve TFS representation. The HiRes Fidelity 120 uses a current steering paradigm to represent spectral fine structure and shows an improvement in frequency resolution over the HiRes strategy, although it is unclear if this provides significant benefits to speech or music perception [109, 110]. The Fine Structure Processing (FSP) strategy bases the pulse triggering pattern in a particular channel on the zero crossings of the respective band waveform to encode TFS; its effectiveness for enhancing speech or music perception is yet to be demonstrated [111, 112]. Nie et al. [113] proposed to convert TFS into a frequency modulation signal and use it to frequency modulate the pulse rate; they showed by vocoder simulations that the frequency modulation information was beneficial to speech recognition in noise. Laneau et al. [114] suggested modulating the channel envelope at the input signal's F_0 with 100% modulation depth; this "F0mod" method showed improvement over the advanced-combination-encoder (ACE) strategy on music perception and Mandarin tone perception but no advantage for sentence recognition [115]. Finally, analog strategies can be used to transmit TFS, but these strategies suffer from increased electrode interaction, which ultimately reduces listeners' perceptual abilities.

In addition to the lack of TFS, contemporary CIs provide little or no representation of harmonics of complex sounds, such as those produced by human voice and musical instruments; this might also contribute to CI users' difficulties in perceiving music and recognizing speech in noise [116, 106]. The lower harmonics of complex sounds can be resolved or separated on the basilar membrane and elicit a more salient and accurate pitch percept than the unresolved harmonics (e.g., [117]. In normal-hearing listeners, pitch cues have been found to be important for separating mixed speech signals (Summers and Leek, 1998). In

short-electrode CI users (hybrid listeners), pitch cues obtained from low-frequency acoustic stimulation have been shown to benefit speech recognition in a competing-talker background [118]. In aggregate, the above studies suggest that providing low-frequency harmonics might be beneficial to CI users.

To encode TFS and harmonic information for CI users, we have proposed a harmonic-single-sideband-encoder strategy (HSSE; [119, 59]) that explicitly tracks the harmonics of complex sounds and linearly transforms harmonics into modulators conveying both amplitude modulation (AM) and TFS cues to each electrode. During unvoiced segments of speech, the fast-oscillating TFS is converted into a slowly-varying yet still noise-like signal and then preserved in HSSE modulators. The main distinction of HSSE from CIS-like strategies is that the former uses F_0-synchronized frequency downshift operations to extract the modulators for electrodes, whereas the latter use nonlinear incoherent approaches to extract temporal cues, e.g., the Hilbert envelope or half/full wave rectification followed by a lowpass filter. Frequency downshift is a linear operation and introduces no distortion to the resulting modulators [92, 5], whereas incoherent approaches incur nonlinear distortions to the extracted temporal cues.

A mathematical analysis of the nonlinear distortions caused by incoherent processing can be found in Flanagan [67]. As a simple illustration, Figure 3.1 shows two tones at 800 and 1000 Hz, respectively. Combining the two tones together results in a bandpass signal. By swapping the amplitudes of the two tones, a different bandpass signal can be constructed. The spectrum and the waveform of each signal are displayed in the 1st and 2nd column, respectively. Despite the evident difference between the two signals, they are mapped to identical temporal cues by taking the Hilbert envelope. The spectrum of each signal's envelope is displayed in the 3rd column. Comparing a signal's spectrum with its envelope spectrum, one can hardly tell how a particular frequency component in the envelope is related to any specific tone in the signal, i.e., the envelope contains distortion components relative to the original signal. In contrast, if both bandpass signals are transformed by a frequency downshift operation like in HSSE, their uniqueness would be retained because the transformed signal contains the original tone components but no distortion component. More details on HSSE are provided in Section 3.2.

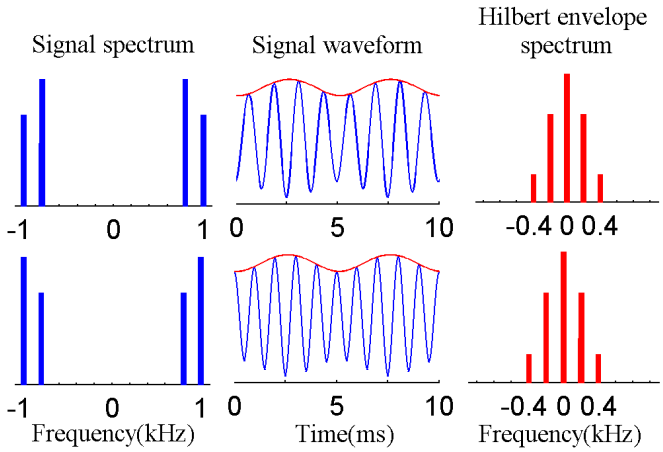

Figure 3.1: Illustration of the distortions introduced by incoherent operation. Note that the Hilbert envelope of a signal does not represent any specific tone component in the original signal, rather contains frequency components that are not present in the original signal (i.e., distortions).

The aim of this study was to investigate the potential benefits of HSSE for speech perception. In experiment 1, vocoder simulations of HSSE and CIS were implemented, respectively; their effects on sentence recognition in noise were compared using normal-hearing listeners. CIS was chosen as a comparison baseline because CIS is the basis of the encoding approaches used in virtually every modern CI speech processor. Nearly all recent strategies resemble CIS in terms of temporal envelope extraction. Although these strategies might be able to deliver more temporal or spectral fine structure cues, they seem to produce similar speech and music recognition performance as CIS [110, 111, 112]. Typically CI users' perception performance can be gauged by a 4 to 8 bands CIS simulation [53]. To prevent vocoders from presenting information that might not be accessible to CI users, both CIS envelopes and HSSE modulators were lowpass filtered at 300 Hz to correspond to the pitch saturation limit in electric hearing. The hypothesis was that under the same spectral and temporal constraints, listeners would recognize speech better with HSSE than with CIS vocoders, due to the advantage of HSSE in extracting non-distorted harmonic and TFS information.

Using the same vocoder simulations as in experiment 1, the effect of HSSE processing

on Mandarin tone identification was compared with that of CIS encoding in experiment 2. Due to the important differences between acoustic and electric hearing, the neural discharge patterns evoked by HSSE- and CIS-encoded Mandarin tone stimuli were also examined, using a population model of electrically-stimulated auditory nerve fibers [62]. This population model was chosen because the model's single-fiber response properties–e.g., spike latency, jitter, and relative refractory period–have been shown to closely correspond to the response quantities measured *in vivo* (e.g., [120]). Moreover, the normalized response thresholds of a population of diameter-distributed model fibers have been shown to match that of the same number of *in vivo* fibers [62], suggesting that the model may be used to approximate the aggregate responses of the auditory nerve. Our reasoning was that if the advantage of HSSE can be observed in both vocoder and neural response simulations, then it is likely to be beneficial to CI users.

3.2 Experiment 1:speech recognition in noise with simulated HSSE and CIS strategies

3.2.1 HSSE processing

To encode harmonics for CI users, the F_0 of a given speech signal was first estimated such that F_0's harmonics can be analyzed. Given the frequencies of harmonics and the frequency spacing of a vocoder, each harmonic was matched to a particular channel based on spectral correspondence. Next, frequency downshift operations were used to transform harmonics into HSSE modulators for their respective channel. Finally, each HSSE modulator was lowpass filtered at 300 Hz to limit the temporal information accessible to vocoder listeners.

F_0 *estimation*

To track F_0, an incoming signal was first segmented into short frames such that constant F_0 can be assumed within each frame. The frame size was empirically chosen as 20 msec to handle the possible F_0 range of human voices (> 50 Hz). There was a 10-msec overlap between contiguous frames. To further refine F_0 estimates, each frame was combined with its preceding and succeeding frames to produce a smooth F_0 trajectory during each 40-msec

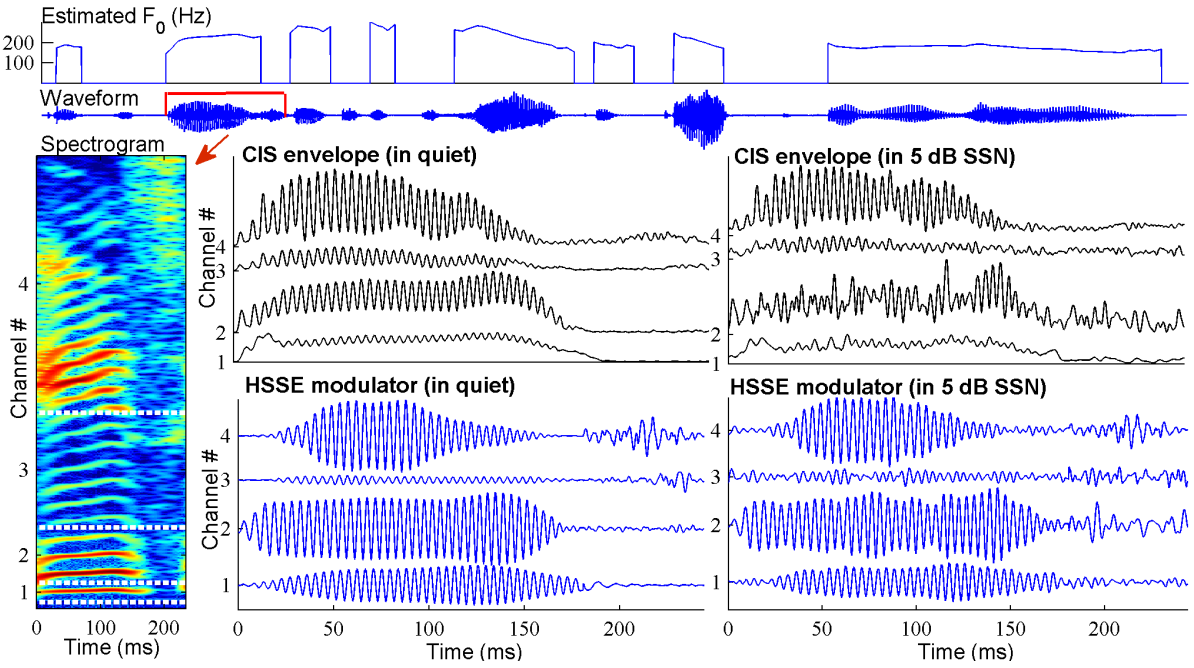

Figure 3.2: Visual comparison between the 300-Hz-wide CIS envelopes and HSSE modulators extracted from the "eas" segment of "easy" in quiet and in noise under 4-channel condition. The spectrogram of the "eas" segment is displayed on the left. SSN is short for speech-shaped steady state noise. The waveform in the 2nd row represents an IEEE sentence saying "it's easy to tell the depth of a well", for which the estimated F_0 contour is displayed in the 1st row. The speakers pitch is about 230 Hz.

window. For simplicity, the following description is focused on the processing of a single frame.

A least-squares-harmonic model was used to track F_0 in this study [121]. Given a signal frame, this method first detected its underlying harmonic structure in the frequency domain and then derived the F_0 accordingly. Pilot studies showed that this technique can reliably track F_0 even in adverse situations, e.g., at a signal to noise ratio of -10 dB. During unvoiced segment, an F_0 value was assigned by linearly interpolating the F_0 estimates from surrounding voiced frames. For example, in the top row of Figure 3.2, the gaps in the estimated F_0 contour (related to unvoiced speech) would be linearly interpolated to generate a continuous F_0 trajectory. In this way, both voiced and unvoiced speech can be processed in the same harmonic-centered fashion, yet the concept of harmonic is only meaningful for voiced speech. It will be shown later that although an interpolated F_0 was assumed for unvoiced speech, the encoded TFS would be noise-like.

Harmonic analysis

Given the detected F_0, a particular harmonic $h_k(t)$, with t being the time index and k the harmonic index, can be modeled as the following sinusoid:

$$h_k(t) = a_k(t)\cos(2\pi F_0 t + \varphi_k(t)) \tag{3.1}$$

where $h_k(t)$ represents the harmonic amplitude, kF_0 is the harmonic frequency, and $\varphi_k(t)$ represents the phase information.

The amplitude information is related to the envelope, while the harmonic frequency and phase information are related to the TFS. If $h_k(t)$ is an actual harmonic extracted from voiced speech and fluctuates nearly periodically, then $\varphi_k(t)$ stays approximately constant to let the overall TFS oscillate regularly at a rate of kF_0. In contrast, if $h_k(t)$ is from unvoiced speech and fluctuates irregularly, then $\varphi_k(t)$ varies randomly over time to cause the overall TFS to oscillate irregularly (McAulay and Quatieri, 1986).

Harmonic selection

To extract the HSSE modulator for a particular channel, the first step is to identify which harmonics are contained in the channel. For example, the spectrogram in Figure 3.2 shows how speech intensity (color scale) varies as a function of time and frequency: the evenly-spaced frequency components represent harmonics, with the bottom one representing F_0. The 4-channel corner frequencies are overlaid on the spectrogram as dashed horizontal lines. One can see that because the channel numbers used in this study are comparatively small (4 and 8), most bands are broad and often contain multiple harmonics.

Due to auditory masking, the strongest harmonic within a channel usually dominates the perception for the related spectral region, whereas the weak components are masked [116]. Similarly, in HSSE processing, the harmonic with the largest magnitude in a channel was selected as the representative of the channel. The magnitude of each harmonic was estimated as the spectrum magnitude at the associated harmonic frequency. For instance, among the three harmonics contained in channel 2, the 2nd harmonic is the strongest and will be selected accordingly.

Frequency downshift

Provided that the k^{th} harmonic was selected for a given channel, it would be transformed into the modulator for the respective channel by a frequency downshift operation. Specifically, the spectrum of $h_k(t)$ would be transposed from its original location kF_0 to the F_0, which is equivalent to multiplying $h_k(t)$ by a F_0-dependent complex exponential function in the time domain [59].

To facilitate the frequency-shifting analysis, the harmonic model in Equation (3.1) was converted into the following analytic form:

$$h_k(t) = Re\{a_k(t)e^{j(2\pi kF_0t+\varphi_k(t))}\} \tag{3.2}$$

where "Re" means taking the real part of a complex signal and symbol $j = \sqrt{-1}$. The analytic representation of $h_k(t)$ was first multiplied by a complex exponential function $e^{-j2\pi(k-1)F_0t}$. The real part of the complex result was then taken to yield the transposed

harmonic $\tilde{h}_k(t)$:

$$\begin{aligned}\tilde{h}_k(t) &= Re\{a_k(t)e^{j(2\pi kF_0t+\varphi_k(t))}\times e^{-j2\pi(k-1)F_0t}\} \\ &= a_k(t)cos(2\pi F_0t+\varphi_k(t)) \end{aligned} \tag{3.3}$$

To differentiate it from the traditional non-negative envelope, $tildeh_k(t)$ was called the HSSE modulator in this study. Comparing Equation (3.1) and (3.3), one can see that conveys the same AM cues as the original harmonic but oscillates at a much slower rate. For voiced speech, $tildeh_k(t)$ would oscillate regularly at the rate of F_0 instead of kF_0. During the unvoiced segments, $tildeh_k(t)$ would appear noise-like because $\varphi_k(t)$ varied randomly over time and caused the overall fluctuation to be irregular, although an interpolated F_0 was used in the frequency downshift operation.

HSSE modulator extraction

Figure 3.3 A shows the functional blocks of HSSE processing. The incoming sound was first filtered into N channels. Within each channel, the strongest harmonic was identified (as described above, yet not included in Figure 3.3 A) and then frequency downshifted, represented as multiplications between band signals and complex exponential functions. As a result, the strongest harmonic within each channel was transposed to the F_0. Next, each transposed band signal was passed through a filter to keep only the strongest harmonic in the modulator. Because the F_0 of human voice is typically above 50 Hz, each transposed signal was first highpass filtered at 50 Hz (3rd-order Butterworth), and then lowpass filtered at 300 Hz (3rd-order Butterworth) to limit the temporal information in the modulator. The combined effect is equivalent to a bandpass filter (50-300 Hz, 3rd-order Butterworth), as shown in Figure 3.3 A. Finally, the real part of the filter output was taken to yield the HSSE modulator for each channel, as implied in Equation (3.3).

As a visual example, Figure 3.2 shows the extracted HSSE modulators of a speech signal whose spectrogram is displayed on the left. For the voiced segment, the extracted HSSE modulators oscillate at a common rate of F_0 and exhibit coherent AM cues across channels, suggesting that they are from the same source. For the unvoiced segment, noise-like TFS

cues are conveyed in the extracted HSSE modulators. One can see that the spectral profile of the speech signal is represented in the relative amplitudes of HSSE modulators between different channels.

In electric stimulation, where a non-negative stimulation signal is required, HSSE modulators can be further half-wave rectified. Alternatively, it is also possible to deliver HSSE modulators without rectification, e.g., by analog stimulation. Given that the purpose of vocoder simulation is to investigate the maximum potential of a strategy, it is not sensible to include rectification in the HSSE simulation, because rectification is an incoherent operation (introduces distortions) and the goal is to instead investigate the potential benefits of non-distorted TFS information to speech perception. In addition, there is a difference between acoustic simulation and electric stimulation regarding how rectification can be used. In acoustic simulation, if a 300-Hz-wide modulator is rectified, it must be lowpass filtered again to constrain the temporal information within 300 Hz; whereas in electric stimulation, it is unnecessary to lowpass filter a rectified modulator, given that the modulator starts as a 300 Hz signal. Because rectification does not necessarily make the simulation more closely resemble CI percepts and distorts the TFS cues, it was not included in the HSSE simulation.

3.2.2 Acoustic simulation of HSSE and CIS

The simulation diagrams of HSSE and CIS are shown in Figure 3.3 A and B, respectively. First, the incoming sound was passed through a 4- or 8-channel analysis filterbank (3rd-order Butterworth) spaced from 80 to 6,000 Hz according to the Greenwood map (1990). The corner frequencies used in the 4-channel simulation were set to 80, 384, 1065, 2588, and 6000 Hz. In 8-channel simulation, the filter cutoff frequencies were 80, 202, 384, 657, 1065, 1675, 2588, 3955, and 6000 Hz.

Next, the temporal encoding of each band was executed. In HSSE, a 300-Hz HSSE modulator was extracted, as described above; whereas in CIS, a Hilbert envelope was extracted and then lowpass filtered at 300 Hz (3rd-order Butterworth). Because the pitch saturation limit in electric hearing is typically restricted to 300 Hz, to prevent vocoders from present-

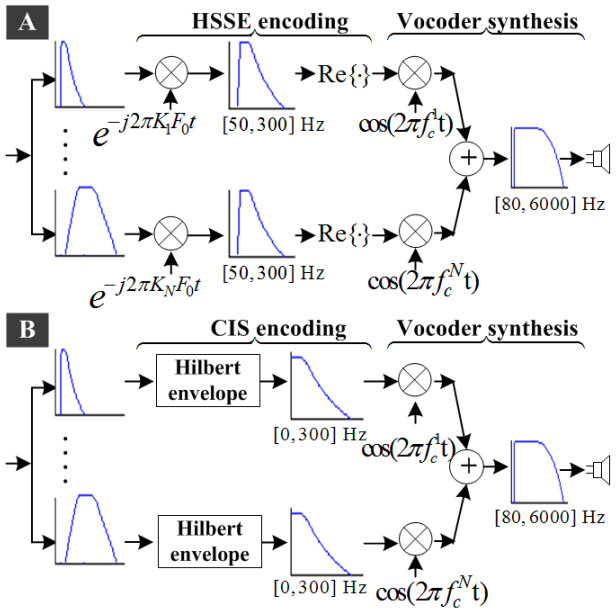

Figure 3.3: Acoustic simulation schemes of HSSE (panel A) and CIS (panel B), with the total number of channels being N. In panel A, the multiplications between band signals and complex exponential functions (e.g., $e^{-j2\pi KF_0 t}$) represent frequency downshift operations; the block $Re\{.\}$ means taking the real part of a complex signal.

ing information that may not be accessible to CI users, both HSSE modulators and CIS envelopes were lowpass filtered at 300 Hz, such that normal-hearing listeners receive similar information as CI users, or at least represent the upper bound of CI performance.

For a visual comparison, the CIS and HSSE encodings of an identical sound, the "eas" segment from "easy", are displayed side by side in Figure 3.2. In the quiet condition, the F_0 cues are conveyed with both strategies yet by different mechanisms. With CIS, the F_0 cues result from the beating of unresolved harmonics, whereas in HSSE the F_0 cues lie in the transformed TFS of selected harmonics. Despite a possible difference in F_0 salience between CIS and HSSE, they seem to convey similar AM cues within each channel. In the noise condition, both AM and F_0 cues are evidently distorted in CIS encoding, with the first two channels being disturbed to a greater extent than the last two channels, because speech-shaped noise produces more interference in low frequencies than in high frequencies. In contrast, the AM and F_0 cues in HSSE encoding appear to be less distorted, because

HSSE modulators are extracted from predominant harmonics that are stronger and thus more resilient to interference [116].

Finally, the temporal encoding of each band is multiplied by a sine carrier at the respective channel center frequency. Tone vocoders were chosen because they are thought to more closely resemble CI percepts, whereas noise carriers introduce random fluctuations in the envelope and distort TFS cues (for review, see [122]). In the last processing step, the modulated carriers are combined and then band-pass filtered again (80—6,000 Hz, 3rd-order Butterworth) to constrain the total information within the analysis spectral range.

3.2.3 Subjects

Ten native English speakers with normal hearing participated in the sentence recognition test. As in all of the tests in this study, subjects were seated in a double-walled, sound-insulated booth (IAC). A program written in MATLAB (Mathworks, Natick, MA) played simulation sounds to subjects and recorded their responses via a graphical user interface. Sounds were presented at 68 dB SPL through an Apple PowerMacG5 sound card connected to a Crown D45 amplifier (Crown Int'l; Elkhart, IN) and a free-standing studio monitor (B&W DM303 speaker from Bowers & Wilkins, North Reading, MA). During the test, the speaker was placed at head level, about 1 m in front of subjects. The University of Washington Institutional Review Board approval and informed consent were obtained from all subjects.

3.2.4 Stimuli and test precedure

Two sets of materials, the Hearing in Noise Test (HINT; [123]) sentences and the Institute of Electrical and Electronic Engineers (IEEE; [124]) sentences, were used to sample a wide range of listening abilities and to avoid a potential ceiling effect during HINT testing (e.g., [125]) and a floor effect with IEEE sentences (e.g., [45]). The target HINT and IEEE sentences were produced by different male talkers with a mean F_0 of 110 and 108 Hz, respectively.

To investigate the effect of HSSE on speech perception, a pilot study was done using

IEEE sentence recognition in quiet: a clear advantage of HSSE over CIS was observed in the 4-channel condition (76% versus 52%), but the performance was too high with both strategies in the 8-channel condition (>90%). HINT sentence recognition was then tested in noise at +10 dB signal-to-noise-ratio (SNR): a ceiling effect was observed again in the 8-channel condition. Thus, a fixed SNR of +5 dB was chosen

Two types of maskers were used in the HINT test: a speech-shaped steady state noise (SSN) and a competing female talker (mean F_0 = 219 Hz). These two types of maskers were also used in the IEEE test; additionally, a male-talker masker (mean F_0 = 136 Hz) was included to explore the effect of masker type on speech perception. The SSN masker was generated by filtering white noise with the target sentence's long-term spectral profile. The competing sentences were selected from the IEEE corpus and varied for each presentation. The masker always lasted longer than the target sentence. The target and the masker were first mixed at +5 dB SNR and then vocoder processed to generate a simulation sound.

To compare the effect of processing strategies, a within-subject test design was used. Each participant listened to both CIS- and HSSE-vocoded speech under the same masker conditions. During the HINT test, eight conditions were used, covering two strategies (CIS and HSSE), two sets of channel numbers (4 and 8), and two types of maskers (SSN and female masker). The 8-channel with SSN masker condition was eventually dropped because subjects' performance was too high (> 90%) to yield meaningful comparison. Previous studies (e.g., [126]) showed similar performance levels. During the IEEE test, results were obtained under six conditions, two strategies (CIS and HSSE) and three types of maskers (SSN, male, and female maskers). Because IEEE sentences are relatively hard to recognize, the channel number was fixed at 8 to avoid the floor effect associated with the 4-channel condition under 5 dB SNR [45].

Five normal-hearing subjects participated in the HINT test and another five, different, normal-hearing subjects participated in the IEEE test. For each subject, the order of test conditions was randomized. Under a given condition, subjects began by listening to two practice sentences to familiarize themselves with the test stimuli. Afterwards, they began the actual test, which consisted of 20 new sentences. Each subject was presented with a different set of sentences for the same condition and was instructed to type in their

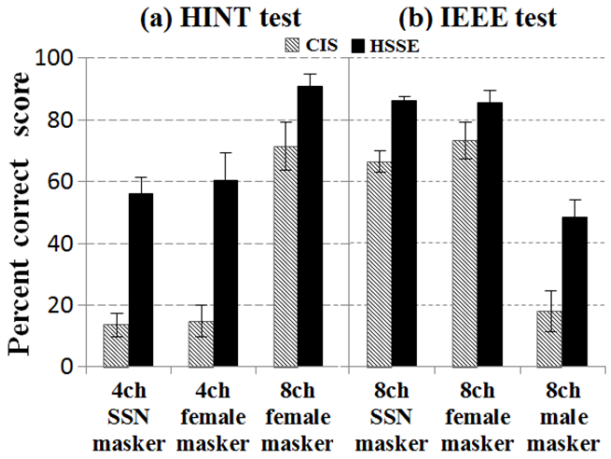

Figure 3.4: The mean intra-subject performance on HINT (a) and IEEE (b) sentence recognition tests with CIS (hatched bars) and HSSE (filled bars) vocoders. The SNR was fixed at 5 dB across all masking conditions. Error bars represent the standard error of the mean.

responses using a computer keyboard. Listeners' performance was evaluated offline by the experimenter as the percentage of keywords correctly recognized.

3.2.5 Results

The average intra-subject performance on the HINT test is shown in Figure 3.4(a). Across all testing conditions, subjects performed significantly better with HSSE than with CIS vocoders. The largest improvement was observed under the 4-channel condition: subjects scored 42% higher with HSSE in steady-state noise [paired $t(4) = 9.7$, $p = 0.001$], and 45% higher against a female competing talker [paired $t(4) = 6.6$, $p = 0.002$]. Under the 8-channel condition with a female-talker masker, subjects also improved significantly with HSSE over CIS vocoders [paired $t(4) = 3.0$, $p = 0.038$]

Figure 3.4(b) shows subjects' average performance on the IEEE test, during which the channel number was fixed at 8. In all of the three masker conditions, subjects scored higher with HSSE than with CIS vocoders. Of the three maskers, subjects performed the worst in the male-masker condition, yet for the same condition, HSSE showed the largest advantage over CIS, approximately 30% [paired $t(4) = 6.8$, $p = 0.002$]. A significant benefit of HSSE

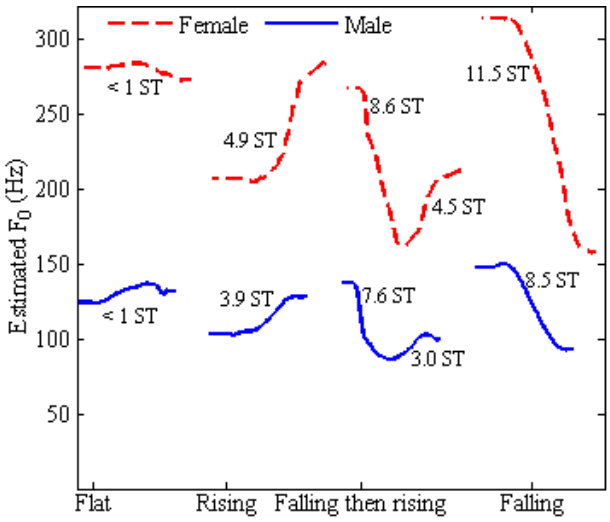

Figure 3.5: The estimated F_0 profile of the female (dashed lines) and the male (solid lines) speaker in pronouncing flat, rising, falling then rising, and falling, respectively. The F_0 variation amount for each tone is displayed along with the F_0 profile. ST stands for semitone.

was also observed for speech recognition in SSN [paired $t(4) = 4.8$, $p = 0.009$] and in the female-masker condition [paired $t(4) = 3.1$, $p = 0.036$].

3.3 Experiment 2: Mandarin tone identification with simulated HSSE and CIS strategies

Understanding of a tonal language depends not only on phoneme recognition, but also on identification of changing fundamental frequencies. For example, in Mandarin Chinese, the monosyllabic word "ji" can be pronounced with either a flat, rising, falling then rising (v), or a falling fundamental frequency. Each word has an entirely different meaning. In experiment 2, the potential effect of HSSE on Mandarin tone identification was investigated, using both acoustic simulation and computational modeling approaches.

3.3.1 Behavioral test

Subjects

Five native Mandarin Chinese speakers with normal hearing participated in experiment 2.

Stimuli and test procedure

The test procedure and stimuli were adopted from Xu et al. [127]. All of the stimuli were vocoder processed in the same way as in experiment 1. The stimuli were drawn from 10 lists. Each list consists of 4 words, all of which have the same syllable but different tonal patterns. The stimuli were pronounced by both female and male speakers. Their F_0 profiles in producing the four words of "ji" list are displayed in Figure 3.5, to show the dependence of the F_0 profile on the gender. To prevent subjects from using duration as one of the identification cues, all syllables in a particular list were made the same in duration through careful selection of multiple recordings.

Tone recognition was evaluated under four conditions, consisting of two strategies (CIS and HSSE) and two sets of channel numbers (4 and 8). The test order of four conditions was randomized within each subject. Under each condition, a four-interval, four-alternative forced-choice paradigm was used. In each particular trial, a word list was first randomly chosen from either the female or the male recordings. One of its four words was then selected, again randomly, as the trial stimulus. Overall, both the 10 female lists and the 10 male lists were presented twice in random order, resulting in a total of 40 lists and 160 trials per condition. Subjects' performance was calculated as the percent of correctly-identified tones.

3.3.2 Computational modeling

Description of the model

Implementation details of the model as well as its parameter set have been presented in [128] and [62], respectively. The input to the model is the electric pulse train generated by a CI processor for a particular electrode. The output is the simulated neural discharge patterns evoked by the input pulse train (see Figure 3.7).

The model is based on the morphology and electrophysiology of spiral ganglion cells (SGCs). In CI users, SGCs serve as a locus where electric stimuli are first converted into neural responses in CI users. Given that subsequent neurological processing can only extract, but not add, information from the incoming stimuli, neural responses at SGCs provide

an upper bound on the auditory information available to CI users. Because the normalized response properties of a population of diameter-distributed model fibers have been shown to match that of the same number of in vivo fibers [62], the same distribution of 250 fibers was used to generate all of the neural outputs in this study.

Generation of electric pulse train

Analogous to the vocoder processing, 8-channel CIS and HSSE implementations were used to generate the electric encoding of a particular stimulus. Because the model was inherently single-channel, neural responses in each spectral channel would be simulated independently. To gauge the best potential of a strategy, the 3rd band ($[384, 657]$ Hz) was selected for simulation, because visual observation suggested that clear F_0 cues were present in this band. With CIS, the F_0 fluctuation were due to beats of unresolved harmonics, while in HSSE the F_0 cues were represented in the TFS of a resolved harmonic.

As described earlier, a 300-Hz-wide CIS envelope and HSSE modulator were first extracted from the 3rd band; they were then logarithmically compressed and converted into an electric pulse train, respectively, using the Nucleus MATLAB toolbox [129]. The negative samples in HSSE modulators were set to zero during log-compression, i.e., HSSE modulators were half wave rectified. To generate a faithful representation of TFS, the per channel stimulation rate was set to 1900 Hz. Each pulse was biphasic and 25-μs wide. The cathodic phase was applied first, followed by an 8-μs gap and an equal-amplitude anodic phase of the pulse.

Stimuli and analysis of results

Due to the high computational cost of simulating neural responses for the whole set of tone stimuli, only the "ji" list was selected, for which both the CIS and the HSSE pulse trains exhibit F_0 cues to allow a further comparison of F_0 encoding with the evoked spike trains. Each stimulus of the list is a 600-msec recording of a female speaker pronouncing the word "ji" with one of the four tonal patterns. The estimated F_0 profile for each stimulus is shown in Figure 3.5.

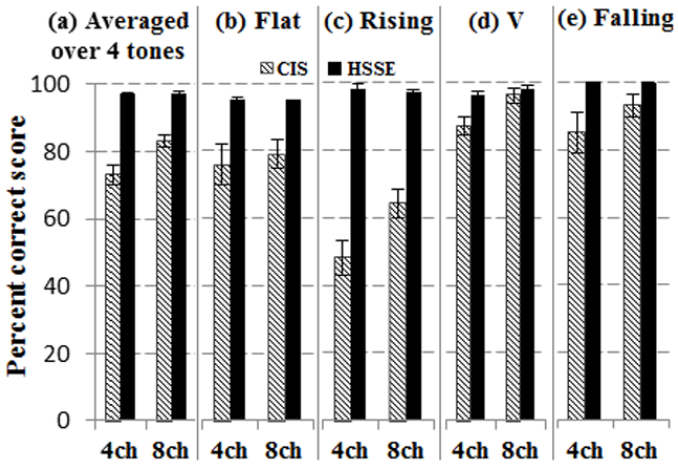

Figure 3.6: (a) The average intra-subject performance on Mandarin tone identification with CIS (hatched bars) and HSSE (filled bars) vocoders. Panel (b)-(e) shows subjects' mean score on recognizing flat, rising, falling then rising, and falling tones, respectively. Error bars represent the standard error of the mean.

To examine the amount of F_0 information captured by the auditory nerve, raster-plots of the simulated neural spikes along 250 fibers were generated and evaluated qualitatively (see Figure 3.7). Furthermore, the intervals between successive spikes were analyzed. Given that the F_0 cues were encoded temporally by both strategies, an increase in the speaker's F_0 would produce a decrease in the interspike intervals (ISIs). The converse also holds true: as F_0 decreases, the spikes should occur more sparsely, producing larger ISIs. Therefore, by measuring how the ISI changes as the speaker's F_0 evolves, a strategy's ability to convey F_0 cues was evaluated.

3.3.3 Results

Behavioral results

Overall, subjects could identify Mandarin tone better with HSSE than with CIS vocoders. Their mean score, averaged across all the four tonal patterns, is shown in Figure 3.6(a). Subjects scored 24% higher with HSSE than with CIS vocoders in the 4-channel condition [paired $t(4) = 8.8$, $p = 0.001$] and 14% higher in the 8-channel condition [paired $t(4) = 8.3$,

$p = 0.001$]. Their mean score on recognizing each individual tonal pattern is displayed in panels (b)-(e), respectively.

With CIS, the recognition performance on rising tone was the worst; HSSE demonstrated the greatest benefit in identifying rising tones. A 3-way repeated measures ANOVA (strategy, channel number, tonal pattern) revealed that processing strategy [$F(1, 4) = 200.3$, $p = 0.001$], channel number [$F(1, 4) = 30.2$, $p = 0.005$], and tonal pattern [$F(3,12) = 19.5$, $p = 0.001$] all had a significant effect on tone identification. There was a significant interaction between strategy and channel number ($p = 0.030$), resulting from the fact that subjects' score with CIS varied substantially as a function of channel number (see panel (c)). A significant interaction was also observed between strategy and tonal pattern, reflecting that subjects' score with CIS was largely affected by tonal pattern. In contrast, subjects, performance with HSSE was consistently good across all of the four tone patterns under both channel conditions (ranged 95%-100%).

Modeling results

Raster-plots Due to space constraints, only the raster-plot for the ji(v) stimulus is provided, which shows both directions of F_0 glide. Figure 3.7 is divided into two subpanels. The CIS encoding and the evoked neural responses are displayed in panel A, while panel B shows the results of HSSE. Within each panel, the top row shows the input electric pulse train; the bottom shows the simulated neural spike train by placing a dot for every occurrence of a spike.

Comparing panel A and B, one can see that the CIS and the HSSE evoked spike trains exhibit similar envelope cues, e.g., the synchronized onset and offset patterns, but they convey different timing cues. The HSSE-evoked spike train displays clear troughs and peaks following the F_0 cues in the HSSE pulse train; whereas such a timing pattern is missing in the CIS-evoked spike train. Although the F_0 cues in the CIS pulse train are visible, the modulation depth is comparatively shallow. Consequently, the CIS stimulation causes saturation in large-diameter fibers ($> 3.6\mu$m), forcing most of the F_0 cues to reside only in low-diameter fibers ($< 3.6\mu$m). In contrast, HSSE might encode F_0 in the duration of a

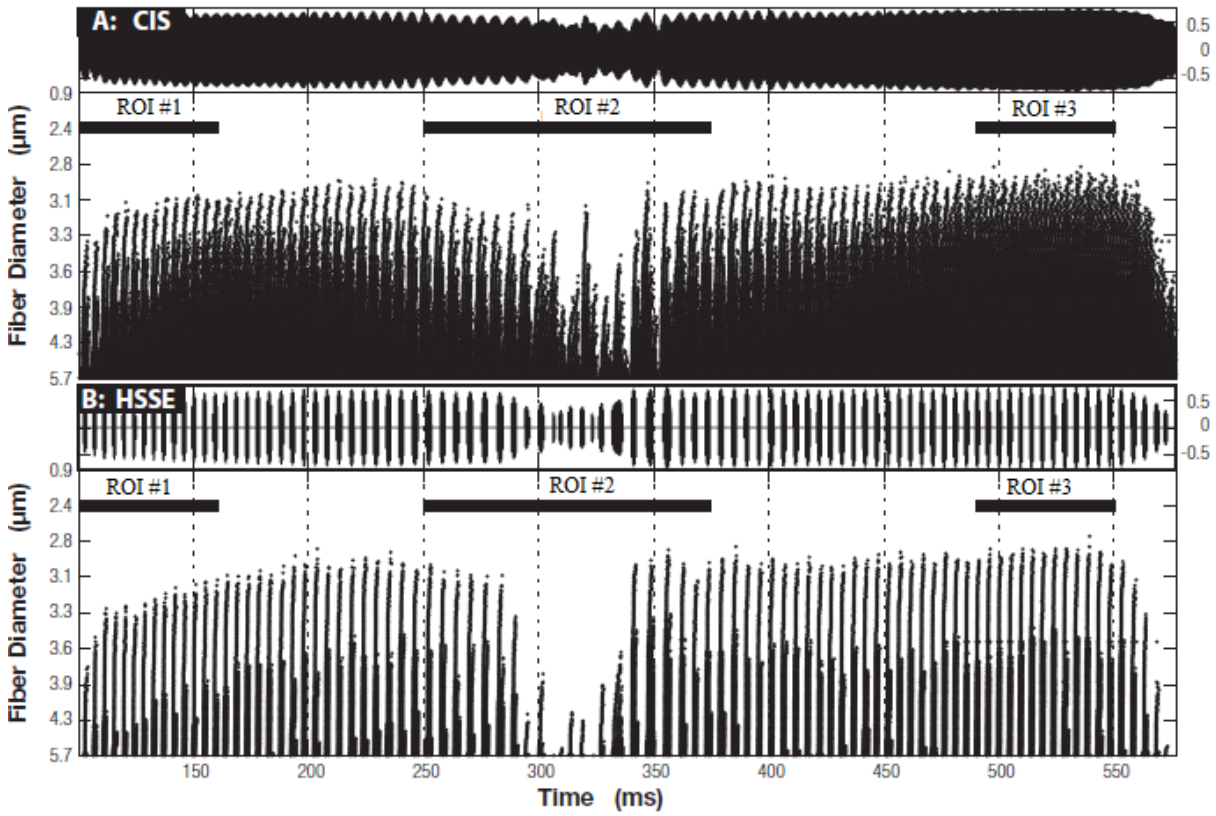

Figure 3.7: Simultated neural discharge patterns under CIS (panel A) and HSSE stimulation (panel B), respectively. Within each panel, the top row shows the electric encoding of stimulus ji(v) by a CIS/HSSE processor for one electrode; the bottom displays the raster-plot of the evoked neural spike train. The regions labeled ROI # 1, 2, and 3 will be sampled in the analysis of interspike intervals.

pulse burst and in the interval between successive bursts. Both properties are captured in the simulated HSSE spike trains: e.g., as the speaker's F_0 decreases between 100 and 290 msec, both the width of the spike bursts and the interval between successive bursts increase noticeably.

Interspike interval (ISI) histogram Because the speaker's F_0 is time-varying (see Figure 3.5), each neural spike train was accordingly time-divided into three segments, designated as regions of interest (ROIs) #1, #2, and #3. The location of each ROI is indicated in Figure 3.7 as black bars placed on top of the raster-plots. For each ROI, the intervals between successive spikes were calculated and the overall ISI distribution is shown in Figure 3.8. Panel A-D corresponds to stimuli ji falt, ring, v, and falling, respectively. Within each panel, the histograms on the left were sampled from the CIS-evoked spike trains, and those on the right were from the HSSE-evoked responses. The peak of each ISI histogram, corresponding to the speaker's F_0 at the respective ROI, is indicated by a dashed vertical line. Because the speaker's F_0 was lower than 500 Hz, only ISIs longer than 2 msec were considered in locating a histogram's peak

Comparing the histograms within each panel, for an identical stimulus, the peak location is the same between CIS and HSSE histograms, but the peak height is notably lower with CIS, sometimes even invisible. This difference in peak height arises because fewer F_0 cues are available in the CIS- than in the HSSE-evoked spike trains. With HSSE, the ISI histograms exhibit clear peaks, correctly capturing the evolution of a particular tonal pattern: the F_0-related peak remained in the same location in response to ji (flat), shifted to the right and left when stimulated with ji (falling) and ji (rising), respectively, and exhibited a clear displacement in response to ji (v). In fact, one can identify the tone supplied to the model based solely on the HSSE yielded ISI distributions. A comparable interpretation is considerably harder with CIS: while it might be possible to deduce the F_0 profile of ji (rising), the absence of histogram peaks makes the remaining identifications difficult.

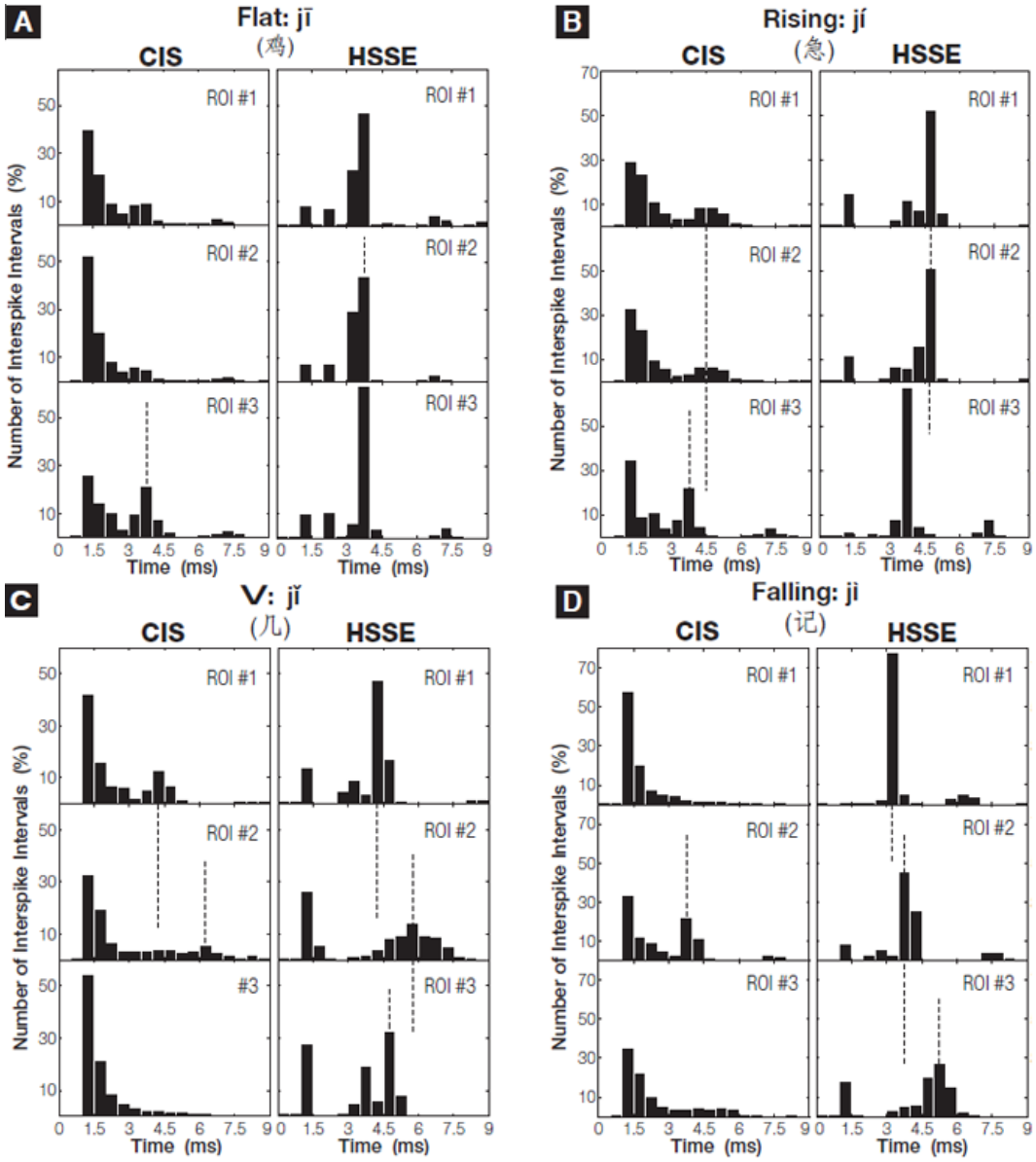

Figure 3.8: Histograms of interspike intervals. Panel A-D corresponds to stimulus ji flat, rising, falling then rising, and falling, respectively. Within each panel, the histograms on the left were sampled from a CIS-evoked spike train and those on the right were from an HSSE-evoked spike train. From top to bottom, the sample regions are sequentially ROI #1, #2 and #3. Vertical dashed lines indicate histogram peaks; only ISIs longer than 2 msec were considered in locating the peaks, because the stimuli's F_0 is below 500 Hz. Each histogram bin is 0.5-msec wide.

3.4 Discussion

3.4.1 The effects of HSSE on simulated speech recognition in noise

This study showed that by encoding non-distorted harmonic and TFS information, HSSE can potentially improve perception of speech in a variety of masking conditions as well as Mandarin tones. During both 4- and 8-channel simulations, HSSE demonstrated a clear advantage over CIS for all listening tasks tested. Specifically, a larger benefit of HSSE encoding was observed in 4-channel than in 8-channel condition, suggesting an interaction between spectral and temporal cues for speech perception. Xu et al. [127] reported that under a given channel condition, Mandarin tone identification with CIS can be improved by increasing the envelope cutoff frequency; the improvement was relatively larger when the channel number was smaller. Stone et al. [130] observed a similar interaction between spectral and temporal cues for speech recognition against a competing talker. Presumably, when the channel number was decreased from 8 to 4, the temporal cues in HSSE can better compensate for the diminished spectral cues for speech perception; although the same temporal cutoff frequency was applied in both HSSE and CIS simulations.

Among the three types of maskers tested, the female and the male talker were found to generate more masking than the SSN masker, which is consistent with previous studies on the effect of masker type on speech recognition with CIs (e.g., [45]). Between the female (mean $F_0 = 219$ Hz) and the male (mean $F_0 = 136$ Hz) talker, the male-masker condition was more difficult, because the target speaker (mean $F_0 = 108$ Hz) was also a male and the F_0 difference between he and the male masker was much smaller than that between he and the female masker. This effect of the F_0 difference on speech recognition in a competing-talker background was also reported by Cullington and Zeng [126]. Presumably, TFS information plays a larger role in speech recognition with a competing talker of the same gender, thus a greater benefit of HSSE encoding was observed in the male-masker condition than in the other two conditions.

There are several possible reasons why HSSE is advantageous over CIS for speech perception. First, HSSE uses frequency downshift operations to transform speech harmonics into modulators, whereas CIS uses incoherent operations to extract channel envelopes. As

illustrated in Figure 3.1, incoherent operations distort the frequency components in a signal, thus speech harmonics cannot be properly represented in CIS as in HSSE. Consequently, the potential benefit of harmonic information to speech perception is more likely to be restored by HSSE than by CIS. Second, HSSE appears to encode F_0 cues better than CIS. In CIS, the F_0 fluctuation is due to the beating of harmonics, which would be obviously distorted in the presence of noise (see Figure 3.2); whereas the F_0 cues in HSSE modulators are extracted from predominant harmonics, which are relatively resilient to noise. Previous studies have shown that the F_0 information is an important segregation cue for speech recognition against a competing talker [131, 118]. The advantage of HSSE relies on the F_0 tracking accuracy, which is generally good at +5 dB SNR if the target and the competing talkers' F_0 are apart to some degree. Therefore, the F_0 cues are more likely to be available to HSSE than to CIS listeners as a segregation cue for speech recognition in noise. Third, HSSE might encode TFS information better than CIS under the 300 Hz cutoff frequency. In the noise condition, because the speech AM cues are obscured by noise in the extracted envelope (see Figure 3.2), additional TFS information is needed to assist the separation of intelligibility-related AM cues [48, 49]. The temporal details in CIS envelopes contain distortions caused by incoherent operations, whereas the TFS cues in HSSE modulators are linearly extracted and thus free of distortions. Presumably, the TFS information in HSSE modulators is more beneficial to speech perception than the information in CIS envelopes.

3.4.2 The effects of HSSE on simulated Mandarin tone identification

Using HSSE vocoders, subjects could reliably identify each tonal pattern ($> 95\%$) under both 4- and 8-channel conditions; whereas with CIS, their performance was comparatively poor under both channel conditions. The fact that subjects scored higher with a 4-channel HSSE vocoder than with an 8-channel CIS vocoder indicates that under a spectral constraint of 4-8 channels, tone identification relies primarily on temporal cues, while place cues can only assist F_0 discrimination to a lesser degree [127].

With CIS, subjects' performance appeared to be dependent on tonal patterns: the recognition score on rising tones was considerably worse than that on falling tones, which was in

line with the performance pattern observed in CI users [109]. Subjects' performance with HSSE, however, was consistently high across all of the four tonal patterns. Presumably, there might be a smaller F_0 change during the evolution of a rising tone than a falling tone (Figure 3.5). Because the F_0 cues conveyed in CIS envelopes were not salient, it was hard for CIS listeners to identify small F_0 changes during a rising tone. For example, Han et al. [109] found that CI users often misperceived rising tones as flat, suggesting that the F_0 cues in CIS envelopes were too weak to elicit an accurate pitch percept.

The difference between CIS and HSSE in temporal F_0 encoding was also demonstrated in neural response simulation. For an identical tone stimulus, the CIS and the HSSE evoked spike trains exhibited a similar AM pattern, but different timing pattern. Although the F_0 cues in the CIS pulse train were visible, the modulation depth was comparatively shallow. Under a constant pulse rate, the CIS stimulation caused saturation in large-diameter fibers ($> 3.6\mu$m), forcing most of the F_0 cues to reside only in small-diameter fibers ($< 3.6\mu$m). In contrast, HSSE may encode F_0 in the duration of a pulse burst and in the interval between successive bursts. Both properties were captured in the simulated HSSE spike trains. As found in the ISI histogram analysis, the dominant interval in an HSSE-evoked spike train corresponded to the stimulus' fundamental frequency period; whereas such a temporal code was not evident in the CIS-evoked responses. Although the spike data were only about one tone list, they provided insights on how the difference between CIS and HSSE in F_0 coding mechanisms would result in a difference in the timing patterns of evoked neural responses. Thus, the modeling results supported the behavioral test results, indicating that HSSE is potentially advantageous over CIS for enhancing tonal pattern identification with CIs.

3.4.3 Implications for cochlear implants

Due to the inherently coarse spectral and temporal resolution in electric hearing, how to encode harmonic and TFS information for CI uses is still an open question. Various strategies have been proposed or implemented for improving TFS representation in CIs, most of which follow a scheme of separately extracting the AM and TFS cues of a band both by nonlinear operations, e.g., the strategy proposed by Nie et al. [113] and the FSP strategy [111].

During electric stimulation, the AM cues are encoded in the amplitude of pulses, while the TFS cues are represented in the timing pattern of pulses. The F0mod strategy follows a similar scheme except that it directly encodes F_0 instead of TFS cues in addition to the AM information [114]. These strategies are like CIS in terms of AM extraction: the operation involved introduces nonlinear distortions and causes misrepresentation of speech harmonics. In contrast, HSSE modulators can closely resemble the original harmonics. Because CI users typically have a small number of effective channels, HSSE delivers only predominant harmonics that are resilient to interference and thus important to speech recognition in noise [116]. The fact that a subset of predominant harmonics—instead of a complete set—can be sufficient to benefit speech perception suggests the feasibility of HSSE with CIs.

Another distinction of HSSE is that the TFS cues in HSSE modulators are never separated from the AM cues except being transformed from high to low frequency as a whole; whereas CIS-like strategies extract AM and TFS cues separately. Schatzer et al. [111] showed that the FSP strategy produced similar performance as CIS on Cantonese tone recognition, regardless of the TFS representation from 100 to 800 Hz, suggesting that the TFS cues encoded in FSP might not be accessible or beneficial to CI uses. Milczynski et al. [115] demonstrated that the F0mod strategy outperformed the ACE strategy on Mandarin tone perception but not on sentence recognition. Typically, CI users can perceive temporal information up to about 300 Hz. Under the temporal constraint of 300 Hz, HSSE produced better sentence recognition and Mandarin tone identification performance than CIS in vocoder simulations. Moreover, the spike data showed that the temporal cues in HSSE modulators can be translated into the simulated neural responses (see Figure 3.7). Given that the models response properties have been shown to closely resemble the response quantities measured *in vivo* [120, 62], these results suggest that HSSE is a promising strategy to enhance speech perception with CIs. The reduced accuracy in phase locking poses a great challenge for CI users to perceive TFS. Compared with the zero crossing information encoded by FSP, the low-frequency TFS cues in HSSE modulators are potentially more accessible and beneficial to CI users.

HSSE is aimed to improve temporal encoding in CIs and the present study has demonstrated its potential benefit to speech perception. However, it might be subject to several

limitations in electric hearing, such as neural degeneration and poor temporal resolution. If there is neural degeneration in patients, their perception is likely to be adversely affected. The extent of neural degeneration would influence the extent to which the improved temporal information could be effectively used, which would result in individually variable benefits for HSSE. On the other hand, to implement HSSE in real-time, an efficient F_0 tracker is required. Pitch tracking is technically solvable in high SNR conditions, e.g., >5 dB, although the tracking process increases the computational cost [132]. Once the F_0 is known, the extraction of each individual harmonic can be executed in parallel. Given the computational resource of a modern processor, it is feasible to implement HSSE in real time.

More than a quarter of the world's population use tonal languages. The relatively poor tone identification performance with CIS-like strategies poses a significant challenge to CI users relying on tonal languages. The proposed HSSE strategy can potentially make a great impact on their life, given that the F_0 cues conveyed by a 4-channel HSSE vocoder can support reliable tone identification and these F_0 cues can translate to the timing patterns of simulated neural responses. In addition to the F_0 cues for voiced speech, HSSE can also preserve noise-like TFS cues for unvoiced speech, possibly producing a more natural speech signal than CIS. By improving the encoding of harmonic and TFS information, HSSE appears to be a promising strategy to improve speech perception with CIs.

Chapter 4

IMPROVED PERCEPTION OF MUSIC WITH A HARMONIC BASED ALGORITHM FOR COCHLEAR IMPLANTS

In Chapter 3, we have demonstrated the potentially large benefit of harmonic and temporal fine structure information for speech perception with cochlear implants (CIs). In this chapter, we will explore their effectiveness on music perception. Many music signals contain multiple instrument sources. However, for most CI users, it is difficult even to perceive basic music elements, such as pitch and timbre. Thus, we will focus on single instrument music signals in the following experiments [1].

4.1 Introduction

Modern CI systems can allow profoundly deaf people to achieve nearly normal performance on speech recognition in quiet. Despite this great success, there remain significant limitations in CI performance on tasks including speech recognition in noise [47], lexical tone discrimination [56], and music perception [133, 134, 135, 57, 136]. Specifically, present-day CI systems are unsatisfactory at conveying two fundamental music elements–pitch and timbre–as indicated by the large performance gap between CI users and normal-hearing listeners on music recognition tasks [137, 138]. The melody recognition test is often used to evaluate listeners' pitch perception, while the musical instrument recognition test is designed to assess their timbre perception. Previous studies have found that both postlingually deaf adult implantees and prelingually deaf pediatric implantees [58] performed significantly worse than their normal-hearing peers on these two tasks. One major factor in these outcomes appears to be the lack of spectral and temporal fine structure information in current CI signal processing strategies [50, 139].

The continuous-interleaved-sampling (CIS) strategy is the basis of the encoding ap-

[1] The work described in this chapter was published as Li, Nie, Imennov, Rubinstein, and Atlas (2013) IEEE Trans on Neural Systems and Rehabilitation Engineering.

proaches used in virtually every modern CI processor. Working like a vocoder [26], it encodes an audio signal by first filtering the signal into 12-22 frequency bands. Each band can be represented as a time-varying waveform that conveys both envelope and fine structure information. The envelope typically varies slowly in time, having amplitude modulations; while the temporal fine structure is fast-oscillating zero crossings and acts as a carrier. Traditionally, the envelope is considered the primary cue for speech recognition [39]. Thus in CIS processing, only the envelope of each band is extracted and then converted into current levels to amplitude modulate a constant-rate pulse train that is delivered to the respective electrode [50]. Modern CIs have at least 12 electrodes, but the effective number of channels CI users have is typically less than 8; whereas a normal human cochlea has about 1000 inner hair cells that can convey spectral information. Despite this great reduction in spectral resolution, CI users can understand speech reasonably well in quiet with just envelope cues from as few as 4 channels. However, for music perception, the lack of spectral and temporal fine structure information presumably reduces the accessibility of important neural codes that give rise to the pitch and timbre percepts, which might partly account for CI users' difficulties in perceiving melodies [140].

Pitch is the perceptual attribute associated most with melodies in music and with intonation in speech. For a harmonic complex sound, such as voiced speech or music, which consists of a series of harmonically related tones that are integer multiples of a common fundamental frequency (F_0), the elicited pitch percept is unified and associated with the F_0 regardless of the relative amplitude of harmonics. Because lower harmonics can be resolved on the basilar membrane and encoded by the auditory nerve at specific places, pitch is assumed to be derived from the place information about individual harmonics [141]. Alternatively, temporal models assert that pitch is related to the time intervals between auditory neural spikes, because higher harmonics cannot be well separated on the basilar membrane and the F_0 of a complex waveform can be derived by pooling timing information across neural fibers regardless of place cues [17]. Both mechanisms can account for pitch perception to some extent, but the lower harmonics seem to play a dominant role, because they can elicit a more salient and accurate pitch percept than the unresolved harmonics [117]. Nevertheless, current CI signal processing strategies do not make effective use of this

pitch coding mechanism: e.g., harmonics generally cannot be separated by the filters of a typical implant processor; even if a harmonic can be separated, its frequency, conveyed in the temporal fine structure, is discarded during the conventional envelope-based encoding scheme. Consequently, the electrically-evoked neural responses fail to provide adequate place or temporal code for pitch derivation, leading to poor pitch perception in CI users [140].

For timbre perception, harmonic and temporal fine structure information are also needed. Timbre is the perceptual attribute by which "a listener can judge that two sounds having the same pitch, loudness, and duration are different" [142]. The major determinants of timbre include: spectral energy distribution, attack time, and fine structure information [143, 144, 145, 136]. One quantitative measure of spectral fine structure, called the spectral irregularity, is based on the relative amplitude of harmonics; this parameter is proven to be one of the dominant cues for timbre recognition [145, 136]. Additionally, the relative timing of onsets and offsets of upper harmonics, as well as their phase coherence (conveyed in the temporal fine structure), have also been found to be important timbre cues [143, 144]. It is thus evident that harmonic information is important for timbre perception. However, with current CI encoding strategies, only the attack time is preserved in the envelopes. The spectral shape, to a lesser extent, can be represented in the relative envelope magnitude across channels. Kong et al. [136] have found that CI users mainly rely on the attack time to recognize instruments, while normal-hearing listeners can take advantage of different cues to recognize a particular timbre. Thus, CI users generally have much lower subjective ratings and poorer recognition scores on timbre perception than normal-hearing listeners [138].

To encode temporal fine structure information for CI users, an important caveat to note is the poor temporal resolution in electric hearing. Most CI users cannot discriminate changes in the repetition rate of an electric waveform above about 300 Hz [24, 108]; whereas in the normal auditory system, temporal fine structure is perceivable up to approximately 5 kHz [140]. To deliver perceptible temporal cues to CI users, Nie et al. proposed to transform temporal fine structure into frequency modulation information and uses it to frequency modulate the pulse rate [113]. The Fine Structure Processing strategy bases the pulse triggering pattern in a particular channel on the zero crossings of the respective

band waveform [111]. Alternatively, the F0mod [146, 147] strategy modulates channel envelopes at the input sound's F_0 to enhance temporal pitch cues. The eTone [132] strategy, based on a harmonic probability measure in each channel, combines F_0 modulated envelopes with the channel signals extracted by the advanced combinational encoder (ACE) [51]. To represent spectral fine structure, more independent functional channels are needed. Current focusing and current steering stimulation paradigms are being investigated, which attempt to increase, respectively, the spectral resolution and the number of distinctive perceptual channels available to CI users by simultaneous application of currents to multiple electrodes [148]. The effectiveness of these new strategies for enhancing music perception has yet to be demonstrated in CI users [50].

Given the importance of harmonic and temporal fine structure information for music perception, a harmonic-single-sideband-encoder (HSSE) strategy has been proposed in our previous studies, which explicitly tracks the harmonics of complex sounds and transforms them into modulators conveying both amplitude modulation and temporal fine structure cues to electrodes [119, 59, 60]. It is hypothesized that by delivering harmonic and temporal fine structure information to CI users, their pitch and timbre perception will improve. To test this hypothesis, the effectiveness of HSSE on melody and timbre recognition was evaluated, using both acoustic simulations in normal-hearing listeners and acute testing in CI users. Additionally, to examine whether the temporal cues encoded by HSSE can be translated into appropriate neural responses, the spike patterns evoked by HSSE were simulated using an auditory nerve model [128, 62]. The stimuli used in all of the experiments contain only a single musical source associated with a particular F_0 that is varying over time. The experimental design is provided in Section 4.3 after a description of HSSE processing in Section 4.2.

4.2 A Harmonic-single-sideband-encoder Strategy

The functional blocks of HSSE are shown in Figure 4.1. To encode harmonics for CI users, the F_0 of an incoming sound is first estimated, such that F_0's harmonics can be analyzed. Given the F_0 estimate and the frequency spacing of electrodes, each harmonic is associated with an appropriate electrode based on spectral correspondence. To encode harmonics

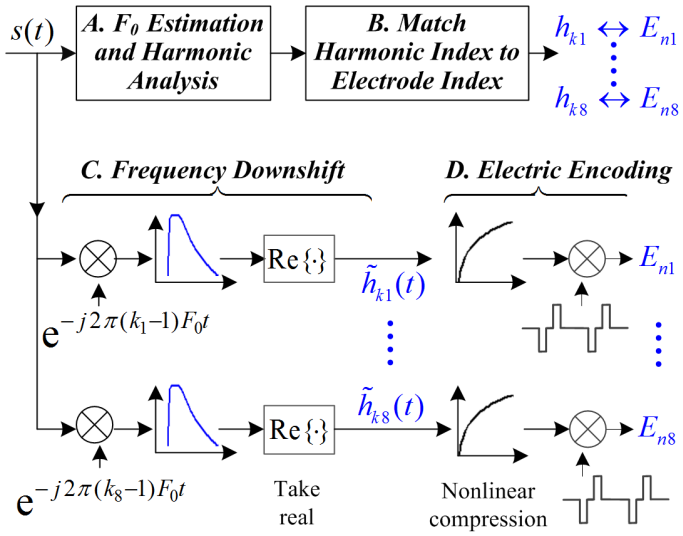

Figure 4.1: Block diagram of the proposed harmonic-single-sideband-encoder strategy. The symbol $h_k(t)$ stands for the k^{th} harmonic, E_n represents the n^{th} electrode, $\tilde{h}_k(t)$ is the modulator derived from the k^{th} harmonic, $j = \sqrt{-1}$, and Re$\{\cdot\}$ means taking the real part of a complex signal.

for CI users, they are first transformed into modulators through frequency shift operations, then logarithmically compressed, and eventually converted into electric pulse trains for their respective electrodes.

4.2.1 F_0 Estimation and Harmonic Analysis

To estimate the F_0 of a music signal, a least-squares harmonic model was used to detect the signal's underlying harmonic structure in the frequency domain [121]. Based on the detection, the F_0 and its harmonics can be derived. Figure 4.2(a) shows the waveform of a guitar note, denoted as $s(t)$. The magnitude spectrum of its first 50 ms is displayed in Figure 4.2(b), where the evenly-spaced frequency components represent harmonics, with the lowest one representing the F_0. Let us denote an individual harmonic as $h_k(t)$, with k being the harmonic index. The signal $s(t)$ can then be represented as:

$$s(t) = \sum_{k=1}^{K} h_k(t) = \sum_{k=1}^{K} a_k(t) cos(2\pi k F_0 t + \varphi_k(t)) \tag{4.1}$$

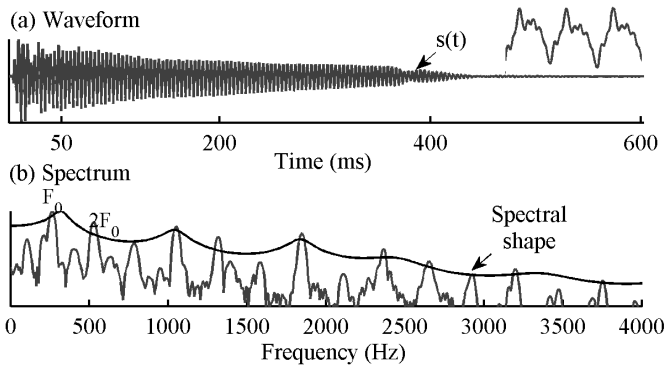

Figure 4.2: (a) Waveform of a guitar note ($F_0 = 262$ Hz). The inset shows the waveform detail over a 12-ms window. (b) Magnitude spectrum of the first 50 ms of the note. The spectral shape is estimated by a 16th-order linear prediction filter and overlaid on the spectrum.

where K is the total number of harmonics under a given sampling rate. Each specific harmonic $h_k(t)$ has its own amplitude modulation $a_k(t)$, frequency kF_0, and phase function $\varphi_k(t)$.

The temporal cues relevant to pitch perception lies in the fine structure information oscillating at harmonic frequencies, while the attacks and decays pertinent to timbre perception are conveyed in the harmonic amplitudes, i.e., $a_k(t)$ as a function of time t for a given index k. Also, the spectral shape can be represented in the relative harmonic amplitudes, i.e., $a_k(t)$ as a function of index k at a given time t. The phase coherence cues for timbre perception are conveyed in $\varphi_k(t)$. If $\varphi_k(t)$ varies randomly over time, causing the temporal fine structure to fluctuate irregularly, then a noise-like quality will be evoked in the auditory perception. In contrast, if $\varphi_k(t)$ stays constant, leaving the temporal fine structure to fluctuate regularly at the rate of kF_0, then a tone-like quality will be elicited.

4.2.2 Match Harmonic Index to Electrode Index

While a complex sound usually contains more than 8 harmonics, only 8 stronger ones were selected for electric stimulation due to several considerations. First, complex tones produced by a musical instrument typically have a formant structure that is determined by

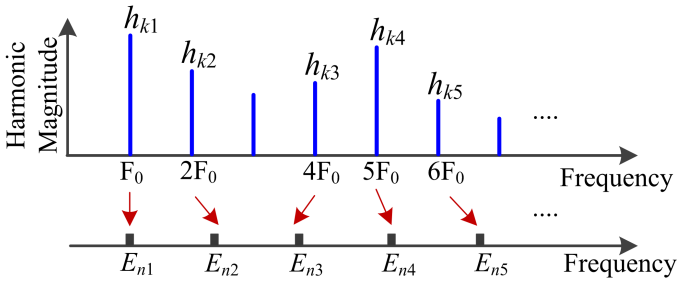

Figure 4.3: Illustration of the harmonic electrode matching procedure. The symbol h_k stands for the k^{th} harmonic, and E_n represents the n^{th} electrode.

the distinctive size and shape of the instrumental air cavity. The harmonics falling in the formant regions are intensified, regardless of the F_0 of the tone. For timber perception, it is more important to deliver the stronger harmonics, because they represent the instrument's formants. Second, CI users typically have less than 8 effective channels. It is therefore more practical to deliver only the predominant harmonics to them.

To identify which harmonics are stronger, one way is to compare the harmonic magnitudes on a note-by-note basis. For instance, one can take the first 50 ms of a note and calculate its Fourier spectrum. As shown in Figure 4.2(b), the magnitude of each harmonic can be estimated as the spectrum magnitude at the corresponding harmonic frequency. The stronger harmonics are then identified as those with larger magnitudes. Alternatively, if the instrument's formant regions are known as a priori information, or estimated during the first few notes, then the predominant harmonics of each note can be identified, in a more efficient way, as those falling in the formant regions. The former approach was used in our experiment, but the latter was more suitable for real time implementation.

Figure 4.3 shows the procedure to identify stronger harmonics and assign them to appropriate electrodes. The harmonic with the largest magnitude was first selected and assigned to an electrode showing the best match in the subject's clinical frequency map. The second strongest harmonic was then selected and assigned. This harmonic electrode matching procedure was executed repeatedly until all the 8 pairs were specified. A selected harmonic electrode pair is denoted as h_k and E_n in Figure 4.3, meaning that the k^{th} harmonic will

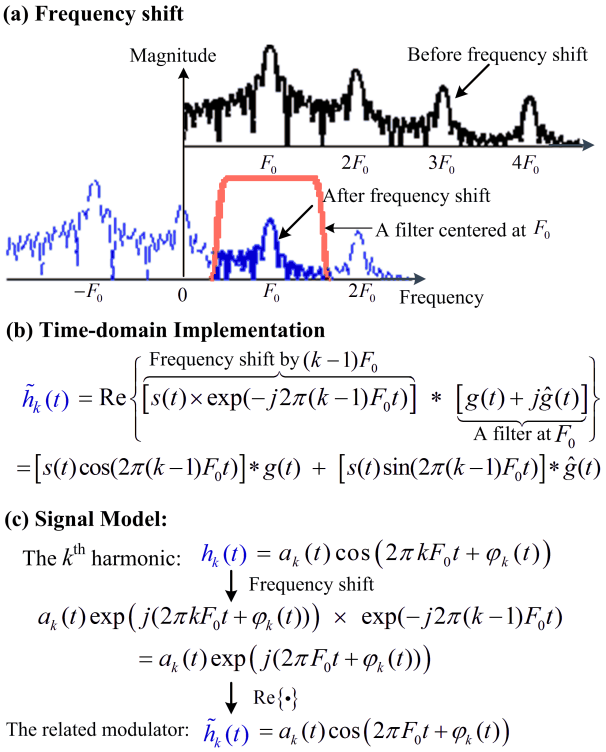

Figure 4.4: (a) Frequency shift illustration: the 3rd harmonic is transposed to the F_0 as a result. (b) Time-domain implementation, which follows the diagram shown in Figure 4.1. The symbol $j = \sqrt{-1}$, "*" represents convolution, $\mathrm{Re}\{\cdot\}$ means taking the real part of a complex signal, and $\hat{g}(t)$ is the Hilbert transform of $g(t)$. (c) Mathematical comparison between the original harmonic and the extracted modulator.

be delivered to the n^{th} electrode during electric encoding. Throughout the duration of a note, the relative magnitudes between harmonics may change. Yet the strategy sticks to the same 8 harmonics, such that the relative timings of their onsets and offsets can be properly represented for timbre perception [144]. Once the F_0 changed from one musical note to another, the matching procedure was started over again to specify 8 new harmonic electrode pairs.

4.2.3 Frequency Downshifting

Given the specified mapping between harmonics and electrodes, each selected harmonic is transformed into a modulator by a frequency shift operation (also known as "coherent

demodulation" [119, 59, 60]), as illustrated in Figure 4.4(a). To extract the k^{th} harmonic, the actual implementation is shown in Figure 4.4(b). First, the input $s(t)$ is multiplied by $exp(-j2\pi(k-1)F_0t)$, such that the k^{th} harmonic is transposed to the F_0. Then, $s(t)$ is passed through a filter, i.e., convolved with the filter's impulse response function $g(t)+j\hat{g}(t)$, where $g(t)$ represents a bandpass filter at F_0 and $\hat{g}(t)$ is its Hilbert transform. Because the information about $h_k(t)$ is only located around F_0, an analytic filter is constructed by $g(t) + j\hat{g}(t)$, to let through only the positive spectrum around F_0. Next, the modulator, denoted as $\tilde{h}_k(t)$, is yielded by taking the real part of the complex filter output. For increased clarity, an equivalent implementation that does not involve complex signal processing is also provided in Figure 4.4(b). In our experiments, $g(t)$ was designed as a 256-tap Finite Impulse Response filter, which was 16 ms long under a sampling rate of 16 kHz.

The mathematical comparison between the original and the extracted harmonic is shown in Figure 4.4(c). Given the k^{th} harmonic, represented as $a_k(t)cos(2\pi kF_0t+\varphi_k(t))$ in equation (4.1), it is first converted into an analytic form, $a_k(t)exp(j(2\pi kF_0t + \varphi_k(t)))$, such that it has a single-sided spectrum. The analytic $h_k(t)$ is then multiplied by $exp(-j2\pi(k-1)F_0t)$ to be transposed to the F_0. Next, the modulator $\tilde{h}_k(t)$ is yielded by taking the real part of $h_k(t)$. Compared with $h_k(t)$, $\tilde{h}_k(t)$ conveys the same amplitude $a_k(t)$ and phase $\varphi_k(t)$, but oscillates at the rate of F_0 instead of kF_0. Because CI users' sensitivity to temporal modulation is generally poor [24], the extracted modulator is more likely to be perceptible to them than the original harmonic.

As shown in [92] and [5], frequency shift is a linear operation that can avoid distortions in the resulting modulators; whereas the envelope extraction methods used in CIS-like strategies, e.g., the Hilbert envelope or full/half wave rectification followed by a lowpass filter, often incur nonlinear distortions in the extracted envelopes [60]. Consequently, harmonics cannot be as represented distortion-free in CIS-like strategies as in HSSE processing.

4.2.4 Electric Encoding

For electric stimulation, the extracted modulators need to be compressed to match the wide range of input acoustic levels to the narrow range of usable current levels. In our

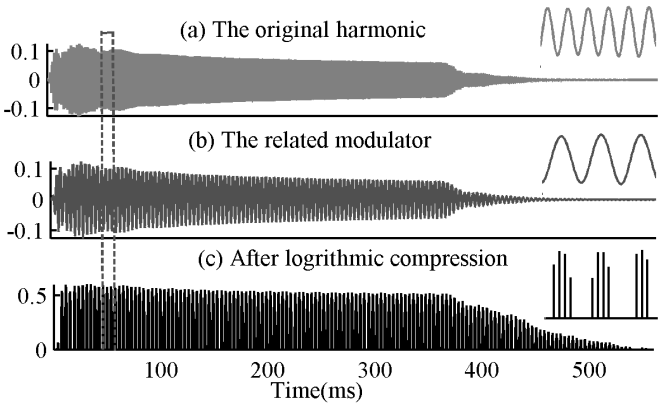

Figure 4.5: (a) Waveform of the 2nd harmonic of the guitar note shown in Figure 4.2. (b) Waveform of the HSSE modulator extracted from the 2nd harmonic. (c) Electric waveform of the log-compressed HSSE modulator. Each vertical line represents a pulse. The inset in each panel provides a detailed view of the respective waveform over a 12-ms window, as indicated by the rectangle overlaid on the waveforms.

experiments, each modulator was logarithmically compressed using a standard loudness growth function with 65 dB input dynamic range [129]. These log-compressed modulators were then delivered to their respective electrodes by constant-rate pulse trains. As in CIS encoding, the pulses for different electrodes were time interleaved so that no simultaneous stimulation occurred at any time.

As a visual example, Figure 4.5(a) displays the 2nd harmonic of the guitar note in Figure 4.2. The associated HSSE modulator is shown in Figure 4.5(b). One can see that the modulator resembles the original harmonic in terms of amplitude modulation, yet oscillates at a slower rate, as indicated by the waveform details. Figure 4.5(c) shows the log-compressed modulator that is carried by a 1900 Hz pulse train. Each vertical line represents one biphasic pulse. The presumed redundant negative values of the modulator were not encoded. Due to compression, the steep decay in harmonic amplitude at around 380 ms is converted into a moderate decrease in pulse amplitude.

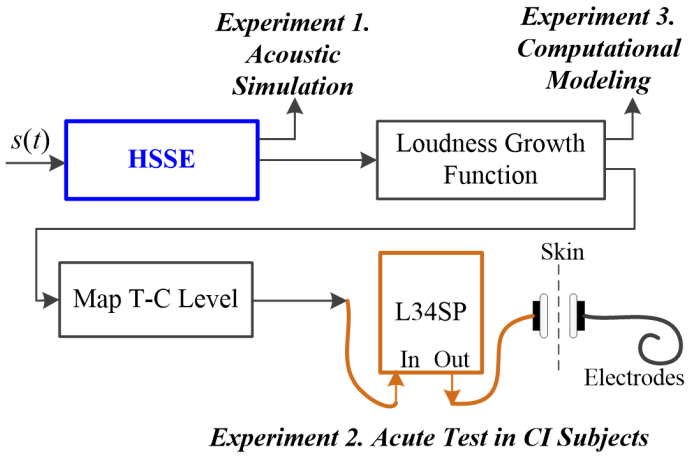

Figure 4.6: Systematic view of the experimental design. The "HSSE" block takes in an audio signal and outputs the extracted modulators. Each modulator is first log-compressed by the "Loudness Growth Function" block and converted into an electric pulse train, the amplitude of which is then mapped into appropriate current levels based on measured threshold (T) and most comfortable (C) loudness levels. The "L34SP" block stands for the Laura research processor of the Nucleus Implant Communicator 2 system, which is used to stream the HSSE pulse trains to a CI subject's electrodes.

4.3 Experimental Design

A systematic view of the experimental design is provided in Figure 4.6. In experiment 1, the effects of HSSE processing on melody and timbre recognition tasks were separately simulated, and then compared with that of the CIS encoding in normal-hearing listeners. The acoustic simulation captures the essence of CI signal processing; however, it does not include the response properties of the electrically-stimulated auditory nerve. In experiment 2, HSSE was acutely tested in CI users using the Nucleus Implant Communicator (NIC) 2 system: subjects' melody and timbre recognition performance with HSSE was measured and then compared with that achieved with their clinical processors. To quantify the temporal cues potentially captured by the auditory nerve, the neural spike patterns evoked by HSSE were simulated and then compared with that evoked by CIS in Experiment 3, using a population model of electrically-stimulated auditory nerve fibers.

4.3.1 Test Materials, Procedures, and Subjects

The melody and timbre recognition tests of the University of Washington Clinical Assessment of Music Perception were used [57]. The melody stimuli consist of 12 familiar songs, e.g., "Twinkle Twinkle Little Star". Each melody is a sequence of digitally synthesized harmonic complex tones. All rhythm cues were eliminated to force listeners to base the recognition of a melody on its F_0 contour. The timbre stimuli consist of sound recordings from 8 musical instruments, e.g., flute and violin. Each instrument played the same five-note sequence at the same loudness level, such that only timbre but not pitch or loudness cues contribute to the recognition of an instrument.

Each test was administered as follows: a participant first received a practice session in which he or she can listen to each test stimulus from a particular set twice. Then the actual test began. The participant was asked to recognize a presented item that was randomly chosen from the stimuli set. Each stimulus was presented three times in a random order. In the end, a score was calculated as the percent of melodies or instruments correctly identified. In all of the tests, participants were seated in a double-walled, sound-insulated booth. For normal hearing listeners, sounds were presented through a speaker at 68 dB SPL. The same speaker was used for CI subjects to measure their performance with the clinical processor. For the acute tests of HSSE, the speaker was not used since stimuli were delivered to CI users through the NIC 2 system after being encoded by HSSE. A program written in MATLAB (Mathworks, Natick, MA) played stimuli to participants and recorded their responses via a graphical user interface.

Five normal-hearing listeners participated in experiment 1. Eight CI subjects, all implanted with the Nucleus CI24 device, attended experiment 2. They are all native English speakers. The University of Washington Institutional Review Board approved the process of recruitment, testing and informed consent for each of them

4.3.2 Experiment 1: Acoustic Simulation

In order for normal-hearing listeners to compare HSSE and CIS, all the melody and timbre stimuli were vocoder-processed to simulate the effect of HSSE and CIS processing,

respectively. Specific stimulus generation details were the same as in [60]. A stimulus was bandpass filtered (3rd-order Butterworth) into 4 or 8 bands. The analysis bands spanned 80-6000 Hz and were logarithmically spaced. To simulate CIS processing, the Hilbert transform was applied in each channel and their Hilbert envelopes were extracted afterwards. In HSSE processing, the strongest harmonic in a particular channel was first identified and then transformed into a modulator, as described in Section 4.2.

To prevent vocoders from presenting information that might not be accessible to CI users, the extracted CIS envelopes and HSSE modulators were all lowpass filtered (3rd-order Butterworth) at 300 Hz to correspond to the pitch saturation limit in electric hearing [24, 108]. Because the F_0 of the test stimuli was between 262 and 523 Hz, in HSSE processing, the selected harmonics were frequency downshifted to half the F_0–as opposed to the F_0–such that the transposed harmonics can be preserved in the resulting 300-Hz-wide modulators. As a result, the F_0 contour conveyed in each HSSE modulator was one octave lower than the original one yet still consistent with it.

For synthesis, the extracted 300-Hz-wide CIS envelope or HSSE modulator from each channel was used to modulate a sinusoid carrier at the respective channel center frequency. The modulated sinusoids were then combined across all bands to generate the simulation sound that would be presented to normal-hearing subjects. In either the melody or the timbre recognition test, each subject was tested under four simulation conditions, covering two strategies (CIS and HSSE) and two channel numbers (4 and 8). The test order was randomized for each subject.

4.3.3 Experiment 2: Acute Test in Cochlear Implant Subjects

The recruited CI subjects are all fitted with the ACE strategy [51], which estimates the incoming signal spectrum using a 128-point Fast Fourier Transform (FFT) under a sampling rate of 16 kHz, providing an analysis frame of 8 ms and an FFT bin spacing of 125 Hz. The FFT bins are then grouped to produce a total of 22 channels, which are typically linearly spaced from 188 to 1312 Hz and then logarithmically spaced up to 7938 Hz. For each channel, the real and imaginary components of the corresponding FFT bins were summed

separately and their norm was computed afterwards to yield the envelope of the associated frame. In each frame, an "n" number of channels with the largest amplitude (typically 8-12 spectral maxima) are selected. The stimulation pulses are interleaved between the selected electrodes accordingly. The optimal number of maxima and the optimal stimulation rate are patient-specific information and saved in a patient's strategy MAP.

In HSSE processing, the same channel spacing of 22 electrodes as in the ACE strategy was used. Given an input signal, HSSE explicitly tracked its harmonics and then transformed 8 predominant harmonics into modulators. As in experiment 1, the extracted modulators were 300 Hz wide. Each modulator was logarithmically compressed and then converted into an electric pulse train, using the Nucleus MATLAB toolbox [129]. To faithfully represent temporal fine structure, the per channel pulse rate was set at 1900 Hz, providing a total stimulation rate of 15200 Hz. Each pulse was biphasic and 25-μs wide. The cathodic phase was applied first, followed by an 8-μs gap and an equal-amplitude anodic phase of the pulse. To deliver the HSSE pulse trains to a CI subject, the Nucleus Implant Communicator 2 system was used, which comprises the CI24RE receiver-stimulator with the Contour electrode array, the Laura L34SP research processor, and a custom fitting program. Before the test began, a subject's threshold (T level) and most comfortable (C level) loudness levels on each electrode were first measured in the fitting program. Based on the measured T-C levels, the amplitudes of HSSE pulse trains were mapped into appropriate current levels, which were then streamed to the subject's electrodes through the Laura processor without any further modification.

The ACE and CIS strategy were both implemented on the L34 processor, which produced roughly the same music perception as the ACE strategy on subjects' clinical processor. However, because CI subjects were already used to their processor, which has been specially optimized for them in terms of the stimulation rate and the number of spectral maxima, they would probably perform better with their own processors. Thus, subjects' clinical processor was chosen as the comparison baseline. As a visual comparison, the electrodograms by ACE and HSSE for a violin note are provided in Figure 4.7(b) and (c), respectively, along with the note's spectrogram in Figure 4.7(a). The spectrogram shows how the note's intensity (color scale) varies as a function of time and frequency: the evenly spaced horizontal lines

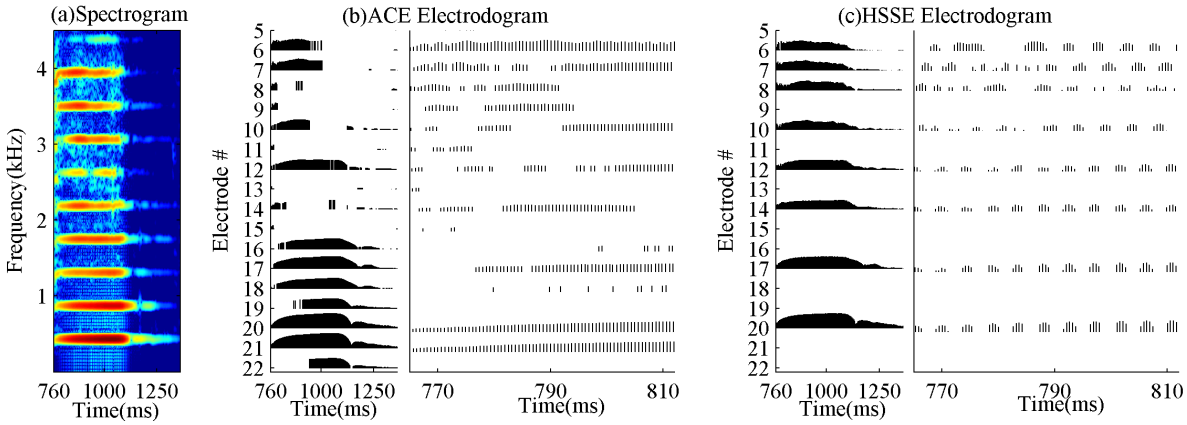

Figure 4.7: (a) Spectrogram of a violin note at 441 Hz. (b) Electrodogram of the note by ACE (8-of-22). (c) Electrodogram of the note by HSSE. A detailed view is presented in (b) and (c), respectively, providing an expanded view of each electrodogram over a 45-ms window.

represent harmonics, with the bottom one representing the F_0. An electrodogram displays how the amplitude of electric pulses, represented as vertical lines, varies as a function of time at each electrode. One can see that the active electrodes are varying in ACE processing, whereas the same 8 electrodes are selected through the note duration in HSSE to transmit 8 predominant harmonics. Despite this difference, the selected electrode subsets are highly overlapped between the two strategies, presumably because both strategies are devoted to represent the predominant components of the note. To show the difference in fine structure encoding, a detailed view of the beginning segment of each electrodogram is also provided. One can see that with HSSE, salient F_0 cues are present at lower-frequency electrodes and noise-like cues are conveyed at higher-frequency electrodes, consistent with the phase coherence feature shown in the beginning of the spectrogram. The same level of fine structure cues are not found in the ACE electrodogram. The distinction of HSSE is also recognizable in comparison with the previously proposed F0mod [147] strategy, which modulates all channel envelopes at the input signal's F_0 regardless of any noise-like features in high frequency. The eTone [132] strategy might be able to preserve noise-like features, but it does not track harmonics as HSSE does. For timbre perception, however, both the harmonic and the noise-like fine structure cues are important and are represented in HSSE.

4.3.4 Experiment 3: Computational Modeling

To examine whether the temporal cues in an electric pulse train can be translated into appropriate neural responses, an auditory nerve model [128, 62] was used to simulate the neural spike patterns evoked by CIS and HSSE, respectively. The model is based on the morphology and electrophysiology of cat spiral ganglion cells. Imennov and Rubinstein [62] demonstrated that the normalized response thresholds obtained from a population of 250 diameter-distributed model fibers match that of the same number of *in vivo* cat fibers. Thus, the same distribution of 250 fibers was used to generate all of the neural outputs in this experiment. The input to the model was the electric pulse train generated by a CIS or HSSE processor for a test stimulus.

Analogous to the vocoder processing, 8-channel CIS and HSSE implementations were used here to generate the electric encoding of the test stimulus. First, the 300-Hz-wide CIS envelope or HSSE modulator for a particular channel was extracted, as described in experiment 1. Next, the extracted signal was logarithmically compressed and converted into a pulse train of 1900 Hz, as described in Experiment 2. Each pulse train was then fed into the auditory nerve model to simulate, respectively, the neural spike patterns evoked by CIS and HSSE for the test stimulus.

Due to the high computational cost of simulating stochastic neural responses, "Twinkle Twinkle Little Star" was chosen as the only test melody in this experiment. It shows a large F_0 jump (from 262 to 392 Hz) at the first 1.5 seconds, thus further reduction in the computational cost can be achieved by simulating only the first 1.5 seconds. Because the auditory nerve model was inherently single-channel, neural responses in each spectral channel would be simulated independently. To gauge the best potential of CIS, the 4th band ([1065, 1675] Hz) was selected for simulation, because visual comparison indicated that better acoustic F_0 cues were conveyed in this band. For HSSE, there were multiple options, so the same 4th band was selected for consistency.

To examine the amount of F_0 information captured by the auditory nerve, a raster-plot of the simulated spike trains along 250 fibers was first generated and evaluated qualitatively (see Figure 4.10). The time intervals between successive spikes were then analyzed. Given

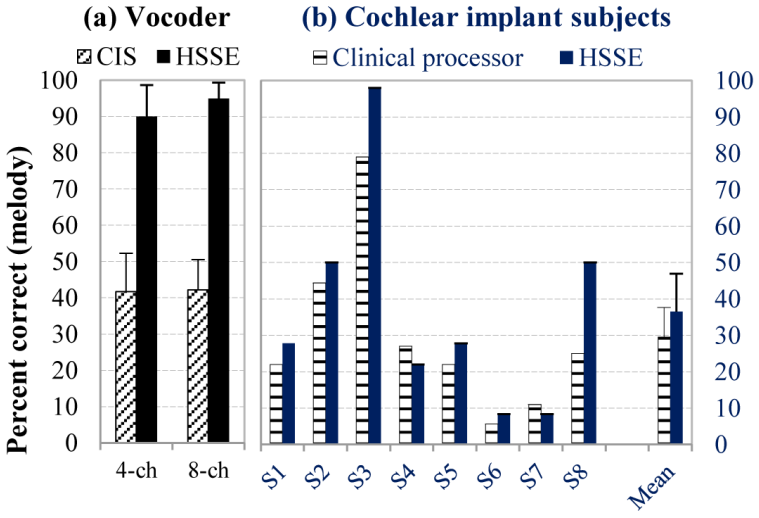

Figure 4.8: Results on melody recognition test. (a) Normal-hearing subjects' mean performance with 4- and 8-channel CIS (hatched) and HSSE (filled) vocoders. (b) Each individual cochlear implant subject's performance and their average. The hatched bars stand for their clinical processors and the filled bars correspond to the acutely tested HSSE. Error bars represent the standard error of the mean.

that the F_0 was encoded temporally with both strategies, an increase in a stimulus' F_0 should produce a decrease in the interspike intervals (ISIs), and vice versa. Therefore, by measuring how the ISI changes as a stimulus' F_0 evolves, a strategy's ability to encode the F_0 can be simulated.

4.4 Experiment Results

4.4.1 Acoustic Simulation Test Results

Figure 4.8(a) shows the average intra-subject melody recognition performance achieved by normal-hearing listeners with CIS and HSSE vocoders. In both channel conditions, subjects scored much higher with HSSE than with CIS: a mean improvement larger than 45% was observed. A repeated-measures analysis of variance (ANOVA) revealed that processing strategy produced a significant effect on melody recognition performance [$F(1, 4) = 50.1$, $p = 0.002$]. Yet within a strategy, no significant performance difference was found between 4- and 8-channel conditions ($p > 0.5$).

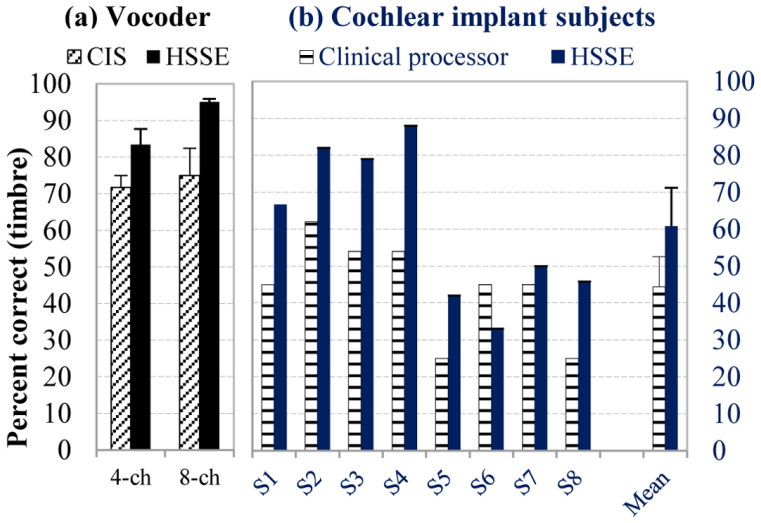

Figure 4.9: Results on timbre recognition test. (a) Mean performance by normal-hearing subjects with 4- and 8-channel CIS (hatched) and HSSE (filled) vocoders. (b) Each individual cochlear implant subject's performance and their average. The hatched bars stand for their clinical processors and the filled bars correspond to the acutely tested HSSE. Error bars represent the standard error of the mean.

Normal-hearing subjects' average performance on timbre recognition is shown in Figure 4.9(a). Unlike in the melody recognition test, subjects showed a slight improvement when the number of channels was increased from 4 to 8, presumably because the spectral shape of a timbre stimulus can be better represented with more channels. In the 4-channel condition, subjects' mean score was 12% higher with HSSE than with the CIS vocoder; and was 20% higher in the 8-channel condition. A repeated-measures ANOVA revealed that the processing strategy was a significant factor for timbre recognition [$F(1,\ 4) = 12.0,\ p = 0.026$], but the channel number was not ($p > 0.1$).

4.4.2 Cochlear Implant Test Results

Each individual CI subject's melody recognition performance as well as their mean score is shown in Figure 4.8(b). With the acutely tested HSSE, some subjects (e.g., S3 and S8) showed a clear improvement, about 20%; while in other subjects, only a minor or no advantage of HSSE was observed. On average, CI subjects scored 30% with their clini-

cal processor and 37% with the acutely tested HSSE. The current data did not imply a significant difference between the two strategies on melody recognition ($p > 0.090$).

Figure 4.9(b) displays each individual CI subjects performance on timbre recognition along with their mean score. With the acutely tested HSSE, most subjects showed an immediate improvement: the biggest improvement observed was 35% (e.g., S4). On average, CI subjects scored 44% with their clinical processor and 61% with the acutely tested HSSE. The current data demonstrated a significant effect of HSSE on timbre recognition [paired $t(7) = 3.3$, $p < 0.013$].

4.4.3 Computational Modeling Results

1) Raster-plots: Figure 4.10 is divided into two subpanels. The CIS-evoked neural responses are displayed in panel A and those evoked by HSSE are in panel B. Within each panel, the top row shows the input pulse train, with each vertical line representing one biphasic pulse; the bottom row shows the raster-plot of the simulated neural outputs by placing a dot for every occurrence of a spike. To clearly display the difference in interspike intervals, a break in the x-axis is introduced, such that the first 60 ms of note 2 (262 Hz) and that of note 3 (392 Hz) can be compared next to each other. In each raster-plot, the diameter of the model fibers is displayed in the y-axis. It is noteworthy that the model responses resemble the spiking behavior observed *in vivo*: e.g., large-diameter fibers exhibit lower firing thresholds and a decreased variability in their response; compared to small-diameter fibers, the absolute refractory period in large fibers is brief, allowing them to fire spikes with greater frequency.

Comparing the CIS- and the HSSE-evoked spike trains for an identical note, one can see that they convey similar envelope but different timing cues. The HSSE-evoked spike train displays clear troughs and peaks, following the F_0 fluctuation in the HSSE pulse train; whereas such a timing pattern is missing in the CIS-evoked spike train. Although the F_0 fluctuation is visible in the CIS pulse train, the modulation depth is comparatively shallow. Using a constant pulse rate of 1900 Hz, the CIS stimulation causes saturation in large-diameter fibers ($> 3.6\mu$m), forcing most of the F_0 cues to reside only in small-diameter

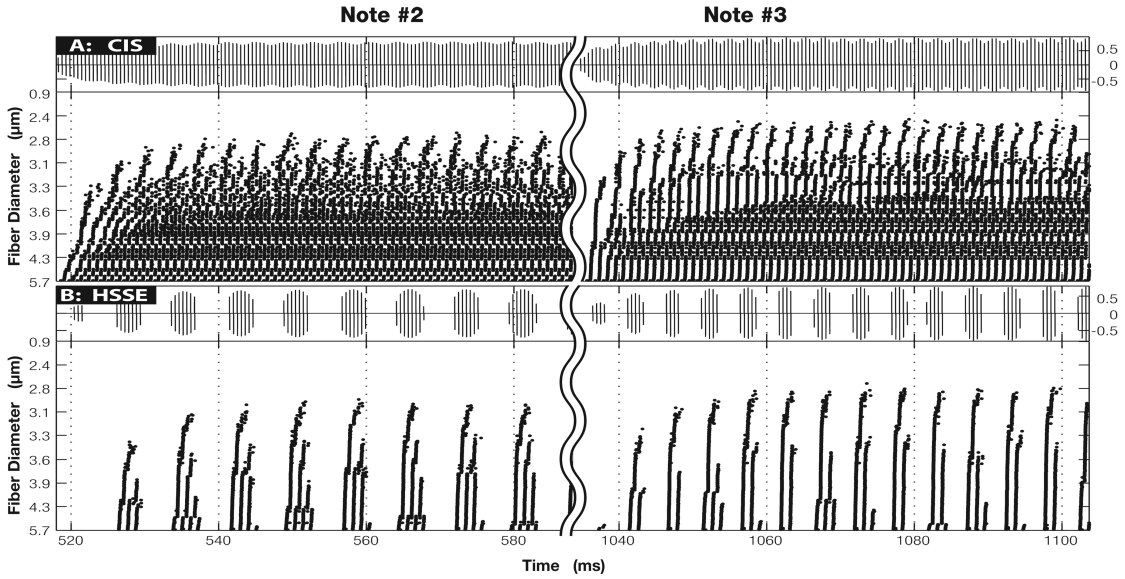

Figure 4.10: Panel A: CIS pulse train (top) and the evoked neural responses (bottom). Panel B: HSSE pulse train (top) and the evoked neural responses (bottom). The underlying stimulus is a melody, "Twinkle Twinkle Little Star". From note 2 to note 3, the F_0 increases from 262 to 392 Hz

fibers ($< 3.6\mu$m). In contrast, HSSE might encode F_0 in two ways: in the duration of a pulse burst and in the interval between successive bursts. As F_0 increases from note 2 to note 3, both the width of the spike bursts and the interval between successive bursts decrease noticeably, potentially providing important timing cues for F_0 discrimination.

2) Interspike-Interval (ISI) Analysis: to evaluate the differences in timing cues available to CIS and HSSE listeners, the histograms of ISIs pooled from all model fibers are plotted in Figure 4.11. The histograms on the left are constructed from the CIS-evoked spike trains, and those on the right are from HSSE. In each histogram, there is a peak between 1.0–1.5 ms, which is related to the fibers' refractory period. The absolute refractory period of a single fiber was $\sim$750 μs [62], imposing a minimum ISI. The relative refractory period can further extend the ISI. Given a stimulation cycle of 526 μs (=1/1900 Hz), the elicited ISIs were all longer than 1.0 ms, producing a peak between 1.0–1.5 ms.

Additionally, there is also an F_0-related peak in each histogram, as indicated by the dashed vertical lines. With HSSE, because the harmonics were transposed to half the F_0,

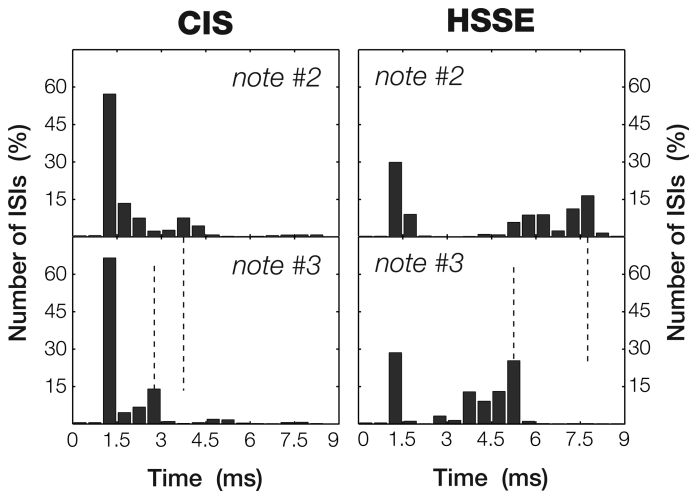

Figure 4.11: Histograms of interspike-intervals pooled from the spike trains shown in Figure 4.10.

the pitch intervals were doubled accordingly. Comparing the top and the bottom rows, one can see that the F_0-related peak shifts to the left, towards shorter ISIs, as F_0 increases from note 2 to note 3. Qualitatively, this shift is already implied in Figure 4.10, which also shows that fewer F_0 cues are present in the CIS- than in the HSSE-evoked spike trains. This difference in available timing cues is manifested as a difference in F_0-related peak height. As can be seen, the F_0-related peak height is much smaller on the CIS side than on the HSSE side. While one might perceive the F_0 change with either strategy, the simulated responses suggest that the change is likely to be much more obvious to HSSE listeners.

The ISI analysis is related to the temporal pitch coding in a single channel. Since the input stimulus used in our simulation was a harmonic tone, the F0mod or the eTone strategy, if included in the experiment, would likely generate similar F_0 patterns as HSSE. However, given a stimulus with noise-like features, there would be a clear difference between HSSE and the F0mod strategy in terms of the electric pulse pattern, and hence the neural spike pattern, because HSSE is more capable of encoding noise-like fine structure cues. With the eTone strategy, the representation of noise-like features would be dependent on the accuracy of the harmonic probability estimation [132].

4.5 Discussion

4.5.1 F_0 Coding with HSSE

To encode the F_0 of a complex sound, HSSE explicitly tracks its harmonics and transforms them into modulators centered at the F_0. Because most CI users can only perceive temporal information up to approximately 300 Hz [24, 108], the harmonics were transposed to half the F_0 in our experiments. Compared with the envelope-based CIS strategy, HSSE represents better the temporal F_0 coding mechanism in the normal auditory system [141, 117], and yielded markedly better performance on the simulated melody recogntion test. The fact that little difference between 4- and 8-channel conditions was found for melody recognition suggests that with a small number of channels, the place cues were too coarse to assist F_0 discrimination. The advantage of HSSE in F_0 coding was also demonstrated by the simulated spike patterns. With CIS, a limited amount of timing cues were conveyed only in the small-diameter fibers, while the large-diameter fibers were mostly saturated. In contrast, clear F_0-related timing cues were observed in the HSSE-evoked spike patterns, suggesting that a better temporal pitch code might be available to HSSE listeners.

In the eight recruited CI users, a broad range of melody recognition performance was observed; yet their mean score with the clinical processor was consistent with previously reported data [138, 57]. With the acutely tested HSSE, some subjects showed an immediate clear improvement, suggesting that the F_0 cues in HSSE modulators were accessible and beneficial to them. On the other hand, many subjects performed as poorly with HSSE as with their clinical processors. There are several possible reasons why some subjects did not show improvement on melody recognition with HSSE. First, there is often a mismatch between temporal and place cues, e.g., a signal with a repetition rate of 500 Hz might be delivered to a place that would normally be tuned to 1000 Hz. Without a correspondence between place and temporal cues, the auditory system may not be able to extract accurate information about frequency. As shown in [149], the correct tonotopic representation of frequency is a necessary ingredient in the neural code for pitch. Second, in CI users there is often partial degeneration of the auditory nerve. This may adversely affect performance simply because there are fewer neurons carrying the temporal information, regardless of its

availability. Third, the melody recognition task is not simply a measure of CI users' pitch perception [135], in that it requires them to recognize familiar melodies with potentially distorted pitch cues and no rhythm cues. Perception of exact pitch intervals is critical to accurate melody recognition, yet CI users may actually perceive mistuned pitch. This might be why subjects S1, S2, and S4 did not recognize melody clearly better with HSSE, but all showed great improvement (>20%) in timbre recognition, where exact pitch intervals were not required.

4.5.2 Timbre Perception with HSSE

On timbre perception, both normal-hearing and CI subjects demonstrated a significant improvement with HSSE over CIS-like strategies. Among the major determinants of timbre, spectral shape was presumably represented by HSSE and CIS with a similar degree of accuracy, because fixed spectral constraints apply to both strategies. The attack time was conveyed in the amplitude modulation, thus it was probably equally accessible to both HSSE and CIS listeners. Therefore, the observed benefit of HSSE on timbre recognition is likely attributable to the encoded harmonic and temporal fine structure cues.

As found in previous studies [144, 145], the relative amplitude of harmonics as well as the relative timing of their onsets and offsets are important timbre cues. In HSSE, the harmonics of complex sounds were explicitly tracked and transformed into modulators for electrodes. Although only the predominant harmonics were encoded due to the limited number of effective channels in CI users, current results suggest that the provided subset of harmonics benefit timbre perception significantly. In contrast, with the CIS and ACE strategy (as well as other recently proposed strategies, e.g., [113, 111, 146, 147, 132]), the representation of harmonics is presumable insufficient, given that these strategies ignore the inherent harmonic structure during encoding. Additionally, the tone- or noise-like quality of an instrument timbre is dependent on the phase coherence cues of upper harmonics [143, 144]. These cues are conveyed in the temporal fine structure, which was largely discarded in CIS and ACE processing. In HSSE, however, the fast-varying temporal fine structure was transformed into low frequencies and preserved in the modulators. As shown

in Figure 4.7, the HSSE-encoded temporal fine structure cues can be either noise-like or F_0-related, reflecting the phase coherence attribute of a particular timbre. It is possible that both the harmonic and the temporal fine structure information assisted HSSE listeners in recognizing a specific instrument timbre.

4.5.3 Implementation of HSSE

In acute testing, CI users only received a limited exposure to HSSE, thus, they might not be able to take full advantage of the information in the pulse pattern. Given more experience with HSSE, they could potentially perform better on melody and timbre recognition tests. For longer duration testing, HSSE may be implemented in real time, given an efficient F_0 tracker. To avoid significant perceptual delay, the F_0 tracker should be able to run with minimal output latency using only a fraction of a processor's computing resource. Also, it should be able to generate accurate F_0 estimates in a wide range of conditions. The feasibility of real-time F_0 tracking has been demonstrated in interactive music applications [150] and in CI speech processing [132]. Once the F_0 is known, the encoding of each individual harmonic can be executed in parallel, for which the computational complexity is comparable to traditional encoding approaches. Given that processors' computing power has been increasing over time, it seems feasible to implement real time HSSE on a modern processor.

Since the test stimuli in the current study contain only a single musical source, it is not clear yet how HSSE can assist CI users in a multi-instrument setting. When multiple sources are present, it is possible to track the F_0 of every single source. Yet how to simultaneously deliver them to CI users is beyond the scope of this work. Alternatively, we can select one particular source–e.g., by tracking the root (lowest) note of a chord, or the predominant F_0 [151] of a leading instrument or singer–and deliver the associated harmonic group to CI users by HSSE. Although it seems less ideal to represent a single source than to deliver all sources, HSSE is still a useful step forward in understanding how to bring an improved musical experience to CI users.

4.6 Conclusion

Although the importance of fine structure information to music perception is well acknowledged, the optimal representation of fine structure information for CI users remains unresolved. The present study has demonstrated that by encoding harmonic and temporal fine structure information, HSSE has possible advantages over CIS-like strategies in delivering timbre cues to CI users. Regarding melodic pitch contour identification, the extent to which HSSE can provide benefit is unclear. However, the vocoder test results suggest that salient and accurate pitch cues are available in HSSE encoding, and these pitch cues can be translated into simulated neural responses. More experience with HSSE could potentially yield higher performance on the melody test. Overall, HSSE seems like a promising strategy to improve music perception with CIs. Improved music perception could have a highly positive impact on CI users' lives, particularly to those deaf at birth or early in life with no previous experience with music [58].

Chapter 5

ON THE IMPORTANCE OF TEMPORAL FINE STRUCTURE TO SPEECH INTELLIGIBILITY IN THE PRESENCE OF NOISE AND SINGLE-TALKER MASKER

5.1 Introduction

By virtue of cochlear frequency analysis, complex broadband sounds (such as speech) are decomposed into a series of relatively narrowband signals, each of which conveys information about the sound by time-varying features. The slow changes in the overall amplitude is called envelope (E), while the more rapid events, such as zero crossings, constitute temporal fine structure (TFS). Fundamental questions about the relative importance of E and TFS to speech perception have been addressed by many studies (e.g., [46, 48, 19, 90, 6]). The conventional methods for speech manipulation use Hilbert transform to represent each subband signal as the product of two components: the Hilbert envelope or amplitude-modulation (which we call "Hilbert AM") function, and a frequency-modulated (which we call "Hilbert FM") sine wave carrier. . It is generally assumed that the decomposition of Hilbert AM and FM provides an appropriate separation of E and TFS, which is, as we will discuss below from a careful engineering standpoint of signal decomposition, not necessarily true.

Mathematically, Hilbert AM and FM are related, rather than independent. Voelcker [83] and Logan [85] showed that certain classes of bandpass signals are determined uniquely (to within a scale factor) by their zero crossings or phase characteristics. This means that one can recover E cues based on Hilbert FM. Given a bandpass signal, by band-limiting its AM component but keeping the FM part unchanged, Ghitza [86] showed that the evoked E cues at the output of narrowband auditory filters were actually not as band-limited as intended. Zeng et al. [87] demonstrated that with a small number of channels, the E cues can be largely recovered from Hilbert FM at the output of simulated auditory filters. Gilbert and Lorenzi [88] argued that when eight or more channels were allocated, i.e., when the analysis

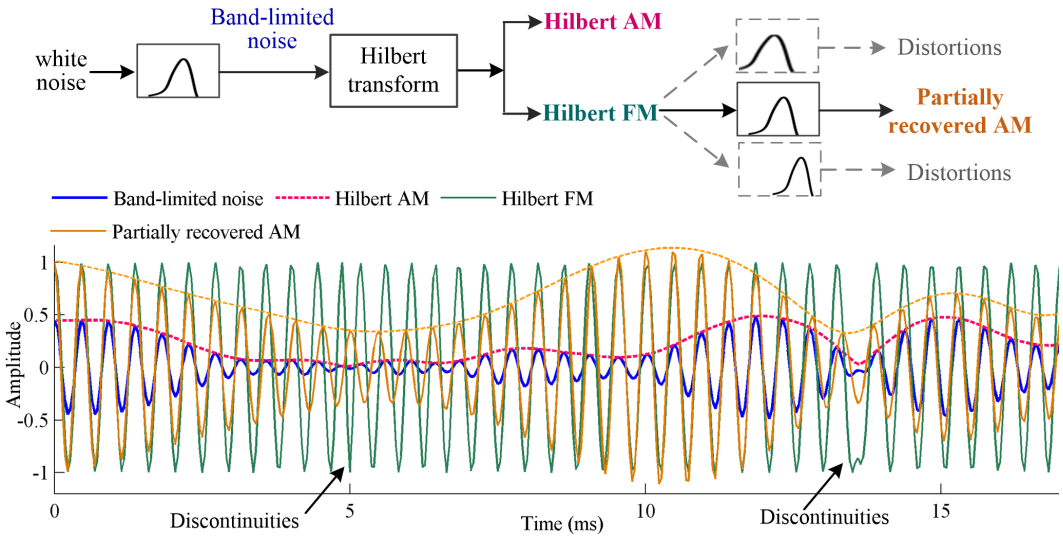

Figure 5.1: Illustration about why Hilbert FM is a flawed representation of temporal fine structure.

bandwidths were no broader than four times the bandwidths of normal auditory filters, the contribution of the synthetically recovered E cues to speech recognition were "essentially abolished".

Besides the coupling between Hilbert AM and FM, another critical issue is that Hilbert FM is a flawed representation of TFS. To demonstrate, we bandpass filter white noise to lie between 2 to 2.3 kHz to generate a band-limited noise signal, which is displayed in Figure 5.1. The corresponding Hilbert AM and FM components are overlaid on the signal waveform. As can be seen, there are discontinuities in Hilbert FM whenever the AM function approaches zero. Although the original signal is perfectly band-limited, the FM component is actually band-unlimited, i.e., it has infinite bandwidth [67]. If we pass the FM component through the original bandpass filter to reband-limit it, its amplitude becomes no longer flat, but fluctuates following roughly the original AM function. This confirms the observation by previous studies that at the output of simulated auditory filters, E cues can be partially recovered from Hilbert FM. Besides, the contents in Hilbert FM will leak into other adjacent channels and introduce distortions. It is problematic that the information extracted from one channel can actually affect other channels. Additionally, there are intermodulation

distortions within the underlying channel as well [152], which we will show in Figure 5.3. However, the discontinuities or distortions associated with an careful engineering view of Hilbert FM are largely ignored in previous studies.

Given that Hilbert FM can lead to E recovery, there have been controversies regarding the relative contribution of TFS to speech perception. It is crucial to study the relative importance of E and TFS in neural coding, because the transformation from acoustic cues to neural responses is complicated and the information conveyed in neural coding potentially has direct relevance to perception. Heinz and Swaminathan [19] analyzed the neural spike trains evoked respectively by Hilbert AM and FM, and showed that TFS contributes to speech perception by recovering E cues. They found that neural E coding was always a primary contributor to speech perception even in the presence of noise, while neural TFS coding had less perceptual salience than previously thought [90]. Recently, Shamma and Lorenzi [6] proposed a simplified peripheral auditory model to quantity neural E and TFS coding. They compared the neural representations evoked respectively by Hilbert AM and FM, and found that neural E and TFS coding contribute comparably to speech intelligibility regardless of speech processing conditions

Interpretation about the results of previous studies is often linked to the fact that Hilbert AM and FM are related. But the issue is ignored that Hilbert FM is a distorted representation of TFS. The discontinuities in Hilbert FM cause distortions when being processed by narrowband auditory filters, which presumably confounded previous studies about the importance of TFS. Following and extending Shamma and Lorenzi [6], the upcoming discussion and results explore the relative importance of neural E and TFS coding to speech intelligibility in different noise conditions. For example, Shamma and Lorenzi [6] only used high SNR speech and they also used only a conventional Hilbert AM and FM vocoding decomposition. To address the concerns about the conventional Hilbert AM and FM decomposition, we proposed a new slow envelope and fast envelope decomposition, using half wave rectification followed by filters reflecting engineering interpretation of neural physiology. The slow envelope represents the temporal cues that can be conveyed in neural responses despite phase locking degradation, while the fast envelope represents the temporal cues whose neural representation critically relies on phase locking. The signal processing

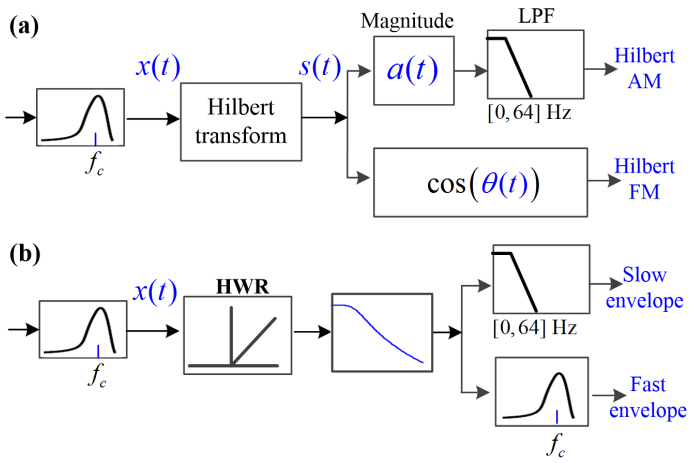

Figure 5.2: Block diagram of signal processing for (a) conventional Hilbert decomposition, and (b) new slow envelope and fast envelope decomposition. HWR is short for half wave rectification.

details are described in next section. We hypothesized that by using fast envelope instead of Hilbert FM as the representation of TFS, the relative contribution of neural TFS coding to speech intelligibility would appear significantly larger. The slow envelope is, other than its method of calculation, identical to conventional Hilbert AM.

5.2 Methods

Following previous studies, the vocoder processing parameters used here were adapted from Gilbert and Lorenzi [88]. Given an input stimulus, it was first bandpass filtered (Butterworth, 36 dB/oct rolloff) into 16 adjacent channels spanning the range 80 8020 Hz. The 16-channel condition was chosen because it was discussed in earlier studies. Each band was then processed in two different ways, as shown in Figure 5.2.

5.2.1 Conventional Hilbert AM and FM decomposition

The signal processing for Hilbert decomposition is depicted in Figure 5.2(a). Given a bandpass signal $x(t)$, its Hilbert transform pair was first obtained and denoted as $\tilde{x}(t)$. Then a complex-valued analytic signal, $s(t)$, was constructed, which can be decomposed into a non-negative envelope function $a(t)$, and an instantaneous phase function $\theta(t)$ as follows:

$$s(t) = x(t) + j\tilde{x}(t) = a(t)\exp\left(j\theta(t)\right) \quad (5.1)$$

$$\text{where } a(t) = \sqrt{x^2(t) + \tilde{x}^2(t)} \quad (5.2)$$

$$\theta(t) = \tan^{-1}\left[\tilde{x}_n(t)/x_n(t)\right] \quad (5.3)$$

Based on its analytic representation, the bandpass signal can be expressed as the product of the Hilbert envelope and a sinusoidal FM carrier:

$$x(t) = Re\left\{s(t)\right\} = a(t)\cos\left(\theta(t)\right) \quad (5.4)$$

As in previous studies, the envelope function $a(t)$ was next lowpass filtered (Butterworth, 36 dB/oct) at 64 Hz to generate the Hilbert AM component for the following modeling experiments; while the sinusoidal carrier $\cos\left(\theta(t)\right)$ was taken as the FM component without any further modification.

To illustrate, Figure 5.3 shows three amplitude modulation signals generated by two beating tones at 900 and 1000 Hz, respectively. From top to bottom, the relative strength between the two tones varies systematically. At each row, the signal spectrum is displayed on the left. The spectra of the corresponding Hilbert AM and FM components are shown in the 2nd and 3rd column, respectively. Their waveforms are depicted on the right.

First, the two signals in the top and bottom rows cannot be differentiated in terms of their envelopes. Their difference lies only in the fast oscillation structure. For example, in one fundamental period, the top signal exhibits 9 cycles, while the bottom signal exhibits 10 cycles. The fine structure seems to be dominated by the stronger component. Second, in the middle row, Hilbert FM has infinite bandwidth; it shows discontinuities whenever the envelope approaches zero. Third, by comparing each Hilbert FM spectrum with the original signal spectrum, we can see that intermodulation distortions are automatically produced within the band, due to the fact that the derivation of Hilbert FM is based on nonlinear operations [152].

According to Voelcker [83] and Logan [85], for these signals, one cannot recover the envelope based on the zero crossings, or reconstruct the zero crossings from the envelope(Voelcker, 1966; Logan, 1977). Given that envelope and fine structure convey very

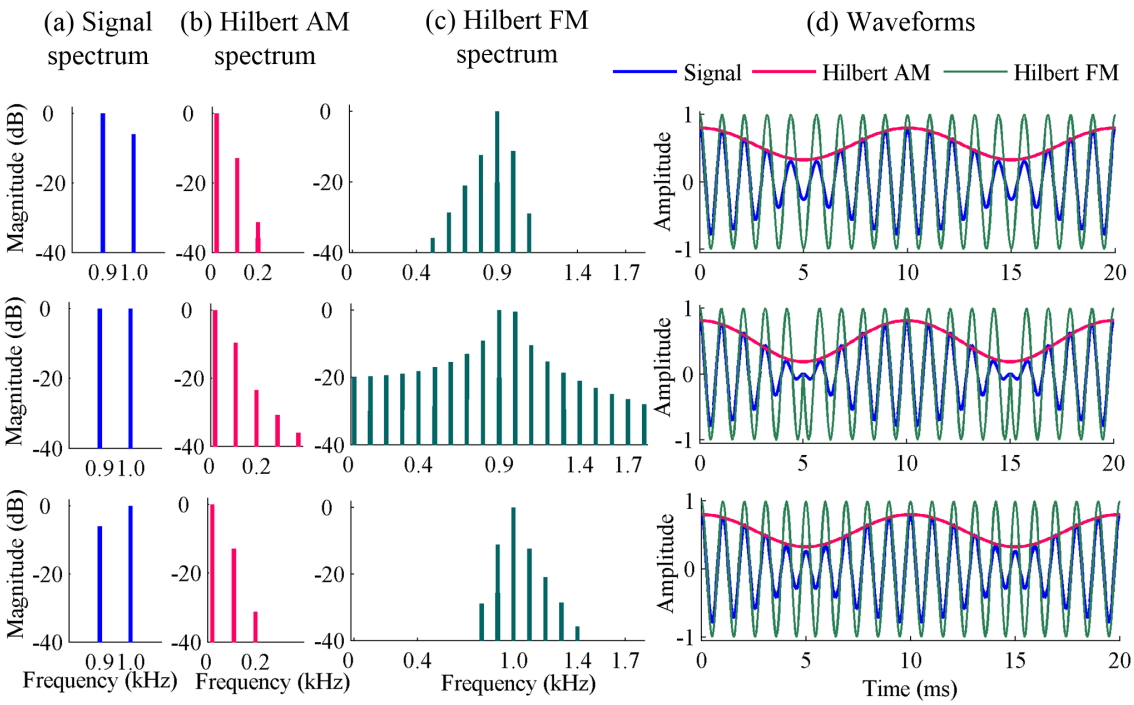

Figure 5.3: From top to bottom, each row corresponds to one amplitude modulation signal generated by two beating tones at 900 and 1000 Hz, respectively. The signal spectrum is displayed in the 1st column. The spectra of the corresponding Hilbert AM and FM are shown in the 2nd and 3rd column, respectively. Their waveforms are depicted in the 4th column. Each Hilbert AM waveform was obtained with a 150-Hz lowpass filter to preserve the modulation rate at 100 Hz.

different information about the signal, they should presumably contribute differently in neural coding. To investigate their relative roles, the definition of TFS poses a key issue. Conventionally, it is assumed that Hilbert FM is a faithful representation of TFS. However, the above analyses suggest that Hilbert FM is actually a distorted representation. Using Hilbert AM and FM decomposition, previous modeling studies found that envelope and TFS contribute similarly in neural coding [6]. Presumably, the distortions in Hilbert FM (e.g., across channel content leakage and within channel intermodulation distortions) confounded the contribution of TFS in previous studies.

5.2.2 New slow envelope and fast envelope decomposition

Figure 5.2 (b) shows the signal processing for a new approach, a slow envelope and fast envelope decomposition. Given a bandpass signal, it is first half wave rectified and then lowpass filtered. Details are described in the below. The available information was then analyzed into two types of temporal cues.

Signal spectrum analysis after half wave rectification (HWR)

Because rectification is a nonlinear operation, for an arbitrary input signal $x(t)$, there is typically no closed-form representation of the signal spectrum after rectification. However, we can gain some insight by expressing the half wave rectified signal, $x_+(t)$, as follows:

$$x_+(t) = \begin{cases} x(t) & \text{if } x(t) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$x_+(t) = \frac{x(t) + \sqrt{x^2(t)}}{2} \tag{5.5}$$

The half wave rectified signal is equivalent to the combination of a scaled copy of the original signal and a quadratic representation of itself. To illustrate, for the three signals analyzed in Figure 5.3, their spectra after half wave rectification are shown in the 2nd column of Figure 5.4. As expected, a scaled copy of the original signal spectrum is detected. Additionally, there are strong components in the low frequency and double frequency regions, which are generated by the square operation. The spectral peak at 100 Hz reflects the amplitude modulation rate at 100 Hz. It should be noted that due to the square root operation, harmonics of signal components are automatically produced.

Slow envelope and fast envelope decomposition

Physiologically, all the temporal cues will be first represented in the voltage variations of inner-hair cells, and later coded in the discharge patterns of auditory-nerve filters [140]. Given that the cell voltage cannot change arbitrarily fast, and neural spikes can only phase lock to stimulus waveforms up to a certain frequency, a lowpass filter was employed after

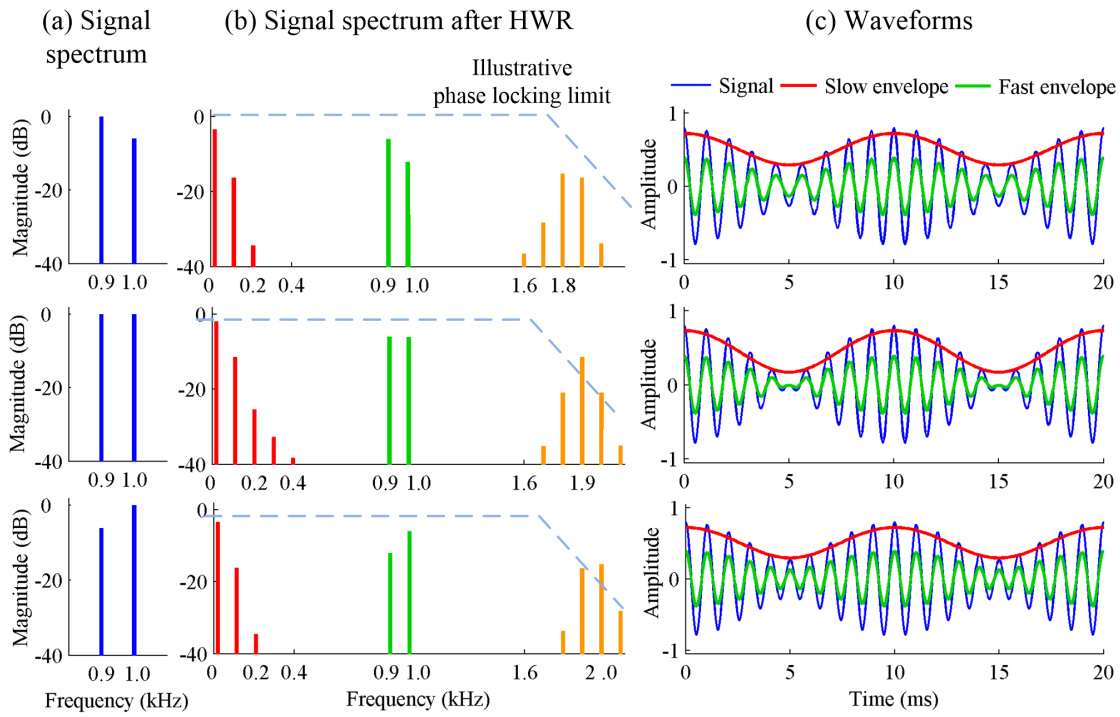

Figure 5.4: The amplitude modulation signals shown in Figure 5.3 are duplicated here following the original order. The signal spectrum is displayed in the 1st column. The signal spectrum after half wave rectification is shown in the 2nd column. The signal waveform is depicted in the 3rd column, overlaid by the corresponding slow envelope and fast envelope waveforms. The slow envelope was obtained with a 150-Hz lowpass filter to preserve the modulation rate at 100 Hz.

half wave rectification (HWR) to reflect these considerations. The cutoff frequency of this lowpass filter is a free parameter, which should be dependent on listeners' temporal coding abilities. In the following modeling experiments, a Butterworth lowpass filter with 1 kHz cutoff frequency and 12 dB/oct rolloff was used.

Up to the phase locking limit, neural spikes can synchronize with the individual cycles of the input stimulus waveform. As a result, TFS is coded in the phase locking patterns of neural responses [13]. On the other hand, the E cues can be conveyed in the average firing rate of auditory-nerve fibers, even in the presence of phase locking degradation [1]. Thus, one fundamental difference between TFS and E is that the former critically relies on phase locking to be represented, while the latter does not. Based on these physiological reasons,

we use the proposed fast envelope to capture the temporal cues that will be represented in neural phase locking patterns, and use the slow envelope to reflect the temporal cues that can be coded in average firing rates.

Specifically, to extract slow envelope, a lowpass filter was applied to the half wave rectified signal. The effect of a lowpass filter is equivalent to a temporal averaging operation. Based on equation (5.5), the slow envelope is essentially generated by a square operation:

$$\sqrt{x^2(t)} = \sqrt{a^2(t)\cos^2(\theta(t))} = a(t)\sqrt{\frac{1+\cos(2\theta(t))}{2}} \tag{5.6}$$

A lowpass filter will keep only the AM information but remove the high frequency oscillations. Mathematically, the above operation is equivalent to the derivation for Hilbert AM. In the following experiments, the lowpass filter was chosen to be identical to that used in Hilbert AM extraction. Thus, slow envelope is identical to the Hilbert AM except a scale difference.

To generate the fast envelope, a bandpass filter was used to extract the scaled copy of the original signal spectrum. For simplicity, the bandpass filter was chosen to be the same as the analysis filter of the underlying channel. Consequently, for the channels centered above phase locking limit (1 kHz assumed), the fast envelope would gradually decline in energy until it vanishes. Yet the slow envelope will not vanish for these high frequency channels.

For the three example signals shown on the left of Figure 5.4, the extracted slow envelope and fast envelope are displayed on the right, overlaid on the respective signal waveform. Provided that the signal frequency is within the phase locking limit, the extracted fast envelope seems like a scaled copy of the original signal.

5.2.3 Comparison between the two decomposition methods

As shown above, the extracted 64-Hz-wide Hilbert AM and slow envelope are mathematically equivalent. This can be confirmed by visually comparing their spectra shown in Figure 5.3 and 5.4, respectively. However, the Hilbert FM and fast envelope are fundamentally different. When being processed by the narrowband auditory filters, Hilbert FM can cause distortions by leaking into irrelevant channels. Additionally, the within channel

intermodulation distortions in Hilbert FM can potentially affect the phase locked responses. In contrast, the fast envelope does not contain distortions; it conceptually corresponds to the temporal cues that will be represented in the neural phase locking patterns within the underlying channel.

For vocoder synthesis, each Hilbert AM component was used to modulate a bandpass noise carrier centering at the respective channel. The modulated carriers were then combined together to generate E-speech. To produce TFS-speech, each Hilbert FM component was multiplied by a constant equal to the root-mean-square (RMS) power of the related band and then added together. Technical details about stimulus generation are given in Gilbert and Lorenzi [88]. Like in previous studies, the forward-backward filtering was used here to achieve zero phase delay.

In contrast, for the proposed new decomposition, slow envelopes were used to synthesize E-speech by a noise vocoder. Fast envelopes were combined together to generate TFS-speech without any further manipulation. Overall, for each input stimulus, there were four types of simulation sounds. For each type of simulation sounds, the relative effectiveness of neural E and TFS coding were evaluated using a simplified auditory model by Shamma and Lorenzi [6]. The global RMS power of each stimulus was equalized to eliminate level variations in the following experiments.

5.2.4 Neural coding simulated by a simplified auditory model

To explore how acoustic cues are reflected in the evoked neural responses, Shamma and Lorenzi [6] proposed a simplified biologically plausible auditory model, which takes in a sound stimulus and outputs a neural representation of the stimulus' spectrotemporal modulations, called an "auditory-spectrogram". As shown in Figure 5.5, their model begins with a "cochlear filterbank" of 128 highly overlapping narrowband filters. Each filter output is then fed into a function module representing the transduction nonlinearity of inner-hair cells. Next, the output of the inner-hair cell module is fed into a lateral inhibitory network [3, 153, 154], which sharpens the cochlear responses by a spatial derivative operation and then properly smooths the interim output to generate a final neural representation, i.e., the

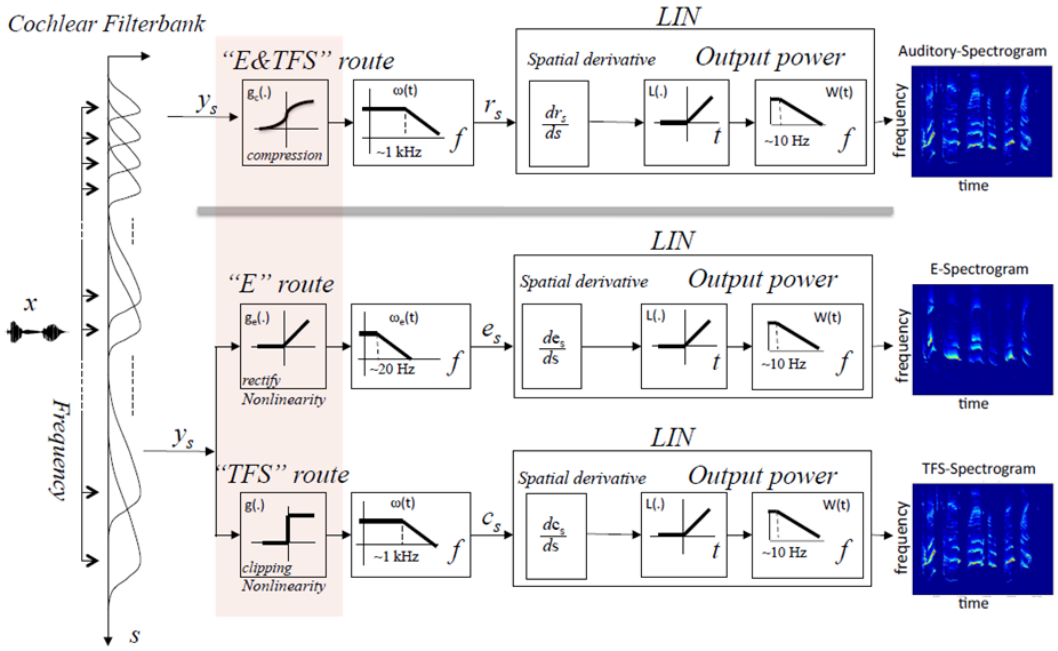

Figure 5.5: Conceptual diagram of E and TFS cues in the early stages of auditory pathway. The top E&TFS route depicts the "normal" processing in which the inner-hair cell applies a compressive nonlinearity followed by membrane lowpass filtering that gradually attenuates phase locked responses. In the bottom pathways, the hair cell nonlinearity is modified to highlight the encoding due to two extremes: an "E" route and a "TFS" route. Figure adopted from Shamma and Lorenzi [6] with few modifications.

auditory-spectrogram.

In normal auditory-nerve responses, the neural E and TFS cues are completely intermingled. However, to separate their relative effectiveness, Shamma and Lorenzi [6] constructed two extreme idealizations of the responses, referred to as "E route" and "TFS route" by manipulating the inner-hair cell module in different ways. To generate the normal auditory-spectrogram, the inner-hair cell module was realized by a sigmoid function that imposes nonlinear compression and a finite dynamic range on the E cues, and a lowpass filter that gradually decreases TFS cues from 1 to 6 kHz. By contrast, to construct the idealized "E-spectrogram", the inner-hair cell module was changed to HWR that allows infinite dynamic range, and a 20-Hz lowpass filter that largely removes TFS cues. For "TFS-spectrogram", the sigmoid compression was replaced by an infinite clipping operation that allows only

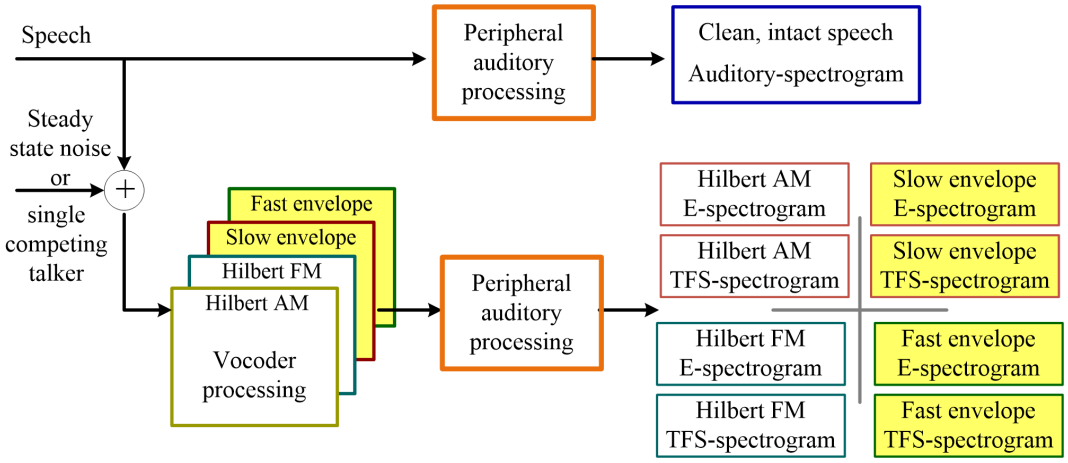

Figure 5.6: Experimental setup. Details about the four "vocoder processing" blocks are described in the previous section. The blocks in yellow represent the proposed new decomposition. The "auditory processing" block was adopted from Shamma and Lorenzi [6].

1-bit dynamic range but preserves zero crossings. It should be kept in mind that the E-spectrogram and TFS-spectrogram do not reflect the normal auditory processing; however, they can be regarded as two extremes of neural coding exploring the tradeoff between dynamic range and bandwidth—two critical and conflicting factors in signal sampling and reconstruction [155].

5.2.5 Simulation procedure and stimuli

To investigate the perceptual benefit of TFS, speech intelligibility is often measured in various noise conditions. In this study, the auditory-spectrogram is regarded as the neural basis of speech intelligibility, because it conveys all the spectrotemporal modulations relevant to speech recognition. Therefore, by comparing the degree of similarity between E/TFS-spectrogram and the normal auditory-spectrogram, the relative contribution of neural E/TFS cues to speech intelligibility can be evaluated.

As illustrated in Figure 5.6, given a speech stimulus, its normal auditory-spectrogram was first computed, which is a function of time and frequency, denoted as $S(t, f)$. In parallel, the stimulus was mixed with a masker, which can be either speech-shaped steady state noise (SSN) or single competing talker, to simulate a certain signal-to-noise ratio (SNR).

The mixture signal was then processed to generate four types of simulation sounds, as described in previous section. For each simulation sound, the corresponding E-spectrogram and TFS-spectrogram were constructed, and then compared with the previously generated auditory-spectrogram. Specifically, the correlation coefficient between the E -spectrogram and the auditory-spectrogram was calculated as followes:

$$r_{E\_clean} = \langle S_E(t,f), S(t,f) \rangle \tag{5.7}$$

where $\langle \cdot, \cdot \rangle$ stands for the two-dimensional inner product calculation, and $S_E(t,f)$ represents the E-spectrogram. The definition of r_{TFS_clean} can be given by replacing the E-spectrogram with the TFS-spectrogram in the above equation. Every spectrogram was normalized to zero mean and unit variance. Technical details about the correlation calculation can be found in Shamma and Lorenzi [6].

The speech stimuli were taken from the IEEE sentence lists. A total of four lists (i.e., 40 sentences) were randomly picked. Each list was produced by a different speaker (two male and two female). The SSN masker was generated by filtering white noise with a target sentence's long-term spectrum. The competing sentence was produced by one of the female speakers and it lasted longer than any target sentence used in the experiment. The SNR varied from -10 to +25 dB. For each type of simulation sound, given a particular masker at a specific SNR condition, a total of 40 data points were obtained for r_{E_clean} and r_{TFS_clean}, respectively. The mean r_{E_clean} and r_{TFS_clean} averaged across all 40 sentences were reported in the following.

5.3 Results

5.3.1 The relative effectiveness of neural E and TFS coding in steady state noise

Figure 5.7 shows the correlation results for speech-shaped steady state noise. The left panel corresponds to the conventional Hilbert decomposition, while the right panel is about the new decomposition. As SNR decreased from +25 to -10 dB, each correlation curve degraded monotonically, reflecting that the neural cues for intelligibility were corrupted by noise interference.

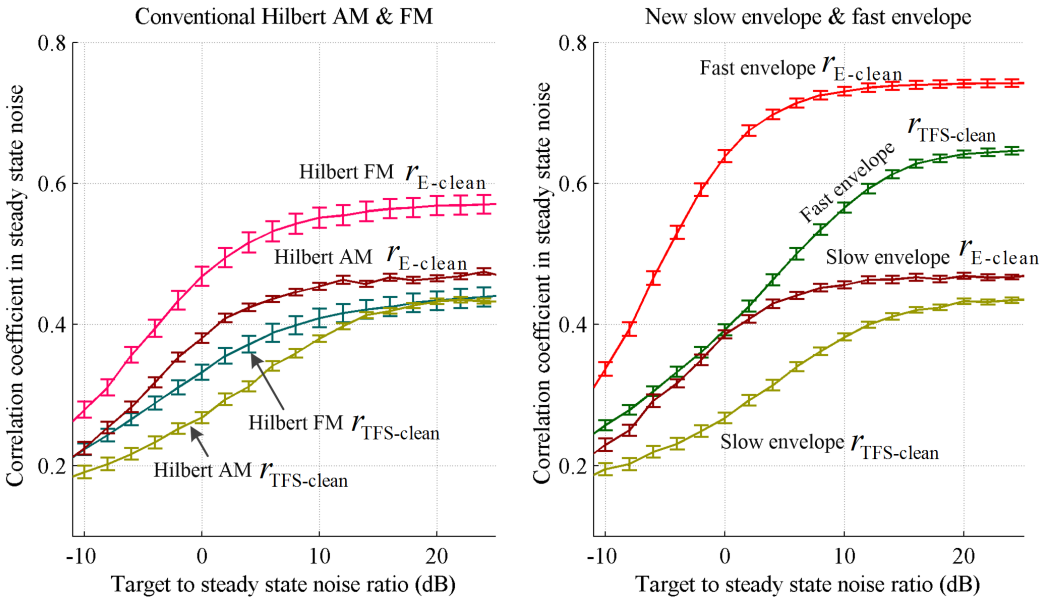

Figure 5.7: The relative effectiveness of neural E and TFS coding in steady state noise as measured by the correlation metric. Error bars represent the standard error of the mean.

The neural E and TFS coding evoked by Hilbert AM produced similar correlation results as those elicited by slow envelope, which is expected because they are mathematically equivalent. Between Hilbert FM and fast envelope, however, significant differences were observed. At a given SNR condition (e.g., > -5 dB), the neural E and TFS coding evoked by Hilbert FM produced much smaller correlation scores than those elicited by fast envelope. In fact, the neural TFS coding evoked by Hilbert FM seemed roughly as effective as those elicited by Hilbert AM (e.g., > 10 dB). This was clearly not the case with fast envelope. Within each type of simulation sounds, the evoked neural E coding produced higher correlation scores than the neural TFS coding across all SNR conditions, suggesting that neural E coding was a primary contributor to speech intelligibility in steady state noise.

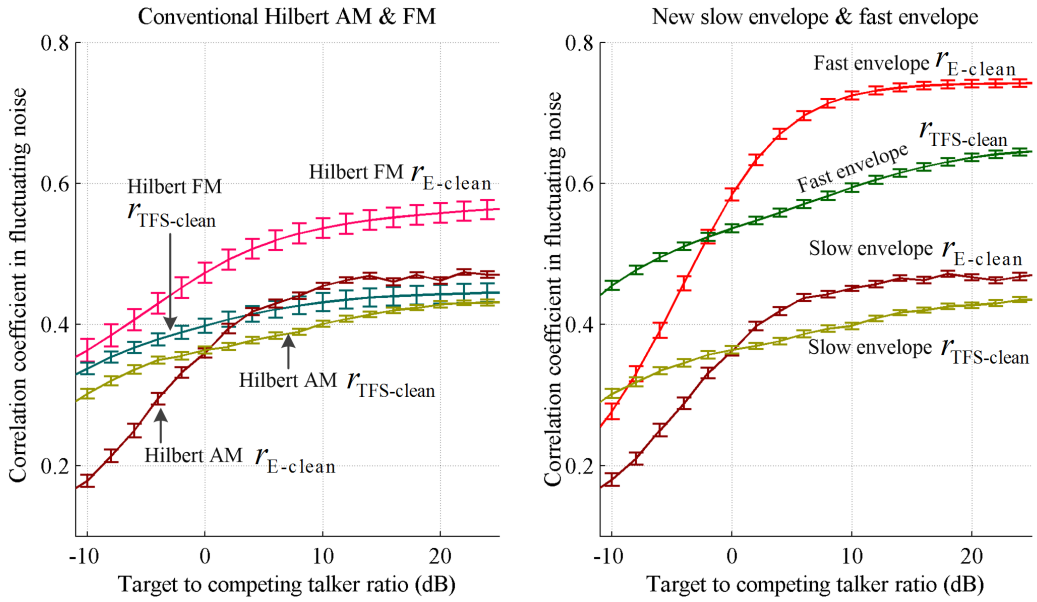

Figure 5.8: The relative effectiveness of neural E and TFS coding in the presence of single competing talker. Error bars represent the standard error of the mean.

5.3.2 The relative effectiveness of neural E and TFS coding in the presence of single competing talker

In the presence of single competing talker, the relative effectiveness of neural E and TFS coding showed different trends than in steady state noise. One major difference is that the correlation scores of the neural E and TFS coding evoked by fast envelope showed a crossover near -2 dB. A similar crossover was also observed with Hilbert AM and slow envelope, but not with Hilbert FM. This suggests that neural TFS coding contributes significantly to speech intelligibility in fluctuating noise.

Between Hilbert AM and slow envelope, there was no significant difference as expected. Like in steady state noise, the neural TFS coding evoked by fast envelope produced much higher correlation scores than that elicited by Hilbert FM. In fact, the neural TFS coding evoked by Hilbert FM seemed just marginally more effective than those elicited by Hilbert AM. With Hilbert FM, the evoked neural E coding produced higher correlation scores than neural TFS coding across all SNR conditions.

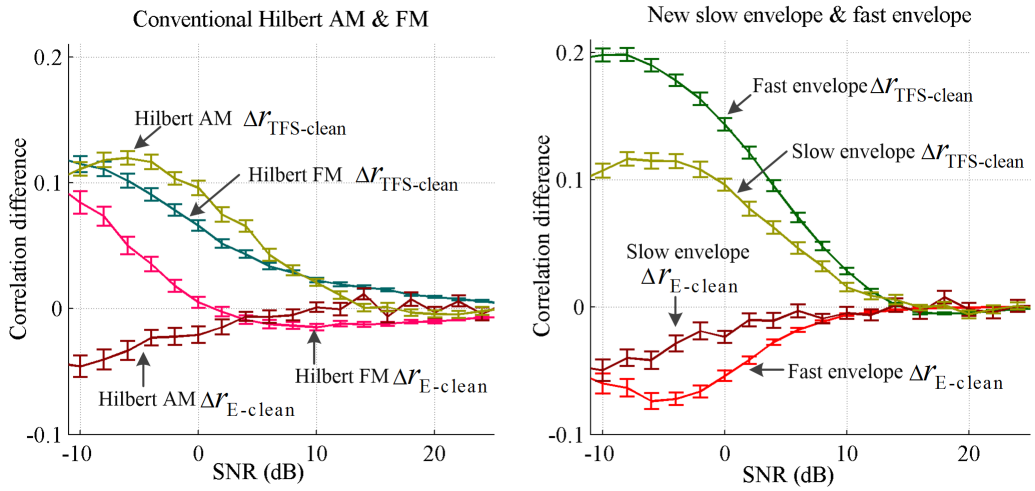

Figure 5.9: : By changing the masker from steady state noise to single competing talker, the increment in correlation score as a function of SNR for each type of neural coding.

5.3.3 Masker effect on correlation results

To analyze the effect of masker type on the correlation results, each correlation curve shown in Figure 5.8 was compared with its correspondence in Figure 5.7. Take the neural TFS coding evoked by fast envelope as an example: the correlation score obtained in the fluctuating masker at a particular SNR was compared with that obtained in the steady state noise at the same SNR. Figure 5.9 shows the difference in correlation score as a function of SNR for each type of neural coding.

As SNR decreased from +10 to -10 dB, the neural E coding evoked by fast envelope exhibited decrement in correlation score as the masker was changed from steady state noise to single competing talker, suggesting that neural E coding degraded faster in fluctuating noise than in steady state noise. By contrast, the neural TFS coding evoked by fast envelope degraded much slower in fluctuating noise than in steady noise, suggesting that neural TFS coding might be a primary contributor to masking release in normal-hearing listeners. Similar trends were found with slow envelope and Hilbert AM. Differently, the neural E coding elicited by Hilbert FM degraded slower in fluctuating noise than in steady noise at negative SNRs. Similar trend was observed by Swaminathan [156].

5.4 Discussion

Due to the concern that conventional Hilbert FM is a flawed, i.e., distorted, representation of TFS, we proposed a more principled engineering definition with connection to presumed physiolog, called the fast envelope. Using both the conventional Hilbert decomposition and the new slow envelope and fast envelope decomposition, we explored the relative contribution of neural E and TFS coding to speech intelligibility in different noise conditions. Overall, the trends observed with the fast envelope differ significantly from those obtained with Hilbert FM as used in conventional vocoding.

5.4.1 Distortions in Hilbert FM confounded the relative importance of TFS

Using the conventional Hilbert AM and FM decomposition, Swaminathan and Heinz [90] found that in steady state noise, the significance of neural TFS coding was only marginal at positive SNRs but became relatively substantial at negative SNRs. Nevertheless, neural TFS coding was constantly a weaker cue compared to neural E coding. This observation also holds true in the presence of modulated noise [156]. Although the neural representations used in their studies were fundamentally different from the auditory spectrograms used in this study, the correlation metrics were not substantially different [6]. Thus, we can make some qualitative comparisons. In line with the previous studies, current results about Hilbert FM also showed that neural E coding was relatively more effective than neural TFS coding in the presence of steady state noise or single competing talker. However, the trends observed with fast envelope differ significantly from this, which we will discuss later.

Shamma and Lorenzi [6] found that in quiet, the neural E and TFS coding invoked by Hilbert FM seemed comparably effective, and they did not differ significantly from those elicited by Hilbert AM. Consistent with their observations, we found that at high SNRs, the neural TFS coding evoked by Hilbert FM seemed comparably effective as the TFS coding elicited by Hilbert AM. This trend can be partly explained by the fact for some bandpass signals, the envelope and TFS are related, thus TFS cues can be partially recovered from Hilbert AM. However, this is not true for many amplitude modulation signals (see Figure 5.3). In fact, by comparing the neural TFS coding evoked by Hilbert FM with that elicited

by fast envelope, we can see that using Hilbert FM, the relative contribution of TFS was potentially estimated as being too low. The discontinuities in conventional Hilbert FM confounded the relative importance of TFS and made it seem insignificant.

5.4.2 The significant role of TFS in masking release was demonstrated using fast envelope

At lower SNRs, normal-hearing listeners can recognize speech better in fluctuating noise than in steady state noise—a phenomena known as masking release. Qin and Oxenham [44] showed that in the presence of single talker masker, normal-hearing listeners need a lower SNR to achieve 50% correctness on speech recognition task than in steady state noise (e.g., -12 versus -7 dB). However, when speech was processed to preserve E cues but discard TFS, the single-talker masker became more difficult than the steady state noise. These trends can be well explained by the observation that the neural TFS coding evoked by fast envelope was more resilient to fluctuating interference, whereas the neural E coding degraded faster in fluctuating noise than in steady noise (see Figure 5.9).

Hopkins et al. [49] investigated the benefit of TFS to speech perception in fluctuating background. To preserve TFS in a channel, they kept the corresponding bandpass signal unprocessed—this idea essentially coincides with the definition of fast envelope. Hopkins et al. showed that as TFS cues were systematically increased from low to high frequency, the SNR required by normal-hearing listeners to achieve 50% correctness was decreased gradually. In this study, the phase locking limit was fixed at 1 kHz. The benefit of TFS for "listening in the background dips" was clearly demonstrated with fast envelope but not with Hilbert FM (see Figure 5.8).

It should be noted that current results do not indicate that neural TFS cues alone can support robust speech recognition. As illustrated in Figure 5.3, E and TFS often convey very different information about the sound, thus they likely play different roles in speech perception. Conceptually, the ability to "listen in the background dips" relies on neural TFS coding to identify the spectral/temporal regions where the target speech overpowers the masker, and depends on neural E coding to recognize the target speech in the regions of these dips. For example, TFS coding is known to be important for pitch perception, which

plays a critical role in talker separation but not necessarily important for speech recognition in quiet [106].

5.4.3 Relevance to TFS deficit in hearing loss

Hypothetically, the new slow envelope and fast envelope decomposition can help to understand the perceptual TFS deficit in hearing impaired listeners. As mentioned before, slow envelope reflects the temporal cues that can be coded in average firing rates, while fast envelope corresponds to the temporal cues represented in phase locking patterns. Lorenzi et al. [157] showed that listeners with cochlear hearing loss can have normal audiometric thresholds but still a reduced ability to perceive TFS information. Presumably, the audiometric threshold, like slow envelope, relies on average firing rate; thus it can be measured as normal in spite of phase locking degradation. Similarly, restoring hearing threshold via hearing aid amplification may recover slow envelope cues, but probably not fast envelope (i.e., TFS) cues. As a result, in spite of the amplification, the benefit of TFS to speech perception in noise is still not accessible.

Henry and Heinz [158] showed that an important consequence of sensorineural hearing loss is that some auditory-nerve fibers grow less responsive to TFS information near their characteristic frequency but increasingly responsive to low-frequency temporal information. This deficit in temporal coding directly affects fast envelope representation. Consequently, the observed benefit of fast envelope to speech perception in fluctuating noise might be lost. For example, Hopkins et al. [49] found that for speech recognition with competing talker, hearing impaired listeners cannot benefit from TFS as much as normal-hearing listener did.

5.4.4 Implications for cochlear implants

Cochlear implant (CI) users rely primarily on E cues, thus they cannot take advantage of the robustness of TFS coding in fluctuating noise. As a result, CI users typically do not receive any masking release; instead, they often perform worse in the presence of a competing talker than in steady state noise [45]. Shannon [24] found that CI users sensitivity to temporal cues declined significantly above 300 Hz. By contrast, neural TFS coding relies on phase

locking up to a few kHz. Consequently, to restore the perceptual benefits of neural TFS coding in CI users is fundamentally difficult.

It is interesting to note that the TFS of a waveform consisting of multiple components is dominated by the stronger component (see Figure 5.3). In the neural responses evoked by speech, the phase locking patterns are often dominated by harmonics, particularly the stronger harmonics near the vowel formants [3]. In practice stronger components are relatively more resilient to interference. This suggests that to encode TFS cues for CI users, the stronger components should be be prioritized. Additionally, given the coupling between E and TFS, it is important to process them as a whole, rather than separately [92]. For example, Li et al. [60] proposed to code the predominant harmonic components and transform the E and TFS cues as a whole to modulators for electrodes. They demonstrated using acoustic simulations that speech perception in noise can be improved. This illustrates the potentially large benefit of providing TFS information in a cochlear implant.

Chapter 6
CONCLUSION

This dissertation studies engineering signal decompositions for careful study of temporal cues in complex broadband sounds, commonly known as envelope and temporal fine structure (TFS). Due to the coupling between envelope and phase, it is problematic to isolate the TFS from the envelope for any signal which is not extremely narrowband. The conventional definition of TFS, Hilbert FM, is flawed for natural signals, such as speech. Nevertheless, many perception and modeling studies have used Hilbert FM to investigate the importance of TFS. To address this concern, we have derived more principled engineering definition of TFS, which allows one to potentially study the importance of TFS in a deeper way. Additionally, we developed original TFS coding strategies for cochlear implant (CI) users and demonstrated the potentially large benefit of TFS for speech perception in noise and music perception with CIs. In the following, we review major results and briefly discuss remaining questions for future work.

6.1 Main results

Given a bandpass signal, its envelope and phase are often related, rather than independent. The coupling between envelope and phase makes it an ill-posed problem to separate the TFS from the envelope. Conventionally, a Hilbert transform is used to represent a bandpass signal as the product of two components, the Hilbert envelope or amplitude-modulation (AM) function and a frequency-modulated (FM) sine wave carrier. This product decomposition is regarded as a separation of envelope and TFS. However, many previous studies found that Hilbert FM can still contain envelope cues (or lead to envelope recovery), which is expected given the unavoidable coupling between envelope and phase. Besides, there are signal discontinuities in Hilbert FM, which makes it a flawed representation of TFS. The discontinuities cause Hilbert FM to be infinite bandwidth. As a result, Hilbert FM will

introduce within-band (intermodulation) frequency distortions and restore some seemingly eliminated but instead distorted envelope when being processed by narrowband auditory filters.

To address the concerns about Hilbert AM and FM decomposition, we proposed to process envelope and TFS as a whole, rather than separately. This idea was reflected in our development of TFS coding strategies for CI users. Perception studies suggest that TFS plays an important role in pitch perception and speech perception in noise, especially in fluctuating background sounds. Due to the inherently coarse spectral and temporal resolution in electric hearing, conventional cochlear implant (CI) coding strategies only transmit envelope cues in a small number of channels. Thus, the perceptual benefits of TFS information are mostly inaccessible to CI users, which potentially contribute to their difficulties in understanding speech in noise and perceiving music. To encode fine structure information for CI users, we proposed a harmonic-single-sideband-encoder (HSSE) strategy that explicitly tracks the harmonics in complex sounds and transforms them into modulators conveying both envelope and TFS cues. A key distinction about HSSE is that it keeps the envelope and TFS cues together during the transformation, such that the extracted modulators do not contain distortions. The effectiveness of HSSE on speech perception in noise, Mandarin tone discrimination, and music perception were evaluated using three approaches, including acoustic simulation in normal-hearing listeners, computational auditory-nerve modeling, and acute test in actual CI users. Significant improvements were demonstrated in most tasks.

The potential advantage of HSSE can be attributed to a few factors. First, HSSE is designed to track predominant harmonics and then convert them into stimulation signals conveying both envelope and TFS cues. The processing hypothetically reproduces the neural response patterns observed in normal auditory systems, where phase locking patterns are often dominated by stronger harmonics. Since there are fundamental differences between electric stimulation and acoustic stimulation in terms of neural coding abilities, to restore the normal neural spike patterns in CI users is almost impossible. However, it is still potentially advantageous to code predominant harmonics, because they are potentially more resilient to interference and they play a larger role in pitch perception. From a practical point of

view, most CI users only have a small number of effective channels. Thus, it is practically important to focus on the predominant harmonics.

Additionally, we proposed a new engineering distortion-free additive view of signal decomposition, the slow envelope and the fast envelope, using half wave rectification followed by filters reflecting engineering interpretation of physiological considerations. The slow envelope is a tool for representing temporal cues that can be coded in the average firing rate of auditory nerve fibers, while the fast envelope instead captures the temporal cues represented in neural phase locking patterns. Compared with Hilbert FM, the fast envelope is a more principled signal-based definition of TFS. Using this new slow envelope and fast envelope decomposition, we investigated the relative contribution of neural envelope and TFS coding to speech intelligibility in different noise conditions. The neural representation was generated by a simplified peripheral auditory model (Shamma and Lorenzi, 2013). In our modeling experiments, the two decomposition methods led to significantly different levels of relative importance of neural envelope and TFS coding, which was different than Shamma and Lorenzi [6] Hilbert vocoder-based conclusion. The trends observed with fast envelope correlate better with previous perceptual study results, which suggest that neural TFS coding plays a significant role in masking release. By contrast, the discontinuities in the previous Hilbert FM approach likely confounded the importance of neural TFS coding and made it seem insignificant.

6.2 Future work

As the number of CI uses grows globally, the question of how to represent fine structure information for CI users becomes increasingly important. The proposed HSSE strategy can potentially contribute to solving this issue. Meanwhile, there are several open questions worth consideration. In high frequency where multiple harmonics may fall into one channel, the beating between harmonics can generate fundamental frequency cues in the envelope. Physiological data show that auditory-nerve fibers can phase lock to the envelope. Typically, sound energy becomes more spread, rather than concentrated, in high frequency. The idea of detecting predominant components in high frequency does not seem appealing in this respect. Thus, it might be desirable to use envelope based coding methods for high

frequencies but harmonic based coding for low frequencies. Second, it is generally assumed that TFS information is important to consonant perception. Given that HSSE is tailored for harmonic processing, it is not clear whether or not the extracted TFS information will contribute to consonant perception. It might be helpful to collect some consonant recognition data and analyze the confusion matrix to examine the effect of HSSE on different phonetic features.

The new decomposition provides a better signal processing tool for investigating the contribution of TFS to speech perception in noise. The slow envelope and the fast envelope corresponds respectively to an new engineering formalism for neural envelope and TFS coding. By varying the cutoff frequency of the lowpass filter (right after the half-wave rectifier), we can simulate different degrees of phase locking degradation, or TFS deficit. Conceptually, we can measure speech recognition performance in noise using both fast envelope speech and slow envelope speech. By comparing their difference in recognition performance, the perceptual benefit of TFS can be evaluated. To conduct speech recognition using fast envelope speech, we need to pay attention to one issue. In the high frequency region where phase locking declines significantly, there will not be much energy in fast envelope. To allow speech recognition, in the channels where fast envelope is not available,the slow envelope can be extracted and combined with fast envelope using vocoder synthesis techniques.

Current automatic speech recognition systems often have a coarse representation of temporal features. The advantage of TFS is missing. Modern recognition systems possess a high level of complexity, which enables them to well represent fine structure information. It is worthy to investigate how TFS can be reflected in robust automatic speech recognition systems. Similarly, given the importance of TFS to speech recognition in noise, the effect of TFS should be taken into account in developing speech intelligibility measures.

BIBLIOGRAPHY

[1] P. X. Joris, C. E. Schreiner, and A. Rees, "Neural processing of amplitude-modulated sounds," *Physiol Rev*, vol. 84, pp. 541 – 577, 2004.

[2] P. X. Joris, D. H. Louage, L. Cardoen, and M. van der Heijden, "Correlation index: a new metric to quantify temporal coding," *Hear Res*, vol. 216-217, pp. 19– 30, 2006.

[3] S. A. Shamma, "Speech processing in the auditory system i: The representation of speech sounds in the responses of the auditory nerve," *J Acoust Soc Am*, vol. 78, pp. 1612 – 21, 1985.

[4] D. D. Greenwood, "A cochlear frequency-position function for several species–29 years later," *J Acoust Soc Am*, vol. 87, no. 6, pp. 2592–605, 1990.

[5] P. Clark and L. E. Atlas, "Time-frequency coherent modulation filtering of nonstationary signals," *IEEE Trans Sig Proc*, vol. 57, no. 11, pp. 4323–32, 2009.

[6] S. Shamma and C. Lorenzi, "On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system," *J Acoust Soc Am*, vol. 133, pp. 2818 – 2833, 2013.

[7] B. C. J. Moore, *An introduction to the psychology of hearing*, Academic Press, 2003.

[8] H. Fletcher, "Auditory patterns," *Reviews of Modern Physics*, vol. 12, pp. 47 – 65, 1940.

[9] H Spoendlin and A. Schrott, "Analysis of the human auditory nerve," *Hear Res*, vol. 43, pp. 25 – 38, 1989.

[10] T. J. Glattke, "Unit responses of the cat cochlear nucleus to amplitude-modulated stimuli," *J Acoust Soc Am*, vol. 45, pp. 419 – 425, 1969.

[11] D. H. Johnson, "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," *J Acoust Soc Am*, vol. 68, pp. 1115 – 1122, 1980.

[12] E. Javel, "Coding of am tones in the chinchilla auditory nerve: Implications for the pitch of complex tones," *J Acoust Soc Am*, vol. 68, pp. 133 – 146, 1980.

[13] P. X. Joris and T. C. Yin, "Responses to amplitude-modulated tones in the auditory nerve of the cat," *J Acoust Soc Am*, vol. 91, pp. 215 – 232, 1992.

[14] T. F. Weiss and C. Rose, "A comparison of synchronization filters in different auditory receptor organs," *Hear Res*, vol. 33, pp. 175 – 180, 1988.

[15] J. M. Goldberg and Brown P. B., "Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localization," *J Neurophysiol*, vol. 32, pp. 613 – 636, 1969.

[16] W. S. Rhode, "Interspike intervals as a correlate of periodicity pitch in cat cochlear nucleus," *J Acoust Soc Am*, vol. 97, pp. 2414 – 2429, 1995.

[17] P. A. Cariani and B. Delgutte, "Neural correlates of the pitch of complex tones ii: Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch," *J Neurophysiol*, vol. 76, pp. 1717 – 1734, 1996.

[18] D. H. Louage, M. van der Heijden, and P. X. Joris, "Temporal properties of responses to broadband noise in the auditory nerve," *J Neurophysiol*, vol. 91, pp. 2051– 65, 2004.

[19] M. Heinz and J. Swaminathan, "Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech," *J Assoc Res Otolaryngol*, vol. 10, pp. 407–23, 2009.

[20] J. Eggermont, "Functional aspects of synchrony and correlation in the auditory nervous system," *Concepts Neurosci*, vol. 4, pp. 105–129, 1993.

[21] M. R. Schroeder, "Modulation transfer functions: definition and measurement," *Acustica*, vol. 49, pp. 179 – 182, 1981.

[22] N.l F. Viemeister, "Temporal modulation transfer functions based upon modulation thresholds," *J. Acoust. Soc. of Am.*, vol. 66, no. 5, pp. 1364–1380, 1979.

[23] S. P. Bacon and N.l F Viemeister, "Temporal modulation transfer functions in normal-hearing and hearing-impaired listeners," *Audiology*, vol. 24, no. 2, pp. 117–134, 1985.

[24] RV Shannon, "Temporal modulation transfer functions in patients with cochlear implants," *J Acoust Soc Am*, vol. 91, no. Apr, pp. 2156–2164, 1992.

[25] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, 1985.

[26] H. Dudley, "Remaking speech," *J Acoust Soc Am*, vol. 11, pp. 169–177, 1939.

[27] S. Rosen, "Temporal information in speech: acoustic, auditory and linguistic aspects," *Phiolos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 336, pp. 367–373, 1992.

[28] B. C. 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 Otolaryngol*, vol. 9, pp. 399 – 406, 2008.

[29] M. Miller, *Representation of stop consonants in the discharge patterns of auditory-nerve fibers*, Ph.d. dissertation, Johns Hopkins University, 1984.

[30] B. Barta, *Testing stimulus encoding in the auditory nerve*, Ph.d. dissertation, Johns Hopkins University, 1984.

[31] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J Acoust Soc Am*, vol. 24, pp. 175 – 184, 1952.

[32] C. G. Fant, *Speech Sounds and Features*, M IT, Cambridge MA, 1973.

[33] R. J. Ritsma, "Frequencies dominant in the perception of the pitch of complex sounds," *J Acoust Soc Am*, vol. 42, no. 1, pp. 191–8, 1967.

[34] M. B. Sachs and E. D. Young, "Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate," *J Acoust Soc Am*, vol. 66, pp. 470 – 479, 1979.

[35] E. D. Young and M. B. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J Acoust Soc Am*, vol. 66, pp. 1381 – 1403, 1979.

[36] M. Miller and M. Sachs, "Representation of stop consonants in the discharge patterns of auditory-nerve fibers," *J Acoust Soc Am*, vol. 74, pp. 502 – 517, 1983.

[37] B. Delgutte, "Speech coding in the auditory nerve ii: Processing schemes for vowel-like sounds," *J Acoust Soc Am*, vol. 75, pp. 879 – 86, 1984.

[38] H. E. Secker-Walker and C. L. Searle, "Time-domain analysis of auditory-nerve-fiber firing rates," *J Acoust Soc Am*, vol. 88, pp. 1427 – 36, 1990.

[39] Robert V. Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, pp. 303–304, 1995.

[40] P. J. Blamey, R. C. Dowell, G. M. Clark, and P. M. Seligman, "Acoustic parameters measured by a formant-estimating speech processor for a multiple-channel cochlear implant," *J Acoust Soc Am*, vol. 82, pp. 38 – 47, 1987.

[41] P. W. Skinner, L. K. Holden, T. A. Holden, R. C. Dowell, P. M. Seligman, J. A. Brimacombe, and A. L. Beiter., "Performance of postlinguistically deaf adults with the wearable speech processor (wsp iii) and mini speech processor (msp) of the nucleus multi-electrode cochlear implant," *Ear Hear*, vol. 12, pp. 3 – 22, 1991.

[42] Blake S. Wilson, Charles C. Finley, Dewey T. Lawson, Robert D. Wolford, Donald K. Eddington, and William M. Rabinowitz, "Better speech recognition with cochlear implants," *Nature*, vol. 352, no. 6332, pp. 236–238, 1991.

[43] H. J. McDermott, C. M. McKay, and A. E. Vandali, "A new portable sound processor for the university of melbourne/nucleus limited multielectrode cochlear implant," *J Acoust Soc Am*, vol. 91, no. 6, pp. 3367–3371, 1992.

[44] Michael K. Qin and Andrew J. Oxenham, "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," *J Acoust Soc Am*, vol. 114, no. 1, pp. 446–454, 2003.

[45] Ginger S. Stickney, Fan-Gang Zeng, Ruth Litovsky, and Peter Assmann, "Cochlear implant speech recognition with speech maskers," *J Acoust Soc Am*, vol. 116, no. 2, pp. 1081–1091, 2004.

[46] Zachary M. Smith, Bertran Delgutte, and Andrew J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, pp. 87–90, 2002.

[47] Fan-Gang Zeng, Kaibao Nie, Ginger S. Stickney, Ying-Yee Kong, Michaels Vongphoe, Ashish Bhargave, Chaogang Wei, and Keli Cao, "Speech recognition with amplitude and frequency modulations," *Proc Natl Acad Sci USA*, vol. 102, no. 7, pp. 2293–2298, 2005.

[48] Christian Lorenzi, Gatan Gilbert, Hlose Carn, Stphane Garnier, and Brian C. J. Moore, "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *Proceedings of the National Academy of Sciences*, vol. 103, no. 49, 2006.

[49] K. Hopkins, B. C. J. Moore, and M. A. Stone, "Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech," *J Acoust Soc Am*, vol. 123, no. 2, pp. 1140–1153, 2008.

[50] F. G. Zeng, S. Rebscher, W. Harrison, X. Sun, and H. Feng, "Cochlear implants: System design, integration, and evaluation," *IEEE Rev in Biomed Eng*, vol. 1, pp. 115–142, 2008.

[51] Andrew E. Vandali, Lesley A. Whitford, Kerrie L. Plant, and Graeme M. Clark, "Speech perception as a fucntion of electrical stimulation rate: Using the nucleus 24 cochlear implant system," *Ear and Hearing*, vol. 21, no. 6, pp. 608–624, 2000.

[52] I. Hochmair, P. Nopp, C. Jolly, M. Schmidt, H. Schosser, C. Garnham, and I. Anderson, "Med-el cochlear implants: state of the art and a glimpse into the future," *Trends Amplif*, vol. 10, no. 4, pp. 201–19, 2006.

[53] L. M. Friesen, R. V. Shannon, D. Baskent, and X. Wang, "Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants," *J Acoust Soc Am*, vol. 110, pp. 1150–63, 2001.

[54] Fan-Gang Zeng, "Trends in cochlear implants," *Trends in Amplification*, vol. 8, no. 1, pp. 1–34, 2004.

[55] A. J. Spahr, M. F. Dorman, and L. H. Loiselle, "Performance of patients using different cochlear implant systems: effects of input dynamic range," *Ear Hear*, vol. 28, no. 2, pp. 260–75, 2007.

[56] L. Xu, X. Chen, H. Lu, N. Zhou, S. Wang, Q. Liu, Y. Li, X. Zhao, and D. Han, "Tone perception and production in pediatric cochlear implants users," *Acta Otolaryngol*, vol. 131, pp. 395–8, 2011.

[57] Robert Kang, Grace Liu Nimmons, Ward R. Drennan, Jeff Longnion, Chad Ruffin, Kaibao Nie, Jong Ho Won, Tina Worman, Bevan Yueh, and Jay T. Rubinstein, "Development and validation of the university of washington clinical assessment of music perception test," *Ear and Hearing*, vol. 30, no. 4, pp. 411–8, 2009.

[58] K. H. Jung, J. H. Won, W. R. Drennan, E. Jameyson, G. Miyasaki, S. J. Norton, and J. T. Rubinstein, "Psychoacoustic performance and music and speech perception in prelingually deafened children with cochlear implants," *Audiol Neurootol*, vol. 17, no. 3, pp. 189–97, 2012.

[59] Xing Li, Kaibao Nie, Les Atlas, and Jay Rubinstein, "Harmonic coherent demodulation for improving sound coding in cochlear implants," in *Proc. ICASSP*. 2010, pp. 5462–5, IEEE.

[60] X. Li, K. B. Nie, N. S. Imennov, J. H. Won, W. R. Drennan, J. T. Rubenstien, and L. E. Atals, "Improved perception of speech in noise and mandarin tones with acoustic simulations of harmonic coding for cochlear implants," *J Acoust Soc Am*, vol. 132, pp. 3387–98, 2012.

[61] X. Li, K. Nie, N.S. Imennov, J.T. Rubinstein, and L.E. Atlas, "Improved perception of music with a harmonic based algorithm for cochlear implants," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, in press.

[62] N. S. Imennov and J. T. Rubinstein, "Stochastic population model for electrical stimulation of the auditory nerve," *IEEE Trans Biomed Eng*, vol. 56, no. 10, pp. 2493–501, 2009.

[63] R. Drullman, J.M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, vol. 95, pp. 1053–1064, 1994.

[64] J. L. Flanagan, D. I. S. Meinhart, Roger M. Golden, and Man Mohan Sondhi, "Phase vocoder," *J. Acoust. Soc. of Am.*, vol. 38, no. 5, pp. 939–940, 1965.

[65] J.L. Flanagan and R.M. Golden, "Phase vocoder," *Bell Syst. Tech.*, vol. 45, pp. 1493 – 1509, 1966.

[66] M. Portnoff, "Time-frequency representation of digital signals and systems based on short-time fourier analysis," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, pp. 55 – 69, 1980.

[67] J. L. Flanagan, "Parametric coding of speech spectra," *J accoust Soc Am*, vol. 68, no. 2, pp. 412–419, 1980.

[68] V. Bargmann, "On a hilbert space of analytic functions and an associated integral transform," *Commun. Pure Appl. Math*, vol. 14, pp. 187 – 214, 1961.

[69] E. Chassande-Mottin, I. Daubechies, F. Auger, and P. Flandrin, "Differential reassignment," *Signal Processing Letters, IEEE*, vol. 4, pp. 293–294, 1997.

[70] F. Auger, E. Chassande-Mottin, and P. Flandrin, "On phase-magnitude relationships in the short-time fourier transform," *Signal Processing Letters, IEEE*, vol. 19, no. 5, pp. 267–270, 2012.

[71] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," *Speech Communication*, vol. 10, pp. 209 – 221, 1991.

[72] P. Sndergaard, R. Decorsiere, and Torsten Dau, "On the relationship between multichannel envelope and temporal fine structure," in *3rd International Symposium on Auditory and Audiological Research, Nyborg*, 2012.

[73] J. S. Lim, *Speech Enhancement*, Englewood Cliffs, NJ: Prentice-Hall, 1983.

[74] M. Portnoff, "Implementation of the digital phase vocoder using the fast fourier transform," *IEEE Tran. on Acoustics, Speech and Signal Processing*, vol. 24, pp. 243 – 248, 1976.

[75] M. Portnoff, "Time-scale modification of speech based on short-time fourier analysis," *IEEE Tran. on Acoustics, Speech and Signal Processing*, vol. 29, no. 3, pp. 374 – 390, 1981.

[76] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Tran. on Acoustics, Speech and Signal Processing*, vol. 32, pp. 236 – 243, 1984.

[77] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, pp. 529 – 541, 1981.

[78] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication*, vol. 22, pp. 403–417, 1997.

[79] K.K. Paliwal and L.D. Alsteris, "On the usefulness of stft phase spectrum in human listening tests," *Speech Communication*, vol. 45, pp. 153–170, 2005.

[80] G. Shi, M.M. Shanechi, and P. Aarabi, "On the importance of phase in human speech recognition," *IEEE Tran. Audio, Speech, and Language Processing*, vol. 14, pp. 1867 – 1874, 2006.

[81] M. Kazama1, S. Gotoh, M. Tohyama, and T. Houtgast, "On the significance of phase in the short term fourier spectrum for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 127, pp. 1432–1439, 2010.

[82] E. Loveimi and S.M. Ahadi, "Objective evaluation of phase and magnitude only reconstructed speech: New considerations," in *10th International Conference on Information Sciences Signal Processing and their Applications*, 2010, pp. 117 – 120.

[83] H.B. Voelcker, "Toward a unified theory of modulation part i: Phase-envelope relationships," *Proceedings of the IEEE*, vol. 54, no. 3, pp. 340–353, 1966.

[84] H.B. Voelcker and Aristides A G Requicha, "Clipping and signal determinism: Two algorithms requiring validation," *IEEE Trans. on Communications*, vol. 21, no. 6, pp. 738–744, 1973.

[85] B. F. Logan, "Information in zero crossing of bandpass signals," *Bell Syst. Tech.*, vol. 56, pp. 487 – 510, 1977.

[86] O. Ghitza, "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," *J. Acoust. Soc. Am.*, vol. 110, pp. 1628–1640, 2001.

[87] Fan-Gang Zeng, K. Nie, S. Liu, G. Stickney, E. Del Rio, Y. Y. Kong, and H. Chen, "On the dichotomy in auditory perception between temporal envelope and fine structure cues," *J Acoust Soc Am*, vol. 116, pp. 1351–1354, 2004.

[88] Gaetan Gilbert and Christian Lorenzi, "The ability of listeners to use recovered envelope cues from speech fine structure," *J Acoust Soc Am*, vol. 119, no. 4, pp. 2438–2444, 2006.

[89] Stanley Sheft, Marine Ardoint, and CHristian Lorenzi, "Speech identification based on temporal fine-structure cues," *J Acoust. Soc. Am.*, vol. 124, no. 1, pp. 562–575, 2008.

[90] J. Swaminathan and M. Heinz, "Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise," *J Neurosci*, vol. 32, pp. 1747–56, 2012.

[91] Les Atlas and Christiaan Janssen, "Coherent modulation spectral filtering for single-channel music source separation," *Proc. IEEE ICASSP*, vol. IV, pp. 461–464, 2005.

[92] S.M. Schimmel and L. Atlas, "Coherent envelope detection for modulation filtering of speech," in *Proc. ICASSP*. 2005, pp. 221–224, IEEE.

[93] B. A. Atal and M. R. Schroeder, "Adaptive predictive coding of speech," *Bell Syst. Tech.*, vol. 49, pp. 1973 – 1986, 1970.

[94] Bishnu S. Atal and Joel R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bitrates," in *Proc. IEEE ICASSP*, 1982.

[95] M. Schroeder and B. Atal, "Code-excited linear prediction(celp): High-quality speech at very low bit rates," in *IEEE ICASSP*, 1985, vol. 10, pp. 937 – 940.

[96] B. S. Atal, "The history of linear prediction," *IEEE Signal Processing Magazine*, vol. 23, pp. 154 – 161, 2006.

[97] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (tns)," in *the 101st AES Conv.*, 1996.

[98] J. Herre, "Temporal noise shaping, quantization and coding methods in perceptual audio coding: A tutorial introduction," in *3rd International Symposium on Auditory and Audiological Research*, 2011.

[99] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *J Acoust Soc Am*, vol. 105, pp. 1912 – 1924, 1999.

[100] R. Kumaresan, "On minimum/maximum/all-pass decompositions in time and frequency domains," *IEEE Trans. Signal Process*, vol. 48, pp. 2973 – 2976, 2000.

[101] M. Athineos and D. Ellis, "Autoregressive modeling of temporal envelopes," *IEEE Trans. Signal Process*, vol. 55, pp. 5237 – 5245, 2007.

[102] Han-Wen Hsu and Chi-Min Liu, "Autoregressive modeling of temporal/spectral envelopes with finite-length discrete trigonometric transforms," *IEEE Trans. Signal Process*, vol. 58, pp. 3692 – 37, 2010.

[103] Gregory Sell and Malcolm Slaney, "The information content of demodulated speech," in *Proc. IEEE ICASSP*, Dallas, TX, 2010, pp. 5470–5473.

[104] Gregory Sell and Malcolm Slaney, "Solving demodulation as an optimization problem," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2051 –2066, 2010.

[105] Richard Turner and Maneesh Sahani, "Probabilistic amplitude demodulation," in *Independent Component Analysis and Signal Separation*, Mike Davies, Christopher James, Samer Abdallah, and Mark Plumbley, Eds., vol. 4666 of *Lecture Notes in Computer Science*, pp. 544–551. Springer Berlin / Heidelberg, 2007.

[106] A. J. Oxenham, "Pitch perception and auditory stream segregation: implications for hearing loss and cochlear implants," *Trends Amplif*, vol. 12, no. 4, pp. 316–31, 2008.

[107] Shuo Wang, Li Xu, and Robert Mannell, "Relative contributions of temporal envelope and fine structure cues to lexical tone recognition in hearing-impaired listeners," *Journal of the Association for Research in Otolaryngology*, vol. 12, no. 6, pp. 783–794, 2011.

[108] Fang-Gang Zeng, "Temporal pitch in electric hearing," *Hear. Res.*, vol. 174, pp. 101–106, 2002.

[109] D. Han, B. Liu, N. Zhou, X. Chen, Y. Kong, H. Liu, Y. Zheng, and L. Xu, "Lexical tone perception with hiresolution and hiresolution 120 sound-processing strategies in pediatric mandarin-speaking cochlear implant users," *Ear Hear*, vol. 30, no. 2, pp. 169–77, 2009.

[110] Ward R. Drennan, J. H. Won, K. Nie, E. Jameyson, and Jay T. Rubinstein, "Sensitivity of psychophysical measures to signal processor modifications in cochlear implant users," *Hear Res*, vol. 262, pp. 1–8, 2010.

[111] R. Schatzer, A. Krenmayr, D. K. Au, M. Kals, and C. Zierhofer, "Temporal fine structure in cochlear implants: preliminary speech perception results in cantonese-speaking implant users," *Acta Otolaryngol*, vol. 130, pp. 1031–9, 2010.

[112] D. Riss, J. S. Hamzavi, A. Selberherr, A. Kaider, M. Blineder, V. Starlinger, W. Gstoettner, and C. Arnoldner, "Envelope versus fine structure speech coding strategy: a crossover study," *Otol Neurotol*, vol. 32, no. 7, pp. 1094–101, 2011.

[113] Kaibao Nie, Ginger Stickney, and Fan-Gang Zeng, "Encoding frequency modulation to improve cochlear implant performance in noise," *IEEE Trans Biomed Eng*, vol. 52, no. 1, pp. 64–73, 2005.

[114] Johan Laneau, Jan Wouters, and Marc Moonen, "Improved music perception with explicitly pitch coding in cochlear implants," *Audiol. Neurootol.*, vol. 11, pp. 38–52, 2006.

[115] M. Milczynski, J. E. Chang, J. Wouters, and A. van Wieringen, "Perception of mandarin chinese with cochlear implants using enhanced temporal pitch cues," *Hear Res*, vol. 285, pp. 1–12, 2012.

[116] C. J. Darwin, "Listening to speech in the presence of other sounds," *Philos Trans R Soc Lond B Biol Sci*, vol. 363, no. 1493, pp. 1011–21, 2008.

[117] T. M. Shackleton and R. P. Carlyon, "The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination," *J Acoust Soc Am*, vol. 95, no. 6, pp. 3529–40, 1994.

[118] C. W. Turner, B. J. Gantz, C. Vidal, A. Behrens, and B. A. Henry, "Speech recognition in noise for cochlear implant listeners: benefits of residual acoustic hearing," *J Acoust Soc Am*, vol. 115, no. 4, pp. 1729–35, 2004.

[119] K. Nie, L. Atlas, and J. Rubinstein, "Single sideband encoder for music coding in cochlear implants," in *Proc. ICASSP*. 2008, pp. 4209–4212, IEEE.

[120] C. A. Miller, P. J. Abbas, and J. T. Rubinstein, "An empirically based model of the electrically evoked compound action potential," *Hear Res*, vol. 135, pp. 1–18, 1999.

[121] Q. Li and L. E. Atlas, "Time-variant least squares harmonic modeling," in *Proc. ICASSP*. 2003, pp. II–41–4, IEEE.

[122] N. A. Whitmal, S. F. Poissant, R. L. Freyman, and K. S. Helfer, "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," *J Acoust Soc Am*, vol. 122, no. 4, pp. 2376–88, 2007.

[123] Michael Nilsson, Sigfrid D Soli, and Jean A Sullivan, "Development of the hearing in noise test for the measurment of speech reception thresholds in quiet and noise," *J Acoust Soc Am*, vol. 95, no. 2, pp. 1085–1099, 1994.

[124] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "I.e.e.e. recommended practice for speech quality measurements," *IEEE Trans Aud Electroacoust.*, vol. 17, pp. 227246, 1969.

[125] M F Dorman, P C Loizou, J Fitzke, and Z Tu, "The recognition of sentences in noise by normal-hearing listeners using simultations of cochlear-implant signal processors with 6-20 channels," *J Acoust Soc Am*, vol. 104, no. 6, pp. 3583–3585, 1998.

[126] H. E. Cullington and F. G. Zeng, "Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects," *J Acoust Soc Am*, vol. 123, no. 1, pp. 450–61, 2008.

[127] Li Xu, Yuhjung Tsai, and Bryan E. Pfingst, "Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses," *J Acoust Soc Am*, vol. 112, no. 1, pp. 247–258, 2002.

[128] H. Mino, J. T. Rubinstein, C. A. Miller, and P. J. Abbas, "Effects of electrode–to–fiber distance on temporal neural response with electrical stimulation," *IEEE Trans Biomed Eng*, vol. 51, no. 1, pp. 13–20, 2004.

[129] B. Swanson and H. Mauch, *Nucleus MATLAB Toolbox 4.20 Software User Manual*, Cochlear Ltd, Lane Cove, Australia, 2006.

[130] M. A. Stone, C. Fullgrabe, and B. C. Moore, "Benefit of high-rate envelope cues in vocoder processing: effect of number of channels and spectral region," *J Acoust Soc Am*, vol. 124, no. 4, pp. 2272–82, 2008.

[131] V Summers and M R Leek, "F0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss," *Journal of Speech, Language, and Hear Res*, vol. 41, 1998.

[132] A. E. Vandali and R. J. van Hoesel, "Development of a temporal fundamental frequency coding strategy for cochlear implants," *J Acoust Soc Am*, vol. 129, pp. 4023–36, 2011.

[133] Kate Gfeller, Christopher Turner, Maureen Mehr, George Woodworth, Robert Fearn, John F Knutson, Shelley Witt, and Julie Stordahl, "Recognition of familiar melodies by adult cochlear implant recipients and normal-hearing adults," *Cochlear Implants International*, vol. 3, no. 1, pp. 31–55, 2002.

[134] Ying-Yee Kong, Rachel Cruz, J. Ackland Jones, and Fan-Gang Zeng, "Music perception with temporal cues in acoustic and electric hearing," *Ear Hear.*, vol. 25, pp. 173–185, 2004.

[135] John J. Galvin, Quian-Jie Fu, and Geraldine Nogaki, "Melodic contour identification by cochlear implant listeners," *Ear Hear.*, vol. 28, no. 3, pp. 302–319, 2007.

[136] Y. Y. Kong, A. Mullangi, J. Marozeau, and M. Epstein, "Temporal and spectral cues for musical timbre perception in electric hearing," *J Speech Lang Hear Res*, vol. 54, no. 3, pp. 981–94, 2011.

[137] Hugh J. McDermott, "Music perception with cochlear implants: A review," *Trends Amplif*, vol. 8, no. 2, pp. 49–82, 2004.

[138] Ward R. Drennan and Jay T. Rubinstein, "Music perception in cochlear implant users and its relationship with psychophysical capabilities," *J Rehabil Res Dev*, vol. 45, no. 5, pp. 779–789, 2008.

[139] Blake S. Wilson and Michael F. Dorman, "Cochlear implants: A remarkable past and a brilliant future," *Hear Res*, vol. 242, no. 1-2, pp. 3–21, 2008.

[140] B. C. J. Moore, "Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants," *Otol. Neurotol.*, vol. 24, no. 2, pp. 243–254, 2003.

[141] R. Plomp, "Auditory psychophysics," *Annu Rev Psychol*, vol. 26, pp. 207–32, 1975.

[142] ANSI, "American national standard acoustical terminology," 1994.

[143] J. F. Schouten, "The perception of timbre," in *Sixth Int. Conf. on Acoustics*, 1968, vol. 1, pp. GP–6–2.

[144] J. M. Grey and J. W. Gordon, "Perceptual effects of spectral modifications on musical timbres," *J Acoust Soc Am*, vol. 63, no. 5, pp. 1493–1500, 1978.

[145] A. Caclin, S. McAdams, B. K. Smith, and S. Winsberg, "Acoustic correlates of timbre space dimensions: a confirmatory study using synthetic tones," *J Acoust Soc Am*, vol. 118, no. 1, pp. 471–82, 2005.

[146] Johan Laneau, Jan Wouters, and Marc Moonen, "Improved music perception with explicitly pitch coding in cochlear implants," *Audiol Neurootol*, vol. 11, pp. 38–52, 2006.

[147] M. Milczynski, J. Wouters, and A. van Wieringen, "Improved fundamental frequency coding in cochlear implant signal processing," *J Acoust Soc Am*, vol. 125, pp. 2260–71, 2009.

[148] B. H. Bonham and L. M. Litvak, "Current focusing and steering: modeling, physiology, and psychophysics," *Hear Res*, vol. 242, pp. 141–153, 2008.

[149] A. J. Oxenham, J. G. Bernstein, and H. Penagos, "Correct tonotopic representation is necessary for complex pitch perception," *Proc Natl Acad Sci U S A*, vol. 101, no. 5, pp. 1421–5, 2004.

[150] P. de la Cuadra, A. Master, and C. Sapp, "Efficient pitch detection techniques for interactive music," in *Proc. International Computer Music Conference*. 2001, ICMC.

[151] M. Goto, "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, pp. 311–329, 2004.

[152] C. Von Urff and F. Zonis, "The square-law single-sidebandd system," *IRE Transactions on Communications Systems*, vol. 10, pp. 257–267, 1962.

[153] S. A. Shamma, "Speech processing in the auditory system. ii: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve," *J Acoust Soc Am*, vol. 78, pp. 1622 – 32, 1985.

[154] R. Lyon and S. Shamma, *Auditory Computation*, chapter Auditory representations of timbre and pitch, pp. 221–270, New York: Springer, 1996.

[155] A. Zakhor and A. V. Oppenheim, "Reconstruction of two-dimensional signals from level crossings," *Proceedings of the IEEE*, vol. 78, pp. 31–55, 1990.

[156] J. Swaminathan, *The role of envelope and temporal fine structure in the perception of noise degraded speech*, Ph.D. thesis, University of Purdue, 2010.

[157] C. Lorenzi, L. Debruille, S. Garnier, P. Fleuriot, and B. C. Moore, "Abnormal processing of temporal fine structure in speech for frequencies where absolute thresholds are normal," *J Acoust Soc Am*, vol. 125, pp. 27–30, 2009.

[158] K. S. Henry and M. G. Heinz, "Effects of sensorineural hearing loss on temporal coding of narrowband and broadband signals in the auditory periphery," *Hear Res*, 2013.

VITA

Xing Li was born in Henan, China. She has been interested in science and engineering ever since she was a kid. She received her BSEE from Nanjing University of Posts & Telecommunications, China, in 2004 June. Then she was accepted to the Graduate School of Chinese Academy of Sciences. On finishing her MSEE in 2007 June, she got the opportunity to study at University of Washington, Settle, for a doctoral degree at Electrical Engineering. At the completion of this writing, she finishes her Phd program and will start working at Microsoft, Redmond.