Multiscal Multirate Spectro-Temporal Auditory Model

Powen Ru

Neural Systems Laboratory
University of Maryland College Park
2001

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Hearing is a very important sensory function to many animals including human beings. In recent years, many modern techniques have been employed to explore this function for the benefit of many different applications. For example, people in the speech recognition field want to reduce the error rate so that machines can understand vocal commands in realistic environments. People in audio the coding field want to reduce the bit rate so that sounds need less memory for storage and need less bandwidth for transmission. People in the speech synthesis field want to improve the text-to-speech performance so that machines can speak naturally. The progress of these technologies depends highly on understanding how human auditory systems interpret sounds. "Sound", according to Oxford American Dictionary, is the "vibrations that travel through the air and are detectable by the ear". In other words, only "detectable" sounds need be taken into account. However, current acoustic processing techniques often include limited amount of signal representation and processing principles that are

motivated by human audition, and thus may spend valuable processing effort on data which is not relevant to the human ear.

The three primary elements of sound in terms of human auditory perception are loudness, pitch, and timbre. Though not very precise, it is well accepted that loudness and pitch are correlated to the magnitude and frequency, respectively, of the acoustic waveform. However, timbre still retains its mystical nature. Even American National Standards Institute (ANSI) is unable to give it a clear definition. Recently, more and more experiments have shown that the shape of the acoustic spectrum is a fundamental cue to the timbre perception of complex sounds (Plomp, 1976) [59]. Unlike loudness and pitch, which are of little importance in speech intelligibility, timbre plays a crucial role in automatic speech processing – such as speech recognition and speech coding. Therefore, understanding the encoding of acoustic spectra is essential for the improvement of high performance sound processing model.

The human auditory system boasts remarkable abilities to detect, separate, and recognize speech, music, and other environmental sounds. Particularly with a view towards applying auditory functional principles to the design and implementation of a man-computer communication link, these capabilities have been the subject of theoretical investigation in recent decades. Since human performance surpasses the performance of automatic speech recognition (ASR), understanding the basic principles behind human speech recognition (HSR) should be a promising approach (Allen, 1994) [3].

The speech signal is a mean by which linguistic information is carried from one destination to another. It is generated by the speech production organs of the speaker with the purpose of being processed by the auditory system of the

listener. Therefore a sound processing model that resembles human auditory system should outperform other signal processing models. As such, it is reasonable to believe that signals are best represented in terms of sensory features. Similarly, the measures of signal quality and the optimal criterion for signal processing tasks should also be perception-oriented.

The investigation of the function of the central auditory system is still a relative young science (Kowalski *et al.*, 1996 [42] [43]; Shamma *et al.*, 1993 [75]; Schreiner and Calhoun, 1994 [69]; Eggermont, 1994 [21]). Wang (1995) has developed a mathematically tractable model to simulate the multi-scale processing found in the auditory cortex [95]. This model is constructed with three axes, namely, tonotopic, scale, and symmetry axis. The tonotopic axis, which arises from the peripheral auditory system, is well understood both physiologically and perceptually. The other axes are newly introduced hence little is known about their psychoacoustic properties. According to Wang's interpretation, important sound aspects like pitch and timbre are well represented in the cortical model. However, further practical applications were still under development prior to this study. Thus, an evaluation is necessary to verify this physiologically-driven auditory model.

1.2 Approach

The goal of this study is to construct a comprehensive auditory model based both on mammalian physiology and human perception. The model is supposed to perform physiological operations and yield useful representations for higher-level processing stages. Thus, each component should be designed to match

available physiological findings. Of course, psychoacoustic data should be taken into account to test and refine the model. Ultimately, the experimental data should be explained by the auditory model. From the viewpoint of evaluation, the model should outperform other conventional approaches in most auditory-related applications.

Sound evokes complex patterns of activity in the peripheral auditory system. This activity codes for the sound in a way that is meaningful to the central nervous system (CNS). The multidimensionality of the percepts must involve a large number of parameters. The coding of these parameters in the responses of the auditory nerve has been a central theme in the neurophysiology of the peripheral auditory system (Sachs and Young, 1979 [67]; Delgutte and Kiang, 1984 [17]). As a result, several response properties have emerged as potential cues from which the CNS may derive the appropriate percepts. These properties include the temporal periodicities, average firing rates, and the distribution of firing rates across the auditory-nerve fiber array.

The complex patterns of activity on the auditory nerve have two nominal aspects: *spatial* in that different tones excite fibers which innervate different cochlear regions, and *temporal* in that responses of fibers to low-frequency tones (less than about 4 kHz) tend to be phase-locked to the waveform of the driving stimulus. Phase locking diminishes at higher frequencies, but fibers may still lock to the envelope modulations due to several interfering harmonics within the bandwidth of the fiber. The spectrotemporal response patterns of the tonotopically ordered auditory-nerve-fiber array are then projected to the CNS, where various neural networks perform further analysis. The primary auditory field (AI) of auditory cortex has been identified in almost all mammals studied. A

fundamental goal in auditory cortical physiology has been to understand how the spectral profile is represented in the firing rate of cortical cells, or, equivalently, how might one predict the responses of a single unit to arbitrary spectral profiles. The coding of sounds on the auditory nerve involves a multitude of spatial and temporal cues. As far as the CNS is concerned, the worth of any cue is ultimately determined by whether it is biologically feasible to utilize it. Most contemporary sound processing models are based on some digital signal processing techniques. This implies that valuable biological behavior is ignored while some redundant information is considered.

It is natural to explore whether the recently discovered response mapping in A1 may be integrated within a functional framework. In building a sound processing system, we make the axiomatic assumption that an acoustical signal can be decomposed to many single unit responses. Recent physiological findings show that the cells in the primary auditory cortex do have selectivity to certain scale-frequency characteristics of acoustical signal. The computation of the analytical process is very similar to the wavelet transform which is a hot topic in the contemporary signal processing world. The wavelet transform is an attractive signal processing technique primarily because of its multi-resolution nature. However, it is still unclear how the central auditory system integrates those responses into useful information.

To put the physiological theory in perspective, psychoacoustic experiments must be conducted to attack this question. Psychoacoustics could be thought as an open-ended science of the human hearing. Though the results from physiological experiments are more reliable than those from psychoacoustic experiments, they merely provide a microscopic view of the auditory system. Moreover, the

physiological approach is hard to apply on human subjects. Thus, psychoacoustics offers a better approach to get a macroscopic view of human auditory systems. Conducting threshold-measuring experiments in this study serves two purposes: first, the resolution of each of cortical axes can be determined; second, the perceptual distance between two arbitrary complex sounds may be predicted based on these experimental results.

Human auditory system is a very complicated structure. In most cases, it receives and interprets the sound produced by the vocal system. It possesses a remarkable ability to recognize sounds at phonetic level. Either the waveform or the spectrum of a phoneme demonstrates wide-range discrepancies due to genders, accents, and emotions. However, the auditory systems are good enough to normalize those features and allow the brain to extract the necessary information. To the contrary, the auditory system can even make use of those redundant features for some other purposes like speaker identification. Therefore, there must exist some relationship between the production system and the perception system.

This multi-disciplined work, driven by physiological findings, involves signal processing techniques, psychophysical methods, and statistical modeling. Wang and Shamma's cochlear model (1994) [94], after some modification, is used as the peripheral auditory model. Their cortical model (1995) [95] is the static part of the newly developed spectrotemporal model. Most psychoacoustical procedures are similar to those in Hillier's ripple detection experiments (1991) [38] and Vranic-Sowers' peak profile experiments (1991) [92]. Their results were also used to test the model. Acoustic tube theory (Temkin, 1981) [87] and Ehrenfest's perturbation theorem (Schroeder, 1967) [71] are employed to compute or predict

the frequency response of the vocal tract. Finally, vowel recognition and musical instrument classification were employed to evaluate the cortical model.

1.3 Overview

An overview of human speech processing pathway is presented in Figure 1.1. The vocal tract of the speaker modified (filtered) the pulses or the noise excited by the glottis. Then, the sound pressure radiated from the lips is transmitted to the listener's ears. The peripheral auditory system transduces this vibration into neural spikes. This action, in turn, changes the evoked potential on the surface of the brain. The block diagram illustrates the backbone of this study which will be elaborated in next few paragraphs.

In Chapter 2, a physiologically-driven auditory model is proposed, which is an extension of Wang and Shamma's cortical model (1995) [95]. Recently available physiological findings on the dynamic auditory processing were used to build this extended cortical model. The model, simulating the spectrotemporal processing of AI, carries out cortical functions like rate-scale tuning and directional selectivity. The auditory pathway is modeled with three levels: the acoustic level (time waveforms), the cochlear level (frequency spectra) and the cortical level (rate-scale representations). The reconstruction for each level is also explicitly discussed.

Chapter 3 describes a series of psychoacoustic experiments from the design stage to the analysis stage. The stimuli were carefully designed to excite desired cells. Ten subjects participated this project. The two-down-one-up (2D1U) method associated with two-alternative-forced-choice (2AFC) paradigm was em-

Figure 1.1: Speech processing pathway: articulatory model, auditory model, and AI cortical model.

ployed to measure subjective thresholds. The thresholds measured under numerous conditions and the distance due to different models are reported. The selected threshold predictions are also given. The emphasis of this chapter lies primarily with the static profile analysis. A preliminary moving ripple experiment is also given for future interest.

In Chapter 4, a simple articulatory model is proposed to seek the correspondence between the speech production system and the perception system. It was found that the articulatory equivalence of cortical translation is stretching/compressing the area function of an acoustic tube. The quasi-equivalence of cortical dilation was also found through more complicated manipulations. A novel vowel space is also given based on perturbation of the area function.

In Chapter 5, some applications or examples are given to evaluate this perception-based auditory model. Vowel recognition experiment was conducted to evaluate the representations. The representations were applied in music to quantify the distances among different instruments.

The conclusion is given in Chapter 6. Some valuable information is also available in the appendices. Appendix A collects numerous auditory-related physiological facts and contains a glossary covering the relevant terms. A description on TIMIT speech corpus, one of the most frequently used databases, is also included. In Appendix B, detail concerning the design of the cochlear model is given. The the filter bank and the compressive function are described in depth. The design of the cortical filter array is explained in Appendix C.

Chapter 2

The Multiscale Multirate Auditory Model

2.1 Introduction

Natural sounds like music and speech are usually characterized by their loudness, pitch, timbre, and onset/offset instants. These descriptions of sound quality have a close relationship to the instantaneous spectral properties of the sound waves. However, the ears are not capable of explicitly computing the Fourier transform to obtain frequency spectra. Instead, a spectrum-like pattern called the *auditory spectrum* is extracted through a series of linear and nonlinear processes.

Physiological, psychoacoustical, and computational studies reveal that the central auditory system has developed an elegant mechanism to extract and represent this spectrotemporal information; the primary auditory cortex (AI) employs a multiscale representation in which the dynamic spectrum is repeatedly represented in AI at various degrees of spectral and temporal resolution. This is accomplished by cells whose responses are selective to a range of spectrotemporal parameters such as the local bandwidth and symmetry of spectral peaks,

and their onset and offset transition rates. Such a representation provides a quantitative description of timbre (quality of sound), and hence can serve as the front-end for higher level processing. Moreover, it may underlie many important "perceptual invariances" such as the ability to recognize speech and melodies despite large change in rate of delivery (Jules and Hirsch, 1972) [41], or the ability to perceive continuous music and speech through gaps, noise, and other short duration interruptions in the acoustical stream. Furthermore, the segregation into different rate-scales such as fast and slow corresponds to the intuitive classification of many natural sounds and music as transient or sustaining; of speech as stop or continuant.

The model consists of two stages, viz., a spectral estimation model and a spectral analysis model. The spectral estimation model was designed to mimic the *cochlea* in the peripheral auditory system. The relevant physiological background and the mathematical formulation are elaborated in Section 2.2. The spectral analysis model mimics the multiscale nature of the primary auditory cortex. The cortical representation is presented in Section 2.2.3. This representation reveals the local bandwidth and the local symmetry of the auditory spectrum. In Section 2.4, dynamic processing is incorporated into an extended model to simulate the spectrotemporal characteristics of cortical responses. This extended cortical model thus analyzes the auditory spectrogram both at different rates and different scales. The composite characteristic response exhibits the directional selectivity which has been discovered by Kowalski *et al.*(1996) in their surgeries [42]. The reconstruction procedure for each stage is also presented in the corresponding section. The above models, driven by the physiological findings, were integrated into a complete multi-resolution model as shown in

Section 2.5 (Figure 2.18).

2.2 Spectral Estimation Model - The Cochlear Model

2.2.1 Peripheral Auditory System

When sound waves impinge upon the eardrum of the outer ear, they cause vibrations which are transmitted via the middle ear to the fluid of the *cochlea* in the inner ear. Consequently, the vibrations produce mechanical displacements on the *basilar membrane*. When evoked by a single tone, the vibrations appear as traveling waves that propagate up to the cochlea from base to apex, reaching a maximum amplitude at a particular point before decaying rapidly. The point at which maximum displacement occurs depends on the frequency of the tone, the lower frequencies propagating further towards the apex of the cochlea. As such, the cochlea segregates incoming frequencies onto different spatial locations in a tonotopically ordered manner along its length. At each point along the membrane, one can measure the displacement as a function of the tone frequency, i.e., a transfer function. In mammalian cochleas, the transfer functions are moderately well tuned, with characteristic frequencies decreasing towards the apex of the cochlea. In humans, above 800 Hz or so, the transfer functions of these "cochlear filters" are roughly related to each other by a dilation. Consequently, along the logarithmic frequency axis, those transfer functions appear approximately invariant except for a translation.

The mechanical vibrations along the basilar membrane are transduced into

electrical activity along a dense, topographically ordered array of auditory nerve fibers. At each point, the membrane displacement causes a local fluid flow which bends *cilia* that are attached to *inner hair cells*. The bending controls the flow of ionic currents through nonlinear channels into the hair cells. The ionic flow then generates electrical potentials across the hair cell membranes. Finally, these electrical potentials are conveyed by the auditory nerve fibers to the cochlear nucleus. Recipient neurons in the cochlear nucleus then reconstruct estimates of the hair cell potentials from the ensemble averages of activity in locally adjacent fibers. In the auditory nerve, the dynamic range between threshold and saturation of activity in a given fiber is limited to $30 \sim 40$ dB (Sachs and Young, 1979) [67]. Temporal fluctuations in any given fiber are limited to frequencies below $4 \sim 5$ kHz due to the low-pass effect of the hair cell membranes. Above these frequencies, the auditory nerve indicates the presence of a particular frequency by a steady increase in the firing rate. The anteroventral cochlear nucleus receives direct input from the auditory nerve and exhibits lateral inhibition.

2.2.2 The Cochlear Model and Auditory Spectrogram

The cochlear model is composed of three major stages, viz., analysis, transduction and reduction (Yang *et al.*, 1992 [98]; Wang and Shamma, 1994 [94]). It can be formulated as following equations:

$$y_1(t,x) = s(t) *_t h(t;x) \tag{2.1}$$

$$y_2(t,x) = g(\partial_t y_1(t,x)) *_t w(t) \tag{2.2}$$

$$y_3(t,x) = \partial_x y_2(t,x) *_x \nu(x) \tag{2.3}$$

$$y_4(t,x) = \max(y_3(t,x), 0) \tag{2.4}$$

$$y_5(t,x) \;=\; y_4(t,x) *_t \mu(t;\tau) \tag{2.5}$$

where x represents the spatial location away from the base of the cochlea. The position-frequency relation is modeled by $x = \log_2 f/f_0$ in *octaves* relative to f_0 Hz. For instance, f Hz is mapped to $x = \log_2(f/1000)$ octaves relative to 1 kHz. This logarithmically transformed frequency is often referred as *tonotopic frequency*. The actual frequency scale of the cochlea is not purely logarithmic for the frequencies below 800 Hz, but rather becomes progressively more linear, especially below 500 Hz (see Section B.1). However, the logarithmic warping is still a fair and simple approximation.

In the *analysis* stage, Eq. (2.1) models the basilar membrane response for a sound signal $s(t)$, where $h(t;x)$ denotes the impulse response of the filter located at x and $*_t$ denotes the convolution in the time domain. In this model, the frequency responses of the filter bank were obtained by dilating the response of a seed band-pass filter, i.e., $H(f;x) = H(f/a;x_0)$, or $h(t;x) = ah(at;x_0)$ in time domain, where the scaling factor $a = 2^{x-x0}$. On the tonotopic axis, the dilation relationship is transformed to a translation so that and all of the filters share one common shape. In the context of signal processing, this analysis scheme is called the *constant-Q filter-bank wavelet transform* (Fliege, 1994) [29]. According to the stochastic analysis in Wang and Shamma (1994), the exact shape of the seed function is not that important [94]. The magnitude of the cochlear filter resembles the measured cochlear filter shape which is a highly asymmetric band-pass filter with moderate slope ($6 \sim 12$ dB/oct) in the low frequency side and much steeper slope ($-50 \sim -500$ dB/oct) in the high frequency side (Allen, 1985) [2]. As for the phase, since the response is essentially of finite duration, a reasonable choice is to make the filter of minimum-phase.

Figure 2.1: Block diagram of the cochlear model.

In the *transduction* stage, Eq. (2.2) models the hair cell response which incorporates fluid-cilia coupling, compressive ionic channels, and membrane leakage. The fluid-cilia coupling is described by a temporal derivative which is computationally equivalent to the so called *preemphasis* (Rabinar and Juang, 1993) [61] on the incoming signal. The nonlinear channel through the hair cell is then modeled by a sigmoid-like function $g(\cdot)$ (see Section B.3). Finally, the leakage of the cell membrane is accounted for by a low-pass filter $w(t)$ which filters out all responses beyond 4 kHz. However, this filter plays a minor role for low frequencies, where most of speech energy resides, and will thus be *ignored.* Due to the linearity of the differentiation, two operations in this stage can be moved to other stage to simplify computation. First, the partial derivative with respect to time axis (i.e., ∂_t) can be directly applied to the incoming acoustic signal $s(t)$. Second, the temporal smoothing function $w(t)$ can be associated to next stage.

In the *reduction* stage, the lateral inhibitory network (LIN) (Shamma, 1985 [72]; Shamma, 1989 [73]) was divided into three steps: tonotopic derivative (Eq. (2.3)), half-wave rectification (Eq. (2.4)), and leaky integration (Eq. (2.5)). The derivative ∂_x simulates the lateral interaction among LIN neurons. The spatial filter $\nu(\cdot)$ models the local smoothing due to the finite spatial extent of the lateral interactions. However, this smoothing may be *ignored* since central auditory system provides more significant smoothing. The half-wave rectifier $\max(\cdot, 0)$ mimics the positive nature of the LIN neurons. Finally, a temporal integration window $\mu(t;\tau) = e^{-t/\tau}u(t)$ is applied to model the slow adaptation of central auditory neurons. Here τ is the time constant and $u(t)$ is the *unit step* function. Unlike the auditory nerve fibers, the central auditory neurons cannot follow rapid modulation higher than 1 kHz. However, τ can be anywhere

between 1 to 128 ms depending on the intended destination of the signal. For example, the neurons in cat auditory cortex respond only in the range from 3 to 26 Hz (Eggermont, 1994) [21] whereas those in cat inferior colliculus can follow modulation rate up to 1 kHz (Langner and Schreiner, 1988) [45]. The time-frequency representation $y_5(t,x)$ is called the *auditory spectrogram* throughout this work. At a given time instant, $y_5(x)$ is called the *auditory spectrum* (Yang *et al.*, 1992) [98].

After ignoring $w(t)$ and $\nu(x)$, the whole model is summarized as

$$y_5(t,x) = \max(\partial_x g(\partial_t s(t) *_x h(t,x)), 0) *_t \mu(t;\tau) \tag{2.6}$$

The block diagram is given in the upper part of Figure 2.1. The signal flow is depicted in solid lines of which the *thin* lines represent 1-D flow and the *thick* lines represent 2-D flow. The bottleneck of the entire process is the cochlear filter bank. This problem can be mediated by following means. For software applications on wide-band audible signal, the transfer function of the filter can be approximated by IIR filter coefficients (see Section B.2). For hardware usage, the cochlea filter bank can be implemented using analog VLSI technology (Lin *et al.*, 1994) [48]. The auditory spectrogram is much smoother (in time domain) than the original signal and can therefore be further downsampled to reduce data rate. The sample period is of the order of time constant τ, e.g., if the sample period is chosen as τ then

$$y_5[n,x] = y_5(n\tau, x) \tag{2.7}$$

Typically, for speech signal sampled at 8 kHz, the appropriate spectrum sampling period is 16 ms. Figure 2.2-(b) depicts an auditory spectrogram induced by the acoustic waveform of a sentence ("Come home right away") spoken by

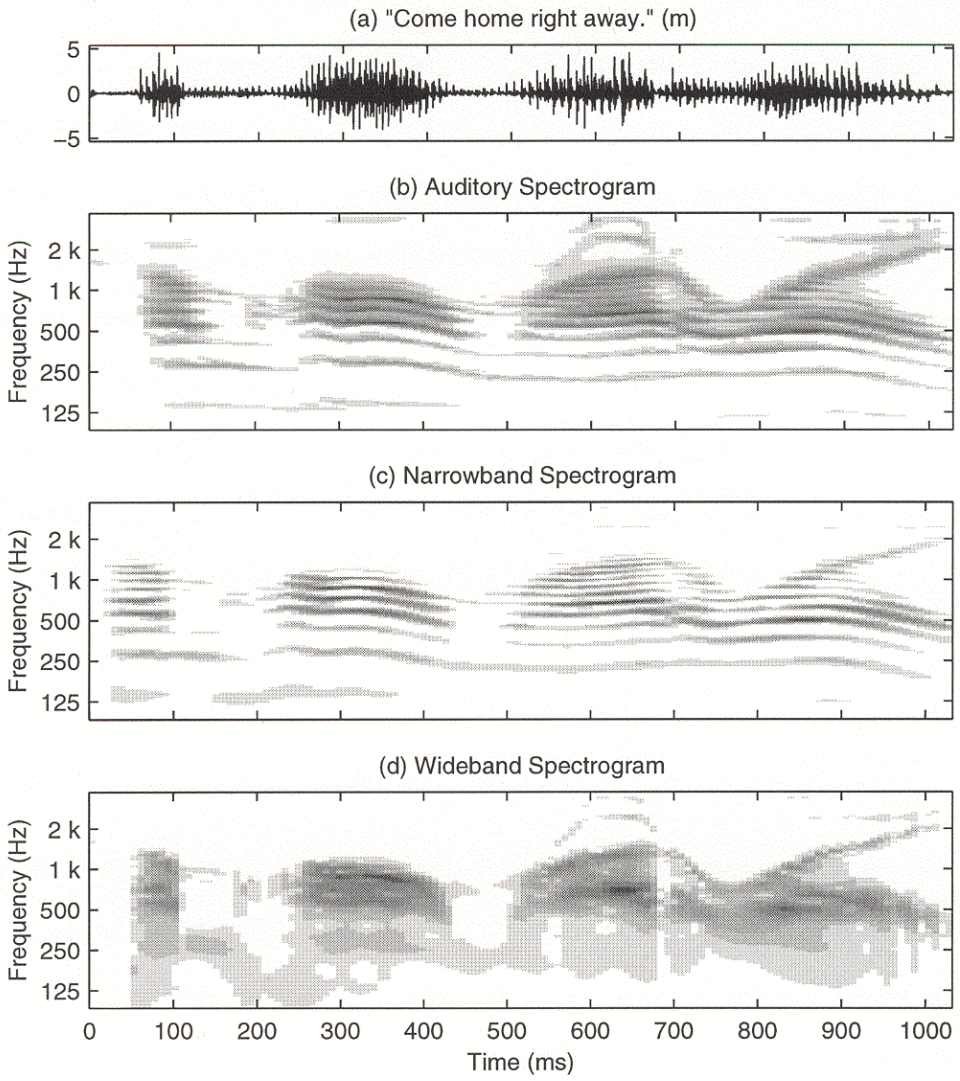

Figure 2.2: (a) The acoustic waveform of a sentence spoken by a male speaker; (b) the corresponding auditory spectrogram; (c) the corresponding narrow-band spectrogram; (d) the corresponding wide-band spectrogram;

a male speaker (shown in Figure 2.2-(a)). In addition, both narrow-band and wide-band spectrograms are shown (Figure 2.2-(c) and (d)). It is well known that the narrow-band spectrogram, using longer time-windowing, displays the fine structure of the spectrum. Thus it amends the extraction of harmonic information. On the other hand, the wide-band spectrogram shows the global shape of the spectrum and hence is employed to highlight the formant traces. However, the two sound features are not well represented in one plot due to the limitation of time-frequency resolution. Compared to the conventional spectrogram, the auditory spectrogram is better for showing both features. In the low frequency range ($100 \sim 500$ Hz), the harmonics appear clearly. In the medium frequency range ($300 \sim 2000$ Hz), the formant traces show up. In the higher frequency range (> 2000 Hz), only the energy distribution can be seen.

2.2.3 Properties and Reconstruction

Self-normalization and Noise-robustness

For a weakly stationary sound source, the auditory spectrum, $y(t, x)$, is a kind of normalized power spectrum. The normalization is driven by the energy distribution of the signal. For each channel x, the output reflects approximately the ratio of the energy of its differential filter to that of its cochlear filter. A spectral peak resolved by the differential filter receives a smaller normalization than the spectral valley. In effect, this difference enhances the peak-to-valley ratio. Thus the auditory spectrum is relatively insensitive to broadband changes in the spectral shape as long as the responses from the differential and cochlear filters are affected similarly. This characteristic effectively limits the dynamic range of the spectra while preserving the peak-to-valley ratio. As a consequence,

the auditory spectra are more robust against noise, preemphasis, and any global spectral tilt. The reader can refer Wang and Shamma (1994) [94] for examples.

Reconstruction

Despite the nonlinearities and temporal integration used to produce the auditory spectrum, a reconstruction of the original signal is still possible. There are several reasons for doing this. First, it is important to show that little information was lost since a fairly accurate reconstruction is still possible. Second, one may need to process or manipulate signals in the spectral domain and then playback the corresponding sounds. Also if the auditory spectrum is invertible, it can then be applied to audio coding, e.g., codec (coder-decoder).

In the analysis stage, the filter bank operation is absolutely linear. Therefore, ideally, the original signal can be perfectly reconstructed from the output of filter bank by means of reverse filtering (Akansu and Haddad, 1992) [1] based on following deductions:

$$Y_1(f,x) = S(f)H(f;x) \tag{2.8}$$

$$\Rightarrow \quad \textstyle\sum_x Y_1(f,x)H^*(f;x) = S(f)\sum_x H(f;x)H^*(f;x) \tag{2.9}$$

$$\Rightarrow \quad S(f) = \textstyle\sum_x Y_1(f,x)H^*(f;x) / \sum_x H(f;x)H^*(f;x) \tag{2.10}$$

where Y_1, S, H are the Fourier transforms of y_1, s, h, respectively. However, this reconstruction has a potential danger. Due to the band-pass nature of the filter-bank, the overall response $\sum_x H(f;x)H^*(f;x)$ resembles a broad band-pass filter so that the gain at both the low- and high-frequency skirts is relatively small. Thus, after reconstruction, any noise or numerical error occurring in those regions will be magnified significantly. To avoid this adverse effect, one

may simply ignore the response at both ends. Most natural sounds are essentially of zero-mean and the human auditory system is less sensitive to the frequencies beyond 4 kHz hence the above simplification will not result in serious perceptual distortion. For better and more efficient reconstruction, the transfer functions of the filter-bank can be weighted, i.e.,

$$H_1(f;x) = w(x)H(f;x) \tag{2.11}$$

where $w(x)$ is a weighting function, such that the overall response is almost unitary within the effective band, i.e.,

$$\sum_x |H(f;x)|^2 w(x) \simeq 1 \tag{2.12}$$

Thus the time waveform $\tilde{s}(t)$ can be synthesized from the projected filter bank response $\tilde{y_1}(t,x)$

$$\tilde{S}(f) = \sum_x \tilde{Y}_1(f,x)H_1^*(f;x) \tag{2.13}$$

$$\tilde{s}(t) = \sum_x \tilde{y}_1(t,x) *_t h_1(-t;x) \tag{2.14}$$

The above operation is referred to the *inverse wavelet transform* (IWT) in the context of signal processing. In the final two stages, two nonlinear operations (i.e., $g(\cdot)$ and $\max(\cdot,0)$) and a severe downsampling (i.e., $t \to n$) are involved, so that the reconstruction from $y_5[n,x]$ back to $y_1(t,x)$ is impossible to obtain directly. Our approach is to apply an iterative method similar to the *convex projection* in Yang *et al.*(1992) [98]. The basic assumption is that the auditory spectrogram $y_5[n,x]$ roughly reflects the local time-frequency (TF) energy distribution. Therefore the guessed $\tilde{y}_1(t,x)$ can be adjusted by the ratio of the target $y_5[n,x]$ over the computed spectrogram $\tilde{y}_5[n,x]$ corresponding to $\tilde{y}_1(t,x)$.

The iteration is summarized as follows (see also the dashed flow in Figure 2.1).

1. Generate unitary white noise, i.e., $\tilde{s}(t) \sim \mathcal{N}(0,1)$ where $\mathcal{N}(0,1)$ denotes the Gaussian distribution with zero-mean and variance 1.
2. Compute $\tilde{y}_1(t,x)$ through $\tilde{y}_5[n,x]$ with respect to $\tilde{s}(t)$.
3. Find the ratio $r[n,x]$ between the target $y_5[n,x]$ and the reconstruction $\tilde{y}_5[n,x]$ (TF comparator). If a specific $\tilde{y}_5[n,x]$ is zero while the $y_5[n,x]$ is not zero, then the corresponding $r[n,x]$ is assigned to be 2.
4. Interpolate $r[n,x]$ to $r(t,x)$ and then use it to scale the filter-bank response, i.e., $\tilde{y}_1(t,x) \leftarrow r(t,x)\tilde{y}_1(t,x)$ (TF adjustor).
5. Reconstruct the time waveform by means of the IWT, i.e., $\tilde{s}(t) = \sum_x \tilde{y}_1(t,x) *_t h_1(-t;x)$.
6. Go to step 2.

A few reconstruction examples are given in Figure 2.3. The first example (Figure 2.3-(a)) is for a sentence spoken by male while the second one (Figure 2.3-(b)) was spoken by a female. The original waveform is given in the uppermost panel followed by its corresponding auditory spectrogram (target). The reconstructed waveform due to the spectrogram is given in the third panel. The resulting auditory spectrogram is also given in comparison with the target spectrogram. In both cases, the errors drop below 3% after about 30 iterations. As the reader can see, the auditory spectrogram of the reconstructed waveform is very close to the original and the temporal envelope is well preserved in the reconstructed waveforms. Though reconstructed waveform is not as clean as the original, an informal hearing test shows that the speech intelligibility is fair, as the reconstructed sentences were well understood.

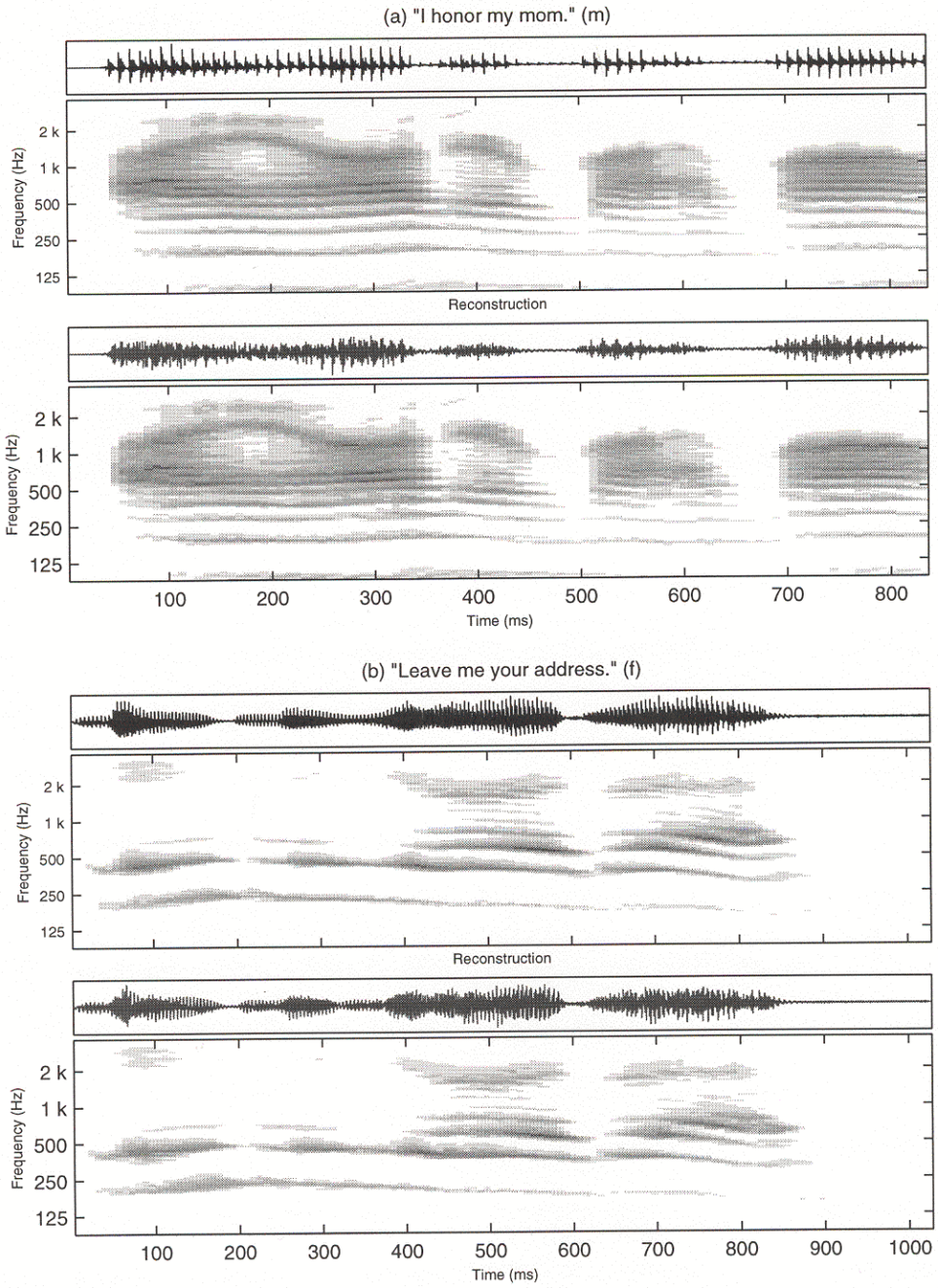

Figure 2.3: Examples of the reconstruction from the auditory spectrogram.

2.3 Spectral Analysis Model - The Cortical Model

2.3.1 Central Auditory System - Static Processing

Generally speaking, the peripheral auditory system acts like a frequency analyzer which estimates the spectrum of the incoming acoustic signal. In the early stages of auditory processing, an enhanced spectral representation is extracted in a series of well understood operations (Wang and Shamma, 1994) [94]. Based on many physiological findings, Wang and Shamma (1995) suggested that the central auditory system serves as the spectral shape analyzer which expands a one-dimensional spectrum into a three-dimensional representation [95]. These dimensions are the *tonotopic axis*, *scale* (local bandwidth), and *symmetry* (local phase). Recently available physiological findings (Schreiner and Calhoun, 1994 [69]; Shamma *et al.*, 1995 [76]; Versnel and Shamma, 1995 [90]) also support the multi-dimensional functionality of the primary auditory cortex.

The aforementioned tonotopic axis is preserved through several central processing stages (Webster, 1992) all the way up to the auditory cortex [97]. However, unlike its essentially one-dimensional nature in the cochlea, the tonotopic axis becomes two-dimensional in AI; many cells tuned to similar frequencies are lined up along the *iso-frequency* planes which run perpendicular to the tonotopic axis (Merzenich *et al.*, 1975 [52]). This suggests that additional features of the auditory spectral pattern are perhaps explicitly analyzed and mapped out along the iso-frequency axis (Heil *et al.*, 1992 [36]; Schreiner and Mendelson, 1990 [70]; Versnel *et al.*, 1995 [90]). Such an analysis occurs in other sensory systems and has been a strong motivation toward the search for auditory analogue. For

instance, an image induces retinal response patterns that roughly preserve the form of the image. The representation, however, becomes much more elaborate in the primary visual cortex (VI), where edges with different orientations, symmetries, and widths are extracted and neurally represented (De Valois and De Valois, 1990 [16]).

It has been reported that cortical cells exhibit a systematic change in the symmetry of their tuning curves (Shamma *et al.*, 1993) [75]. When tested with single tones, neurons along the auditory pathway are found to be selective to a range of frequencies around a BF. Within this range, responses change from excitatory to inhibitory in a pattern that varies from one cell to another in its bandwidth and symmetry around the BF. This response pattern is usually called the *receptive field* (RF) of the neuron. When a broadband signal is used as a stimulus, the cell's response can be thought of as the net effect of all excitatory and inhibitory influences induced by the portion of the spectrum which lies within its RF. Interestingly, the experimental data show that in the center of AI, the RF has a centered excitatory band that is symmetrically flanked by inhibitory side bands. Towards the edges AI, the response area becomes more asymmetric with stronger inhibitory side bands above BF in one direction, and below BF in the opposite direction. Another physiological finding is that cells along the iso-frequency planes vary considerably and systematically in the bandwidth of their tuning (Shamma *et al.*, 1993) [75]. Specifically, neurons in the center of AI are more narrowly tuned compared to those near the edges. Since the notion of the spectral asymmetry is only meaningful within the bandwidth of the neuron's response area, its evaluation must be regarded as a local, and presumably a multiscale operation. From a functional point of view, this implies that the

auditory system may employ a multiscale mechanism to analyze the auditory spectrum, and each scale resolves and extracts information encoded in a specific bandwidth.

Figure 2.4 displays some typical RFs, which were collected from numerous experiments, and their approximations. The process to extract RFs is described in Shamma *et al.*(1995) [77]. These experimental data suggest the multiscale nature of the auditory processing in the primary cortex. The RFs can be fitted by the second derivative of a Gaussian function (see the dashed curves in Figure 2.4 and Section C for details).

$$h(x) = \Omega(1 - 2(\pi\Omega x)^2)e^{-(\pi\Omega x)^2} \tag{2.15}$$

To unify the terminology, a couple of parameters regarding the shape of the receptive field are defined as follows. *Scale* is a quantity to describe the tuning range of receptive fields which can be expressed in terms of envelope (ripple) frequency in *cycles/octave* (for short, in *cyc/oct*; or even shorter, in *c/o*). A schematic example is depicted in Figure 2.5. *Symmetry* represents the shape of the response field which is determined by the strength relationship among the excitatory band and two lateral inhibitory bands. The symmetry of an even (odd) function is 0 ($-90°$). An arbitrary symmetry can be obtained by taking sinusoidal interpolation between the even and the odd function.

Extensive measurements of such RFs have been carried out in AI using a variety of stimuli (Clarey *et al.*, 1992 [14]; Shamma, 1995 [74]). Most directly relevants are those employing broadband spectra with rippled envelopes (Schreiner and Calhoun, 1994 [69]; Shamma *et al.*, 1995 [77]). In these experiments, RF measurements are based on a fundamental assumption that AI responses are essentially linear. To first order, AI responses to broadband spectra are linear in

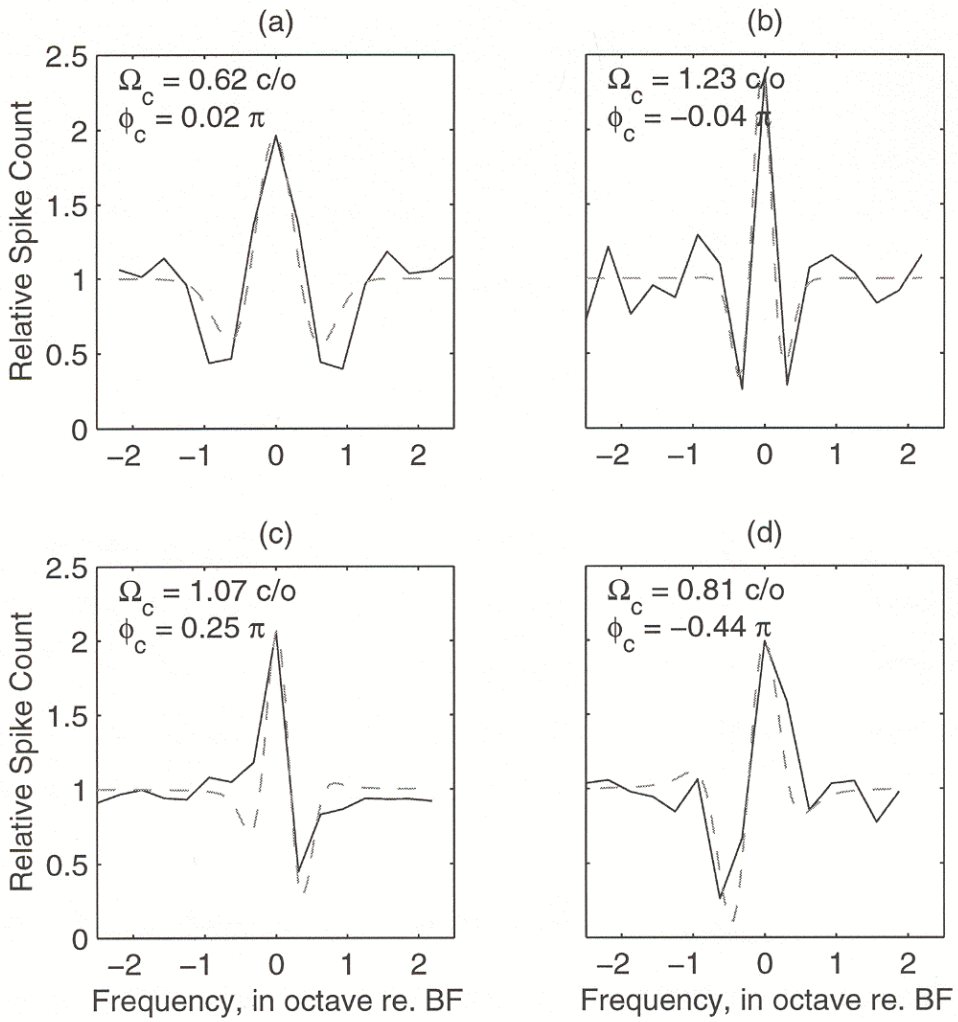

Figure 2.4: Four typical receptive fields show the variety of shape, viz., (a) symmetric shape with wide bandwidth; (b) symmetric shape with narrow bandwidth; (c) asymmetric shape with positive phase; (d) asymmetric shape with negative phase. The dashed curves are the approximations to fit the negative second derivative of a Gaussian distribution density function.

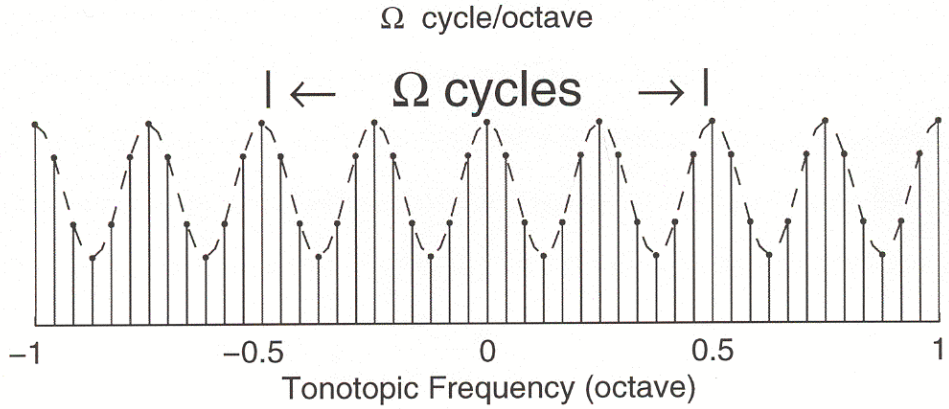

Figure 2.5: The spectrum of a Ω-cycle/octave ripple stimulus

the sense that they satisfy the superposition principle. The linearity in AI has been confirmed in a large number of tests involving spectral profiles composed up to 10 superimposed spectra (Shamma and Versnel, 1995 [76]; Shamma *et al.*, 1995 [77]). Linearity is a powerful simplifying principle that allows one to predict the responses to any arbitrary spectral profile. Thus, a computational model can be built up based on the single unit response.

2.3.2 Pure Spatial Processing Model

The basic operation of the cortical model is a wavelet analysis of the spectral profile (Wang and Shamma, 1995 [95]). On the tonotopic axis, the RF is modeled as

$$\mathcal{RF}(x - x_c; \Omega_c, \phi_c) = h(x - x_c; \Omega_c) \cos\phi_c - \hat{h}(x - x_c; \Omega_c) \sin\phi_c \qquad (2.16)$$

where x denotes the tonotopic frequency in octaves, the seed function $h(x; \Omega_c)$ is a real even function with peak spatial frequency response at Ω_c (in cycle/octave), x_c is the center frequency (in octaves re. 1 kHz), ϕ_c is the characteristic phase

(in radians) which determines the symmetry of the receptive field and $\hat{h}$ denotes the Hilbert transform of the function h. The exact shape of this even function is not important as long as it can manifest the lateral inhibition structure, i.e., a central excitatory band symmetrically flanked by inhibitory side bands. The response of the cell tuned to (x_c, Ω_c, ϕ_c) for an input auditory spectrum $y(x)$ is

$$r(x_c, \Omega_c, \phi_c) =< \mathcal{RF}(x; x_c, \Omega_c, \phi_c), y(x) > \tag{2.17}$$

where $< \cdot, \cdot >$ denotes the inner product of two vectors. The inner product process is equivalent to a filtering process with the spatial impulse response

$$h_{\mathcal{RF}}(x; \Omega_c, \phi_c) = h(x; \Omega_c) \cos \phi_c + \hat{h}(x; \Omega_c) \sin \phi_c \tag{2.18}$$

Thus the response can be efficiently computed by the following convolution:

$$r(x_c, \Omega_c, \phi_c) = y(x) * h_{\mathcal{RF}}(x; \Omega_c, \phi_c)|_{x=x_c} \tag{2.19}$$

The spatial impulse response of the cortical filter with scale Ω_c can be equivalently expressed in the wavelet-based analytical form given by

$$h_w(x; \Omega_c) = h(x; \Omega_c) + j\hat{h}(x; \Omega_c) \tag{2.20}$$

where $h(x; \Omega_c)$ is a symmetric function and $\hat{\cdot}$ denotes Hilbert transformation. The spatial impulse responses for different scales are related by dilation, i.e., $h(x; \Omega_c) = \Omega_c h(x; 1)$. Section C.1 presents some possible candidates for the seed function.

For convenience, let $y(x)$ denote a temporal sample of $y(t, x)$ at a particular time instant. The characteristic cortical response to the spectral pattern $y(x)$ is given by

$$z(x_c, \Omega_c) = y(x) * h_w(x; \Omega_c)|_{x=x_c} \tag{2.21}$$

$$= a(x_c, \Omega_c) e^{j\psi(x_c, \Omega_c)} \tag{2.22}$$

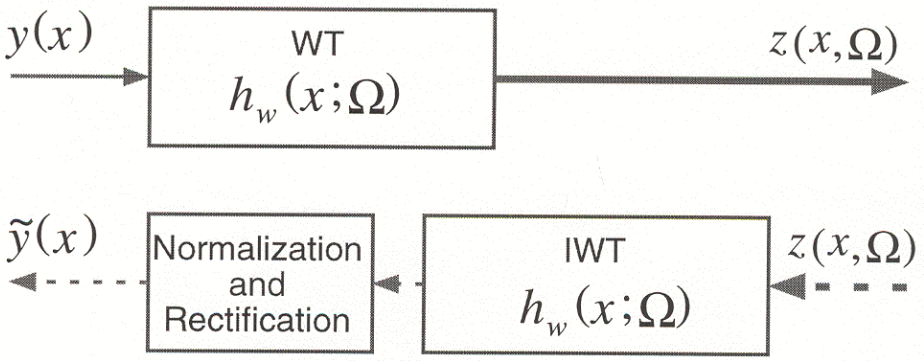

Figure 2.6: Block diagram of the cortical model.

where

$$a(x_c, \Omega_c) = \sqrt{[y(x) * h(x; \Omega_c)]^2 + [y(x) * \hat{h}(x; \Omega_c)]^2}|_{x=x_c} \quad (2.23)$$

$$\psi(x_c, \Omega_c) = \arctan \frac{y(x) * \hat{h}(x; \Omega_c)}{y(x) * h(x; \Omega_c)}|_{x=x_c} \quad (2.24)$$

are the characteristic amplitude and the characteristic phase, respectively. The physical interpretation is that, for all of the cells tuned to (x_c, Ω_c), the cell with $\psi(x_c, \Omega_c)$-symmetry has the maximum response $a(x_c, \Omega_c)$. The raw response of the cells tuned to other symmetries can be obtained by sinusoidally interpolating the real part and the imaginary part of the characteristic response:

$$r(x_c, \Omega_c, \phi_c) = \Re\{z(x, \Omega_c)\} \cos\phi_c + \Im\{z(x_c, \Omega_c)\} \sin\phi_c \quad (2.25)$$

$$= a(x_c, \Omega_c) \cos(\psi(x_c, \Omega_c) - \phi_c) \quad (2.26)$$

where $\Re\{\cdot\}$ denotes the real part and $\Im\{\cdot\}$ denotes the imaginary part.

The block diagram of the cortical model is given in the upper part of Figure 2.6 where the thin line represents 1-D signal flow and the thick line represents 2-D signal flow. Two cortical representations are given in Figure 2.7.

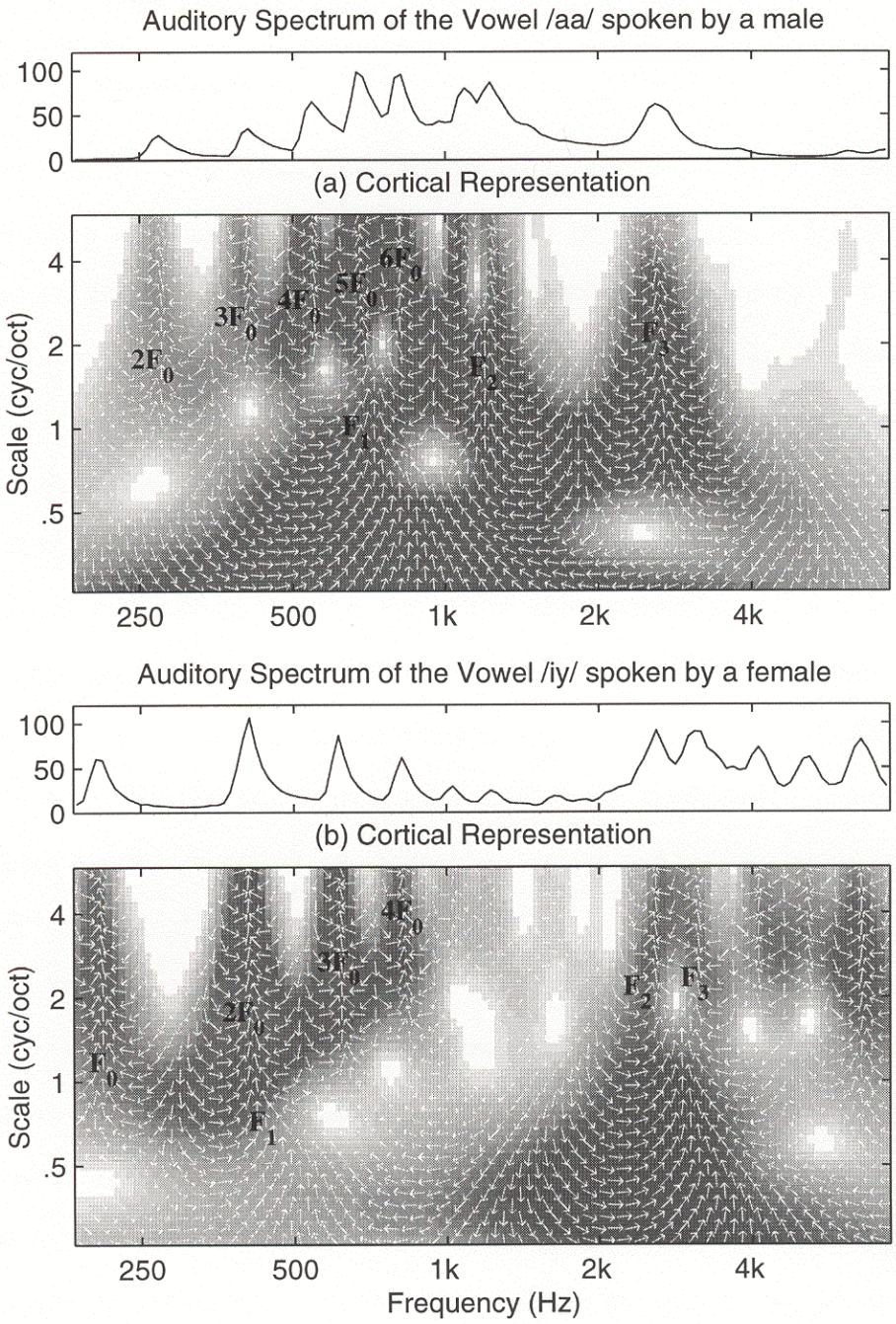

Figure 2.7: The cortical representation of (a) /aa/ spoken by a male; (b) /iy/ spoken by a female. The corresponding auditory spectra in arbitrary unit were superimposed on top of the cortical representations.

2.3.3 Properties and Reconstruction

Cortical Representation

The raw response $r(x_c, \Omega_c, \phi_c)$ is a 3-D array and hence is difficult to illustrate. On the other hand, the characteristic cortical response $z(x_c, \Omega_c)$ is only 2-D and less redundant than the raw response. However, the complex nature of the characteristic cortical response makes the illustration non-trivial. Wang and Shamma (1995) used different colors to indicate the phase and the saturation of the color to indicate the magnitude [95]. This idea does not work well when color is not available. In this study, the magnitude is indicated by the darkness and the phase is indicated by the direction of the white arrows. In order to highlight the spectral peaks, the "$\uparrow$" is chosen to represent a phase of zero degrees. The other phases are applied counterclockwise, e.g., "$\leftarrow$" for $\pi/2$, "$\rightarrow$" for $-\pi/2$, and "$\downarrow$" for π. The size of the arrows also reflects the magnitude. Consequently, in the high magnitude region, the arrows are clearer due to their lengths and the contrast to their surroundings.

The abscissa is the tonotopic axis, covering approximately 5.4 octaves with 24 channel/octave resolution. The ordinate is the logarithmic scale axis, ranging from .25 cycle/octave to 6 cycle/octave with 10 channel/octave resolution. Two examples of naturally spoken vowels obtained from the TIMIT database (see Section A.1) are shown in Figure 2.7. The systematic phase structure of the cortical representation results in the arrows always pointing to the nearest peak. Unlike that in the auditory spectrum or other power-spectrum-like representations, the definition of the peak here has a wider sense. For example, at $f \simeq 600$ Hz in Figure 2.7-(a), the cells tuned to 1 cycle/octave resolve one spectral peak

(F_1, first formant) while the cells tuned to about 4 cycle/octave resolve three peaks, i.e., the 5-th, 6-th, and 7-th harmonics.

Pitch and Formants

The responses related to the harmonics are highlighted by F_0, $2F_0$, $3F_0$, etc, where F_0 is the fundamental frequency. Since the harmonics are equally spaced on the linear frequency axis, the spacing between two adjacent partials on the tonotopic axis decreases logarithmically. The partials will be resolved by exponentially increasing scales therefore forming a line with positive slope. Thus this hyper line effectively shows the pitch information. The higher the hyper line is, the lower the pitch (e.g., male voice, see Figure 2.7-(a)) and vice versa (e.g., female voice, see Figure 2.7-(b)).

Roughly speaking, the response due to pitch resides in the upper-right region, $f/\Omega < 500$. The responses related to the formants are highlighted by F_1, F_2, F_3, etc. However, the trace connecting the formants is usually not a straight line. The locations and the strength of the responses depend on the peak location and the amplitude, which are highly correlated with the shape and the stress of the resonator (e.g., human vocal tracts, musical instruments).

Taking these properties into account, it is easier to imagine how the auditory system is able to extract the physical properties of the sound source through the acoustical signal. Different vowels (due to different vocal manners) result in different "images" on the primary auditory cortex. The "images" are "visible" to the brain, which has adapted in such a way as to distinguish among the patterns. An informal experiment shows that human subjects, after minutes of training, are able to recognize different patterns due to different vowels and

different accents.

Tree Structure and Singularities

The phase contours of the cortical representations are shown in Figure 2.8. The vertical contour lines reveal the one-dimensional nature of the source pattern. The density of contours represents the local slope of the spectrum. Thicker contours are actually composed of many regular contours due to the phase discontinuity at $\phi_c = \pm\pi$. The tree structure is a common feature for all kinds of the wavelet-based analysis models. As shown in Figure 2.8, the tree structure acts as the skeleton of the whole plot and successfully highlights the spectral peaks due to harmonics and formants. Slightly above each bifurcation of the tree structure, there sits a interesting point which will be called a *singular point* or a *singularity* throughout this paper. In the neighborhood of a singular point, the cortical fillers With lower scale will resolve only one spectral peak while the filters with higher scale will resolve two or more peaks. Obviously, the exact location of the singular point will be affected by the characteristics of its surrounding peaks. These characteristics include amplitude, bandwidth, and symmetry. Thus the constellation of the singular points encodes a lot of information of the spectrum. Since the response at the singular points is zero, the feature set is naturally level independent.

The singular points are located at (x, Ω)s where the amplitude $a(x, \Omega)$s are mathematical zeros, i.e., not simply numerical zeros. It is easy to detect them by eye and pick them by hand, e.g., by looking for the bright spots on the cortical representation (see Figure 2.7). However, due to the discrete nature of digital computation, it is not a trivial task for the digital computer to locate the singular

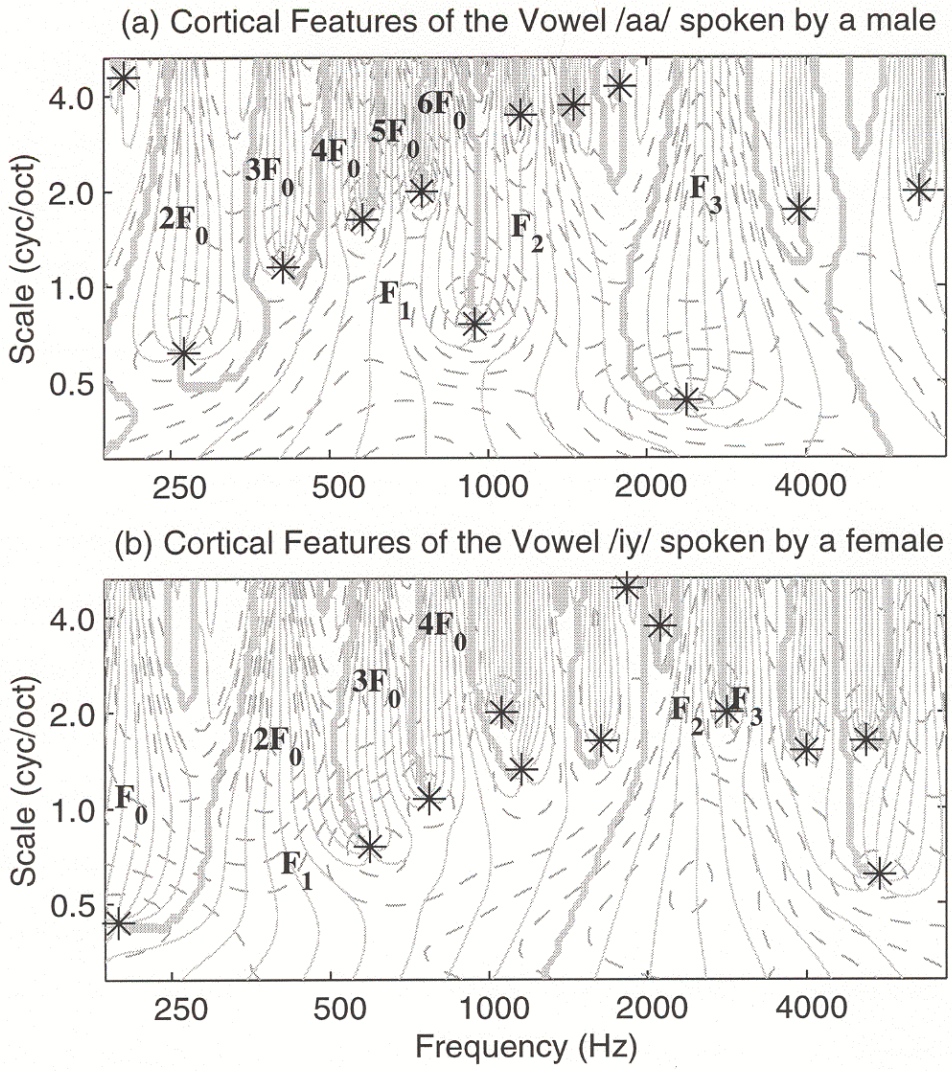

Figure 2.8: The cortical features of (a) /aa/ spoken by a male; (b) /iy/ spoken by a female. The solid lines indicate the phase contour. The dashed lines indicate the magnitude contours. The asterisks indicate the singular points.

points. Fleet (1991) has shown a method to detect the singular points based on the phase discontinuities [27]. The method can be simplified into a two-step algorithm.

1. Collect a set of (x, Ω)s where $a(x_m, \Omega_n)$ is a local minimum.
2. Accept the (x, Ω)s which satisfy $a(x_m, \Omega_n) < \delta\overline{a(x, \Omega)}$ where δ is a small number representing the threshold factor and $\bar{\cdot}$ denotes the mathematical mean.

The asterisks on the two plots in Figure 2.8 were extracted according to this algorithm.

The cortical representation plot provides a novel way to view the simulated human perception of sound. However, it is difficult to view many plots at the same time to check the consistency or to look for invariant features. Because of its three-dimensional nature, it is not suitable to be superimposed in the complex domain. The singularity constellations make the superposition possible. The severe reduction involved cannot promise to preserve all of the information, yet it is useful to see thousands of data in one plot. Figure 2.9 demonstrates the singular points for ten American English vowels extracted from the TIMIT database (see Section A.1). In order to eliminate the singular points due to the harmonics, only four singular points with smaller scale-frequency ratios (Ω/f) were picked. The darkness indicates the population of certain (f, Ω) points.

The relationship between the singular points and the acoustic characteristics can be explored by observing the location of specific singular points. The average formant values listed in Table 5.1 are used to explain this point. The first formant ratio of a male-spoken /aa/ is about 1.79 (1177/658) which is roughly .84 octaves ($\log_2 1.79$). On the cortical representation, ideally these two formants

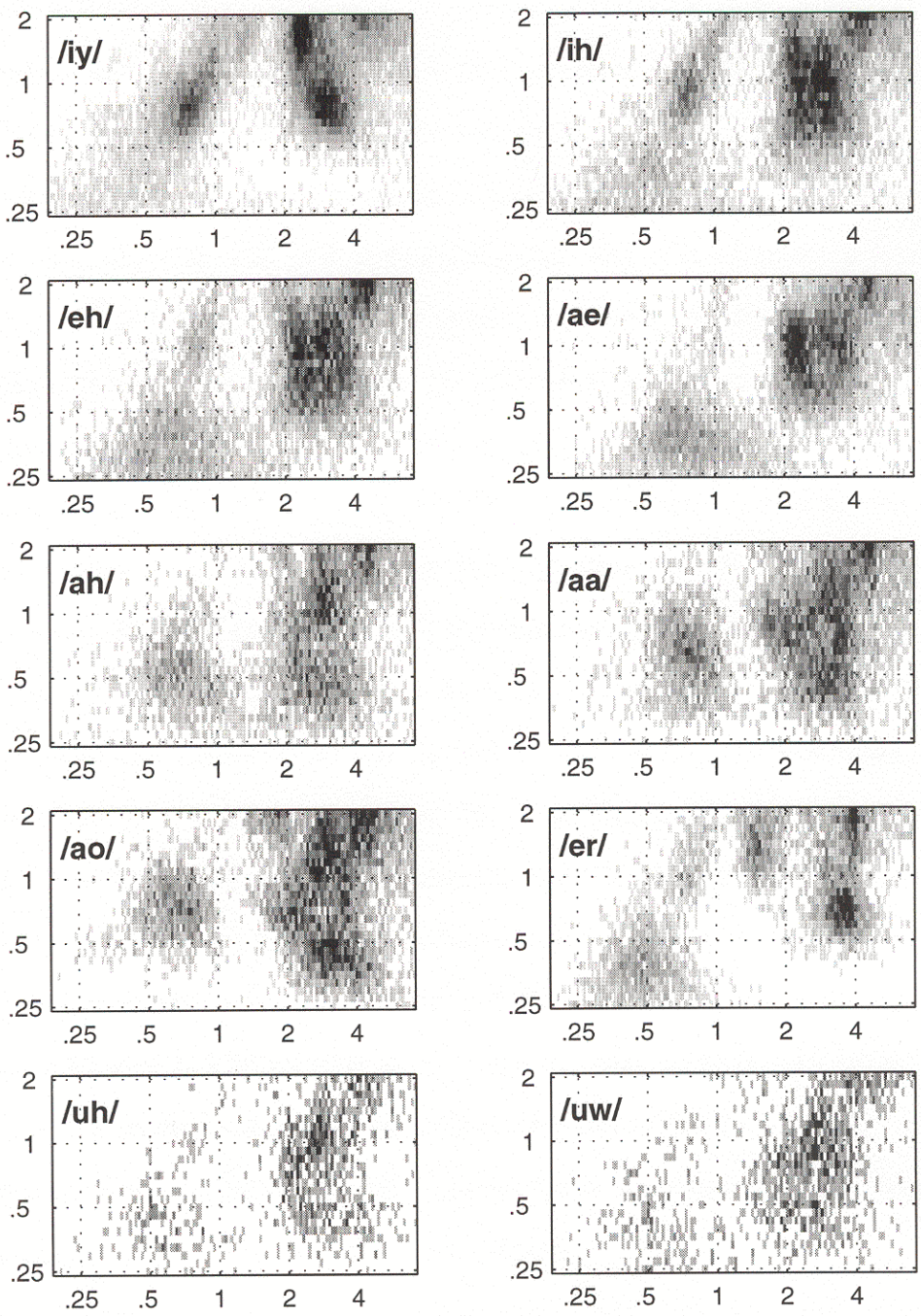

Figure 2.9: The singularity constellations of ten American English vowels. The ordinate indicates the scale axis (in cyc/oct). The abscissa indicates the frequency axis (in kHz).

will be resolved as two peaks at 1.19 cycle/octave (1/.84) but as one peak at .59 cycle/octave. Realistically, the scale to resolve just one peak should be even lower due to the dullness of the trough in the auditory spectrum. The actual location is also affected by the local bandwidth and the relative amplitude. In any case, there should exist a singular point within the frequency-scale tile ($650 < f < 1180$, $.5 < \Omega < 1.2$). By similar reasoning, the frequency-scale tile for this particular point due to female /aa/s (with average formants: $F_1 = 746$ Hz and $F_2 = 1366$ Hz) is ($760 < f < 1370$, $.5 < \Omega < 1.2$). Referring to Figure 2.9, the most likely location is ($f = .8$ kHz, $\Omega = .7$ cycle/octave) and almost all of the interesting points are confined by the deductive tile. For the vowel /ao/, with a lower center of gravity (COG), the singular point is roughly at the same scale but at a lower frequency. For an acute vowel like /ae/, by considering both females and males ($F_1 = 692$ Hz and $F_2 = 1790$ Hz), the confining tile is roughly ($690 < f < 1790$, $.3 < \Omega < .7$). This matches Figure 2.9 very well. For the vowel /eh/, the interesting singular point is slightly lower in frequency. For the vowel /ah/, its location is between that of /aa/ and that of /ae/. For a diffuse vowel like /iy/ ($F_1 = 420$ Hz and $F_2 = 2185$ Hz), the confining tile is roughly ($.69 < f < 1.79$, $.2 < \Omega < .4$). This singular point is very close to the border of the representation and sometimes falls out of the scope. Since the first two formants are too far apart, the singular points contributed by the harmonic peaks are more apparent than the interesting one.

In Figure 2.9, a cluster found in higher scale for both /iy/ and /ih/ exemplifies this. Essentially, this interesting point moves from high scale down to lower scale as the constriction of the vocal tract changes from the back to the front cavity (from /aa/ $\rightarrow$ /ah/ $\rightarrow$ /ae/ $\rightarrow$ /eh/ $\rightarrow$ /ih/ $\rightarrow$ /iy/). For the rounded vowels,

this singular point is located in the lower frequency region ($\sim$ .5 kHz) because the first two formants are relatively lower than those of un-rounded vowels.

The constellation is a very compact feature of the cortical representation; using this feature can greatly reduce the size of the cortical representation. This not only makes storage easier but also makes identification more straightforward. Our brain might use the information of the singularities to drive the muscular system in order to control the vocal tract shape to produce a desired sound. Further research is necessary to find the mapping from the constellation to the acoustic spectrum.

Spectrum Reconstruction

Basically, the primary purpose of the cortical model is to analyze the auditory spectrum. In the cortical representation, the speech features like formants and pitch are well separated. The model provides a way to manipulate the auditory information at the cortical level, e.g., to remove certain scales, to shift a pattern along one of the axes, to emphasize some interesting region, or to change the phase of the response. In order to investigate what kind of change in the auditory spectrum will result in certain change in the cortical representation, one needs to invert the cortical representation back to the auditory spectrum.

In the scale domain, the cortical representation is

$$Z(\Omega;\Omega_c) = Y(\Omega)H_w(\Omega;\Omega_c) \tag{2.27}$$

where Z, Y, and H_w are the Fourier transforms of z, y, and h_w respectively. The cortical filter bank can be designed to be almost unitary, i.e., the overall gain is flat with the magnitude 1 over a reasonably wide range. If the auditory spectrum $Y(\Omega)$ is within the effective band, it can be reconstructed by integrating

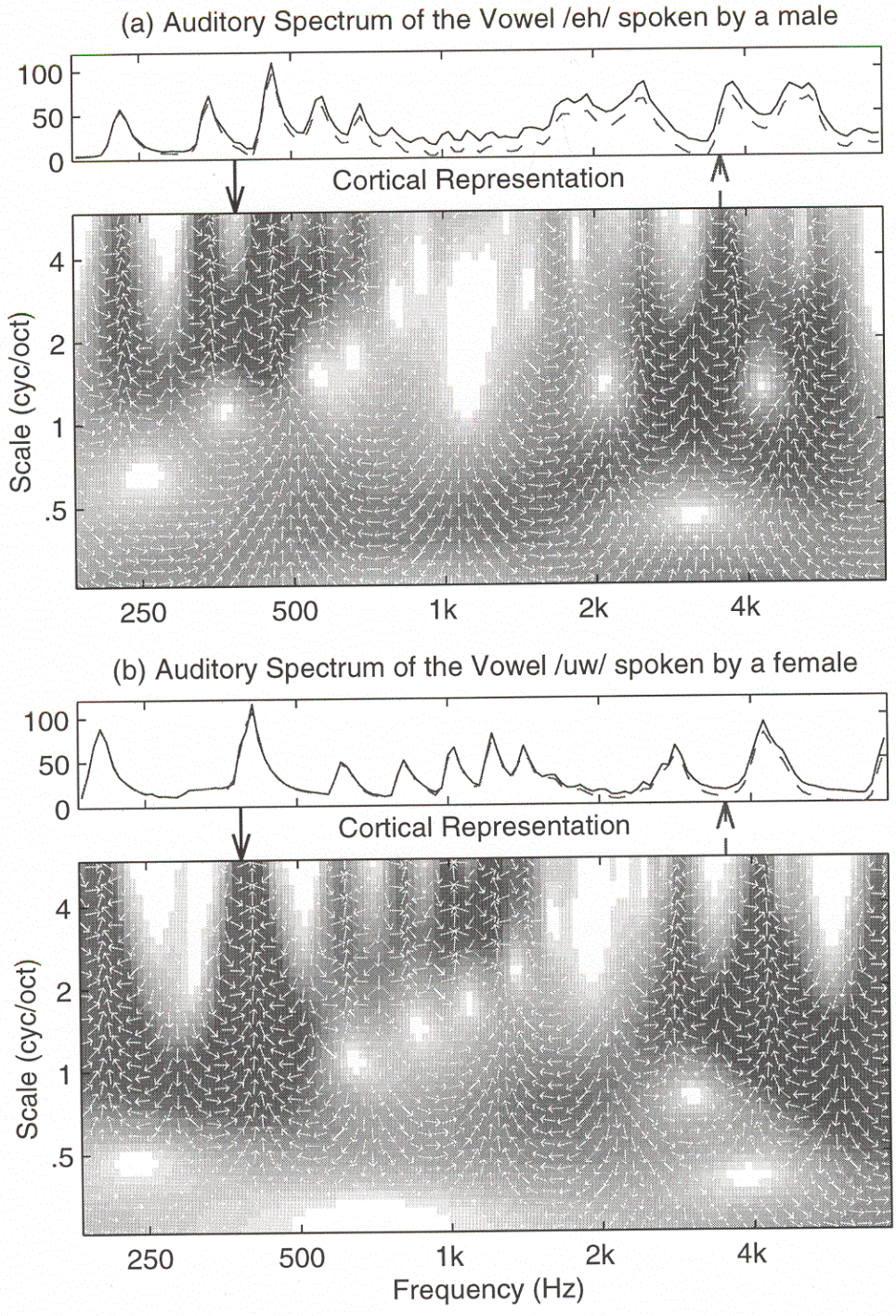

Figure 2.10: Examples of the reconstruction from the cortical representation. The reconstructed auditory spectra are plotted in dashed lines.

the response after reverse filtering,

$$\tilde{Y}(\Omega) \simeq \sum_{c} Z(\Omega, \Omega_c) H_w^*(\Omega; \Omega_c) \tag{2.28}$$

Assuming the number of channels is K, due to the band-pass nature of the cortical filter, the effective band is roughly from Ω_1 to Ω_K. Thus, reconstructing an auditory spectrum in this way will reduce the scale components outside the effective band. Actually, the higher scale components contain little information and are often desired to be suppressed. Unfortunately, for the auditory spectrum, the scale components below Ω_1 cannot be neglected. The normalization, at least for the low scale components, is therefore inevitable. Ideally, the perfect reconstruction can be achieved by integrating the response after reverse filtering with normalization,

$$\tilde{Y}(\Omega) = \sum_{c} Z(\Omega, \Omega_c) H_w^*(\Omega; \Omega_c) / \sum_{c} H_w(\Omega; \Omega_c) H_w^*(\Omega; \Omega_c) \tag{2.29}$$

However, the cortical filters are band-pass filters; hence $H_w(0; \Omega_c) \simeq 0$ for all Ω_cs. As a result, any noise outside the effective band will be magnified by the denominator. One practical solution is to make the first band-pass filter a pure low-pass filter and the last filter a pure high-pass filter. In this way, a perfect reconstruction (Akansu and Haddad, 1992) [1] can be achieved. The resulting overall gain will have less dynamic range; therefore the reconstruction is more stable. One other practical problem is that the reconstruction $\tilde{y}(x)$ may not be real and positive. Actually, even if the cortical representation has not been altered, the reconstruction is essentially complex in most of cases due to the imperfectness of the numerical inverse Fourier transform. A simple treatment can be applied by taking the rectified real part of the reconstruction, i.e., $\max(\Re\{\tilde{y}(x)\}, 0)$. Two examples are given in Figure 2.10. As the reader can see, the reconstruction is

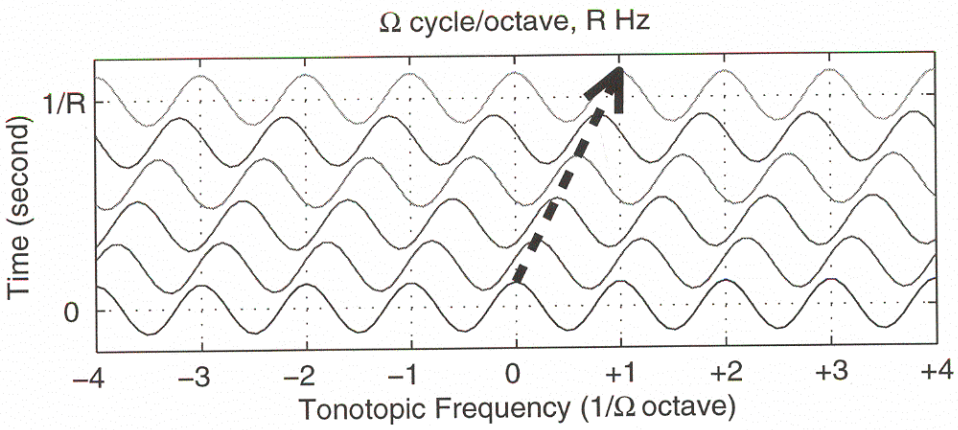

Figure 2.11: A moving ripple of rate R Hz and scale Ω cycle/octave.

almost the same as the original spectrum.

2.4 Spectrotemporal Analysis Model - The Extended Cortical Model

2.4.1 Central Auditory System - Dynamic Processing

Complex spectra such as those of speech and music are broadband and dynamic. It has been successfully demonstrated that AI units are essentially linear. Therefore, the response can be predicted based on the measured ripple transfer function (Shamma and Versnel, 1995) [76]. The linearity is valid for both stationary and moving ripples (e.g., Figure 2.11). The transfer function of up-moving ripples, as well as that of down-moving ripples, are separable by quadrant (Depireux *et al.*, 1998) [19]. Thus the multiscale cortical representation described in Section 2.3.2 can be further extended to represent dynamic features of complex spectra that change in time. Such spectra can be conceptually considered as a weighted sum

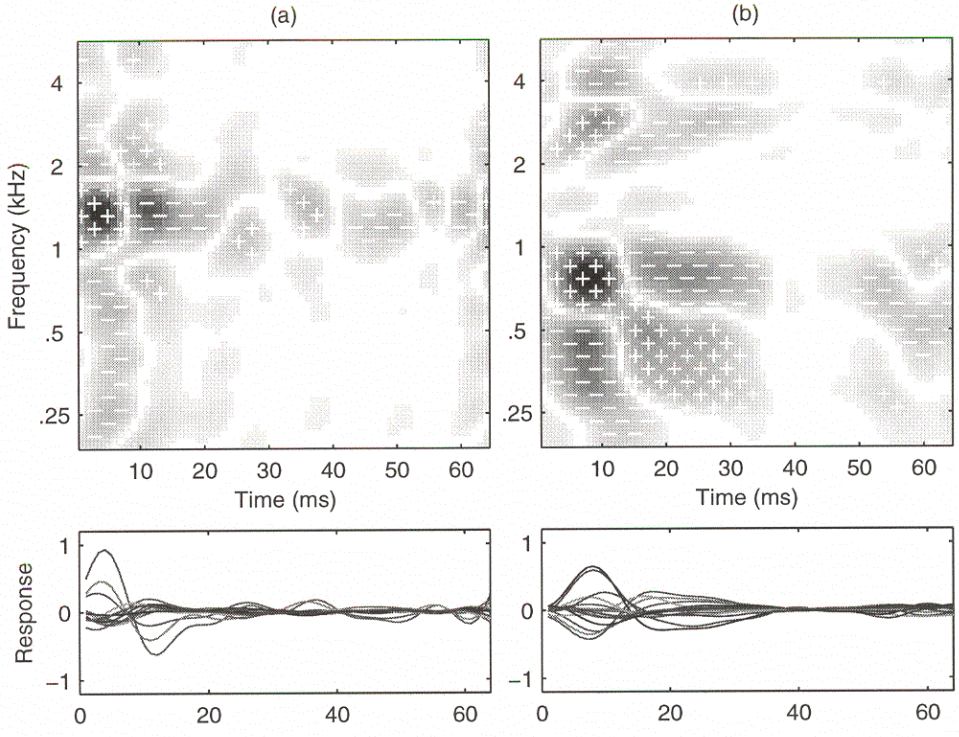

Figure 2.12: Two sample spectrotemporal response fields: (a) a cell with symmetric receptive field and no directional preference; (b) a cell with asymmetric receptive field, longer latency, and preference for downward movement.

of ripples which are moving in time at various velocities and directions. Similar measurement methods can be devised to obtain temporal transfer functions, from which impulse responses can be derived. The impulse responses reflect the dynamic properties of AI units.

Recordings from many AI cells reveal that they respond to modulations in the spectral envelope in a linear and temporally selective manner, being tuned to rates in the range of $2 \sim 16$ Hz (Kowalski *et al.*, 1996 [42] [43]). One possible interpretation of these findings is that AI units have impulse response (IR) functions with a range of dilations analogous to the range of different bandwidths exhibited by the RFs. Specifically, it is assumed in this hypothesis that for any given RF, there are different units with a range of IRs, each encoding the local dynamics of the spectrum at a different rate-scale. That is, there are units exclusively sensitive to slow modulations in the spectrum, while others are tuned to moderate or fast changes. This temporal decomposition is analogous to the multiscale spectral representation produced by the RFs. Figure 2.12 demonstrates two typical spectrotemporal receptive fields (STRFs). The upper panel uses darkness to represent the strength of the response and $\pm$-signs to indicate the polarity. The lower panel displays the superimposed temporal response curves for every 1/4 cycle/octave. A receptive field of symmetric shape is given on the left side. On another side, an asymmetric receptive field, with longer latency and downward-moving selectivity, is given.

Since the primary purpose of the human auditory system is to process speech sounds, it should be adapted to the range of parameters that the articulatory system may produce. For example, the average duration of vowels in continuous speech is about 80 ms (see Figure 2.13). According to the statistics, /ae/ (as in

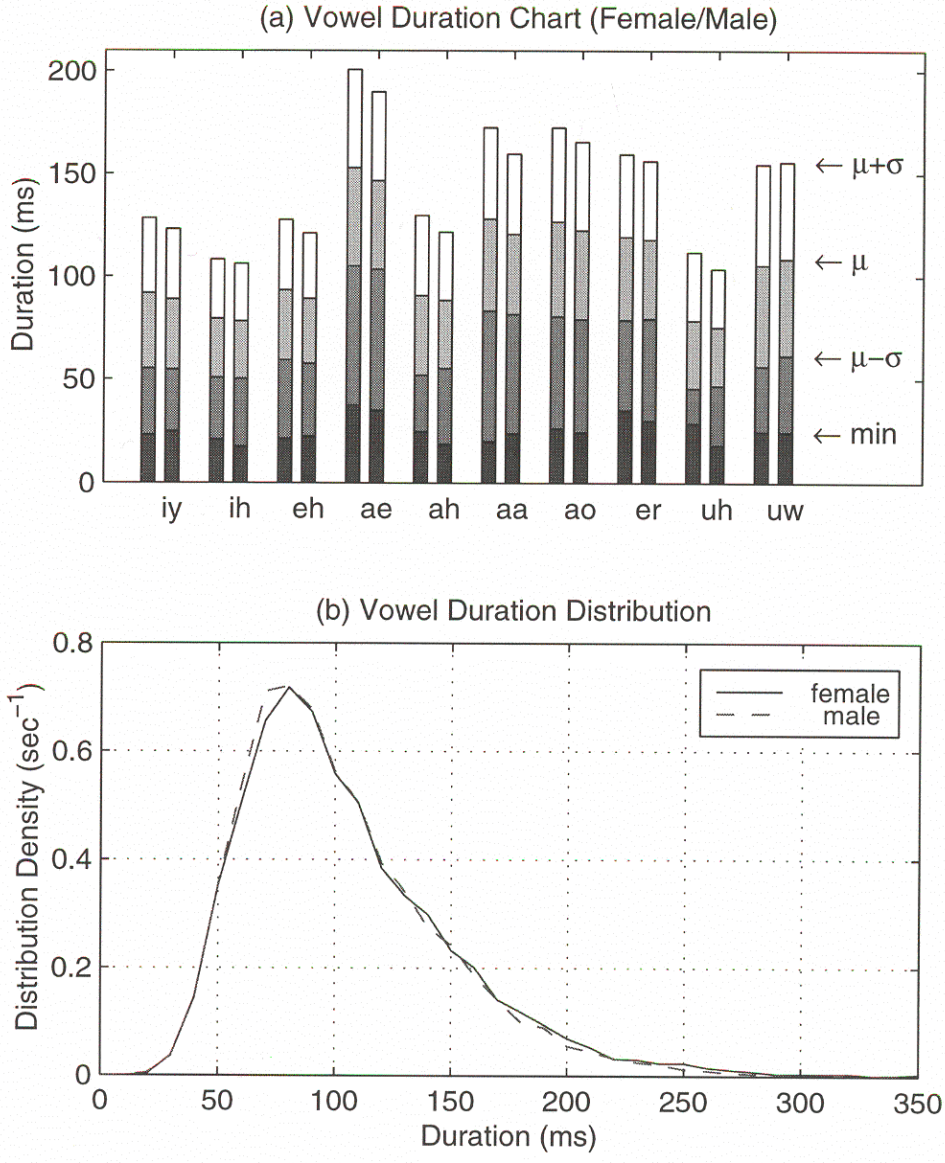

Figure 2.13: (a) The duration of vowels: female (left-hand side) and male (right-hand side). The lowest water-level (min) indicates the minimal duration. The second highest one (μ) indicates the mean duration. The highest and the second lowest ($\mu \pm \sigma$) indicate the deviations. (b) Distribution of duration of all vowels: female (solid) and male (dashed).

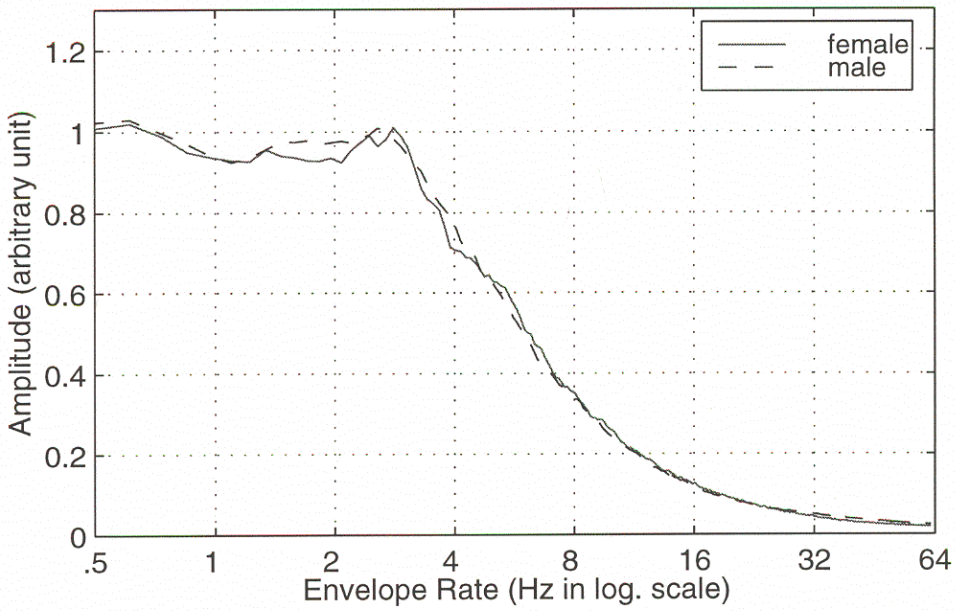

Figure 2.14: Average spectrum of the envelope of the speech waveform for female (solid) and male (dashed).

"hat") is the longest vowel while /ih/ (as in "hit") and /uh/ (as in "hood") are the shortest. This holds for both males and females, even though it often seems that females speak more rapidly. The range of vowel durations, $40 \sim 200$ ms, yields an equivalent range of modulation rates, $5 \sim 25$ Hz.

More evidence from natural speech is the syllabic rate, which is very low compared to the characteristic frequency of any cochlear filter. Thus, a higher-level rate processing mechanism must exist in our central auditory system. Figure 2.14 shows the average spectrum of speech envelopes, produced by averaging magnitude function of speech signals which were smoothed by a 16-ms Hamming window [1]. The plot shows that the dominating rate is approximately $2 \sim 4$

[1] According to Oppenheim and Schafer [39], the approximate width of main-lobe is $8\pi/M$, where M is the number of taps. Thus the main-lope width is about 250 Hz. The measured

Hz, which matches Houtgast and Steeneken's measurement (1985) [39]. The above statistics are based on the TIMIT speech database which contains 6300 sentences (see Section A.1). A pilot experiment shows that thresholds on modulation strength behaves like U-shape curves against the rate axis (see Section 3.5 for details). The most sensitive region is somewhere between 2 and 4 Hz.

2.4.2 Spectrotemporal Processing Model

Based on the evidence mentioned in the preceding section, a spectrotemporal processing model can be extended from the pure spectral processing model. The cells tuned to different traveling rates are modeled by a set of temporal bandpass filters. The impulse response $g(t; R_0)$, which is real, causal, and stable, was chosen to mimic the measured response (Figure 2.12). For any other rate, the impulse response is obtained by dilating the seed response, i.e., $g(t; R_c) = (R_c/R_0)g((R_c/R_0)t; R_0)$. The details about designing the seed temporal filter is elaborated in Section C.2. The phase of the filter is employed to describe the range of latencies and shapes of the actual impulse responses. Analogous to the scale wavelet, this phenomenon can be modeled by means of Hilbert transform.

$$g_{\mathcal{RF}}(t; R_c, \theta_c) = g(t; R_c)\cos\theta_c + \hat{g}(t; R_c)\sin\theta_c \tag{2.30}$$

Therefore the spectrotemporal response at a cell c for the pattern $y(t, x)$ is

$$r(t, x; R_c, \theta_c, \Omega_c, \phi_c) = y(t, x) *_{xt} [g_{\mathcal{RF}}(t; R_c, \theta_c) \cdot h_{\mathcal{RF}}(t; \Omega_c, \phi_c)] \tag{2.31}$$

$$= y(t, x) *_{xt} [g \cdot h \cos\theta_c \cos\phi_c + g \cdot \hat{h} \cos\theta_c \sin\phi_c + \hat{g} \cdot h \sin\theta_c \cos\phi_c + \hat{g} \cdot \hat{h} \sin\theta_c \sin\phi_c] \tag{2.32}$$

3-dB cutoff frequency is approximately at 47.3 Hz)

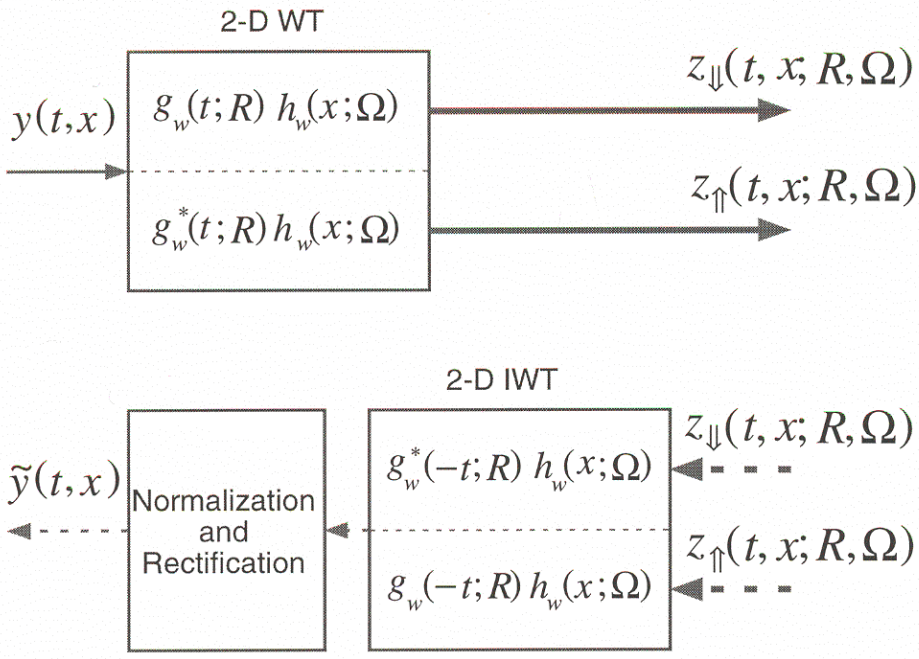

Figure 2.15: Block diagram of the extended cortical model.

2.4.3 Implementation and Reconstruction

This multi-resolution auditory model can be thought as a filter array composed of dilation-related spectrotemporal filters. For audio signal processing, the rate axis R can be discretized into R_is ranging from 2 to 32 Hz with 2-channel/octave resolution, and the scale axis Ω discretized into Ω_js ranging from .25 to 8 cycle/octave with 2-channel/octave resolution. The block diagram is shown in the upper part of Figure 2.15. It is practical to compute the spectrotemporal response in the two-dimensional Fourier domain (R, Ω):

$$Z_{\Downarrow}(R,\Omega;R_i,\Omega_j) = Y(R,\Omega)G_w(R;R_i)H_w(\Omega;\Omega_j) \tag{2.38}$$

$$Z_{\Uparrow}(R,\Omega;R_i,\Omega_j) = Y(R,\Omega)G_w^*(-R;R_i)H_w(\Omega;\Omega_j) \tag{2.39}$$

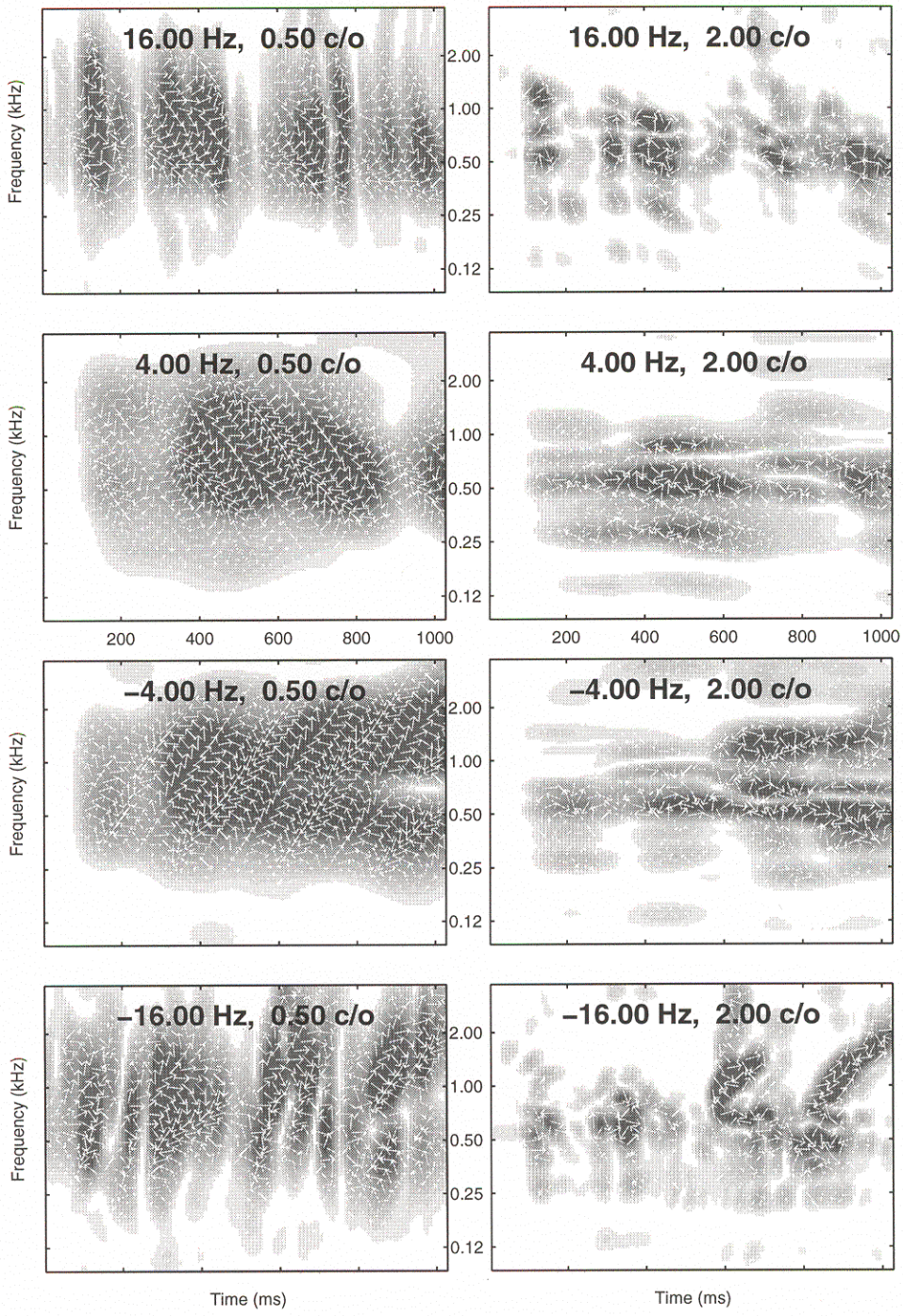

Figure 2.16: Spectrotemporal representation for an auditory spectrogram (see text).

Spectrotemporal Cortical Representation

Basically, the cortical representation is a multi-scale multi-rate time-frequency representation. According to the above specification, a time-frequency representation consists of 2-by-9-by-11 (direction-rate-scale) elements. Each individual time-frequency representation is a *complex* matrix. The size and the complexity make it difficult to have the entire representation in a page without color. Therefore, only the abstract of the spectrotemporal representation is shown (Figure 2.16). The source spectrogram is given in Figure 2.2-(b) or Figure 2.18-(b). A panel represents a cluster of cells (with different BF) tuned to the same rate and the same scale. The response level is indicated by the darkness. The best phase is highlighted by the white arrows.

The low-rate-low-scale panel ($\pm$4 Hz, .5 cyc/oct) is an "approximation" image which gives the global energy distribution. The high-rate-high-scale panel ($\pm$16 Hz, 2 cyc/oct) is a "detail" image which exhibits the fine structure. The high-rate-low-scale panel ($\pm$16 Hz, .5 cyc/oct) is a "vertical" image in which the temporal edge is clearly shown. The low-rate-high-scale panel ($\pm$4 Hz, 2 cyc/oct) is a "horizontal" image which enhances the spectral edge. All of these terms in quotes are frequently used in the wavelet transform literature. In addition, this representation demonstrates the directional selectivity of AI. The negative-rate panel selects upward-moving spectra while the positive-rate panel prefers downward moving spectra. All of these phenomena were reported in recent physiological findings (Kowalski *et al.*, 1996 [42]; Depireux *et al.*, 1998 [19]).

In order to demonstrate the model without any loss, no reduction was performed in the production of the representation. Since the entire representation is composed of many *complex* matrices, the required memory for storage is 198×2

times the size to store the source spectrogram. However, according to the wavelet literature (e.g., Fliege, 1994 [29]; Akansu and Haddad [1]), it is theoretically possible to have the same size as the input spectrogram. For example, one can downsample the smoother representation to certain degree or use less precision to encode the response.

Rate-Scale Plot

Since this model is a multi-dimensional model, we may collapse some of the dimensions to focus on the interesting axes. For example, if we collapse both time and frequency axes, we should find a large response around a particular (R, Ω) in the rate-scale plot for a single moving ripple. More peaks should be found for the multiple moving ripples or most dynamic acoustic signals. A progressing rate-scale presentation is given in Figure 2.17. This is an alternative presentation of Figure 2.16 which originated from the spectrogram in Figure 2.2-(b) or Figure 2.18-(b). The rate-scale response $\zeta(R, \Omega)$ at a time instant was generated using following formulas:

$$\zeta_{\Downarrow}(R_c, \Omega_c; t) = \int |z_{\Downarrow}(t, x; R_c, \Omega_c)| dx \tag{2.40}$$

$$\zeta_{\Uparrow}(R_c, \Omega_c; t) = \int |z_{\Uparrow}(t, x; R_c, \Omega_c)| dx \tag{2.41}$$

Each individual rate-scale representation consists of two small plots. The upper one is of positive rate (downward moving) and the lower one is of negative rate (upward moving). The ordinate indicates the rate (in Hz) and the abscissa indicates the scale (in cycle/octave). Throughout the entire series of plots (Figure 2.17), the color scale is consistent. In order to enhance the response peaks, less-than-average response in each time instant was blanked.

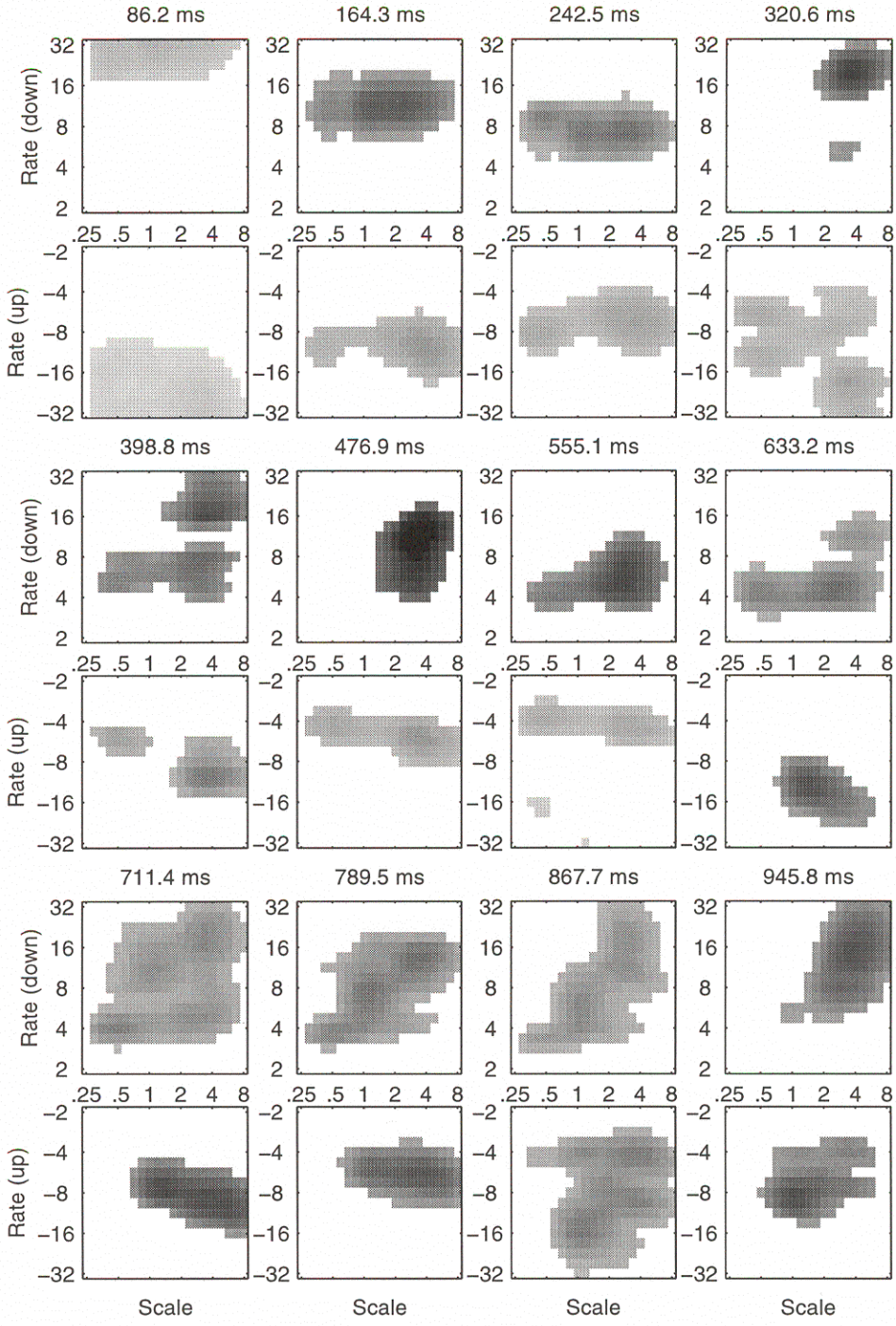

Figure 2.17: Progressing rate-scale representation (see text).

This plot is useful to track the movement of acoustic features. The reader can track the global spectral movement in the low scale region and detailed spectral movement in the high scale region. In Figure 2.17, harmonics and formants travel downwards around 500 ms from the start of the sound; hence the stronger response is shown in the upper panel (which prefers to downward movement). Around $600 \sim 700$ ms, the response in the lower panel turns stronger because harmonics and formants move upwards. During this period, while /ai/ as in "right" was pronounced, the second formant rises from 800 Hz to 1600 Hz whereas the first formant merely changes within a limited range. Thus the response due to the global spectral movement transfers from high scale to low scale. At the end of the sentence, the harmonics move down but the COG of the formants move up. Therefore the high scale response appears in the upper panel (downward) and the low scale response appears in the lower one (upward).

Spectrogram Reconstruction

The spectrogram can be perfectly reconstructed by the following formula.

$$\tilde{Y}(R,\Omega) = \frac{\sum_{i,j,\Updownarrow} Z_{\Updownarrow}(R,\Omega;R_i,\Omega_j) G^*_{\Updownarrow} H_w(\pm\Omega;\Omega_j)}{\sum_{i,j,\Updownarrow} |G_{\Updownarrow} H_w(\Omega;\Omega_j)|^2} \tag{2.42}$$

where $G_{\Downarrow} \equiv G_w(R;R_i)$ and $G_{\Uparrow} \equiv G^*_w(-R;R_i)$. However, the dc edge of the squared sum transfer function is zero due to the zero-mean nature of the individual transfer function. Similarly, the overall gain in the high rate (scale) is relatively small. This will magnify any kind of noise in those region, e.g., numerical truncation errors and quantization errors. One practical way is to make the first band-pass filter to be a pure low-pass filter and the last filter a pure high-pass filter. The resulting overall gain will have less dynamic range; therefore the reconstruction is more stable. Another practical strategy is to weight the filters

before decomposition such that the overall gain is roughly unitary everywhere. By doing so, perfect reconstruction can be achieved without post-normalization. As in the pure spatial processing model, the reconstruction $\tilde{y}(t,x)$ may not be real and positive due to numerical problems. The rectified real part of the reconstruction, i.e., $\max(\Re\{\tilde{y}(t,x)\},0)$, can be taken as the final reconstruction. The block diagram is presented in the lower part of Figure 2.15. A reconstruction schema is shown in Figure 2.18 (see (b) to (c) to (d)). The reconstructed auditory spectrogram (Figure 2.18-(d)) looks like its origin (Figure 2.18-(e)). In this case, the error percentage is below 1%.

2.5 Summary

A comprehensive auditory model including spectral estimation and analysis is proposed. The model was abstracted by a schematic plot, i.e., Figure 2.18. The input for the model is a speech waveform ("Come home right away", Figure 2.18-(a)). This was transformed by the cochlear model into an auditory spectrogram (Figure 2.18-(b)) in which each row represents the average spike count carried by an auditory nerve fiber. Then, the cortical model processed the spectrogram using a two-dimensional filter-array (small panels in Figure 2.18-(c)). Each individual filter is tuned to a specific rate-scale which is determined by the receptive field of a particular cortical cell. As shown in Figure 2.18-(c) (big panels), the signal can be viewed from many aspects. A filter tuned to low rate extracts the temporal envelope while a filter tuned to high rate captures the transition. In the spectral dimension, a filter tuned to low scale highlights the global spectral shape while a filter tuned to high scale shows the fine structure of the spectrum. The

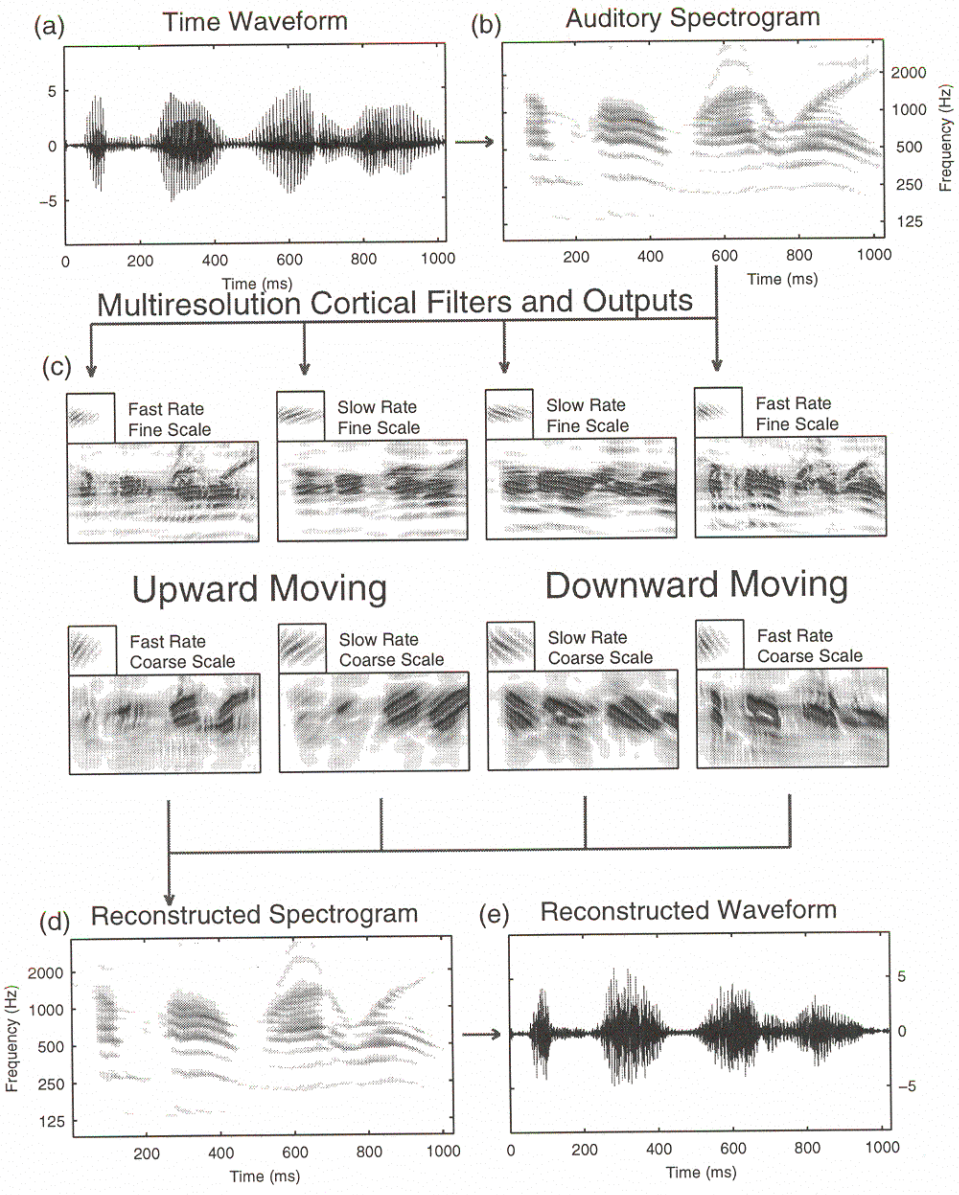

Figure 2.18: Overview of the auditory model.

"spectrally" directional selectivity of the composite response is evident in recent physiological findings (Kowalski *et al.*, 1996 [42]). Thus this multirate multi-scale model, capable of processing dynamic signal, with arbitrary instantaneous spectra, mimics the way the human auditory system processes sounds.

A reconstruction algorithm is available for each stage. Since some non-linear operations involved in the forwarding processing, the inversion process cannot perfectly reproduce the original signal. However, the reconstructed signal (Figure 2.18-(e)) is not substantially degraded since the intelligibility is acceptable. This allows researchers to manipulate a signal at cortical level, e.g., to remove or select certain rates and scales in order to synthesize a signal with desired spectrotemporal characteristics. Automatic audio morphing (Slaney *et al.*, [79]) is then possible at the cortical level by looking for the intermediate cortical representation between those of two sounds.

Speech is probably the most frequent received meaningful signal to our auditory system. Thus, the sound analysis taking place in the auditory system should be optimally adapted to the human voice. This indicates that it may be beneficial to encode the features extracted from this physiological-driven model for the purpose of speech coding. This model also demonstrates a way to separate sounds into different acoustic aspects, e.g., fine-coarse spectral structure, rapid-slow temporal evolution, and up-down moving. This suggests a natural way to handle speech recognition or speaker identification. The drawbacks of applying this model are its heavy computation complexity due to the filter-bank in the cochlear model, and the huge memory requirement due to the spectrotemporal response from the filter-array in the cortical model. The computation load can be relieved by modeling the cochlear filters as IIR filters. As for the

memory requirement, adequate downsampling should be applied to reduce the representation. Theoretically, the data size can be as low as that of the source spectrogram. Certainly, newer signal processing techniques should be explored in further research.

Appendix C

Cortical Filter Design

C.1 Spatial Filter

The spatial response of a cortical cell tuned to 1 cycle/octave can be described by the second derivative of a Gaussian pdf (probability density function) with zero-mean and variance $2/\pi^2$.

$$h(x) = (1 - 2(\pi x)^2)e^{-(\pi x)^2} \tag{C.1}$$

The normalized Fourier transform is

$$H(\Omega) = \Omega^2 e^{1-\Omega^2} \tag{C.2}$$

Another alternative is the Gabor function.

$$h(x) = e^{-x^2/2\sigma^2}\cos(2\pi x) \tag{C.3}$$

Its Fourier transform is

$$H(\Omega) = e^{-2\pi^2\sigma^2(\Omega-1)^2} + e^{-2\pi^2\sigma^2(\Omega+1)^2} \tag{C.4}$$

Both seed functions are depicted in Figure C.1-(a). The negative second derivative of a Gaussian function is represented with solid lines in the upper

panel. The Gabor function is shown with dashed lines. Their scale spectra are shown in the lower panel. Their 3-dB bandwidths are .83 and .82 Ω_0 respectively. One major difference between the two functions is that the former has no dc value whereas the latter has. Since both impulse responses are symmetric with respect to 0, their Fourier transform correspondences are pure real. The responses of the other cells can be obtained by dilating (compressing) the seed function.

$$H(\Omega;\Omega_c) = H(\Omega/\Omega_c) \tag{C.5}$$

$$h(x;\Omega_c) = \Omega_c h(\Omega_c x) \tag{C.6}$$

C.2 Temporal Filter

The temporal cortical impulse response can be modeled as an exponentially decaying sinusoid. For example, at R = 1 Hz,

$$g(t) = e^{-\beta t}\sin 2\pi t, \quad \text{for } t \geq 0 \tag{C.7}$$

$$G(R) = \frac{2\pi}{\beta^2 - 4\pi^2(R^2-1) + j4\pi bR} \tag{C.8}$$

where $\beta = 1$ in this work.

The seed function is shown in Figure C.1. The 3-dB bandwidth is roughly $.33R$. Note that the temporal response is not of zero-mean; since we intend to use this seed function to construct a minimum-phase filter, all of the magnitudes should be greater than zero prior to taking the logarithm. Analogous to the spectral response, the response of other cells can be obtained by dilating (compressing) the seed function.

$$G(R;R_c) = G(R/R_c) \tag{C.9}$$

$$g(t;R) = R_c g(R_c t) \tag{C.10}$$

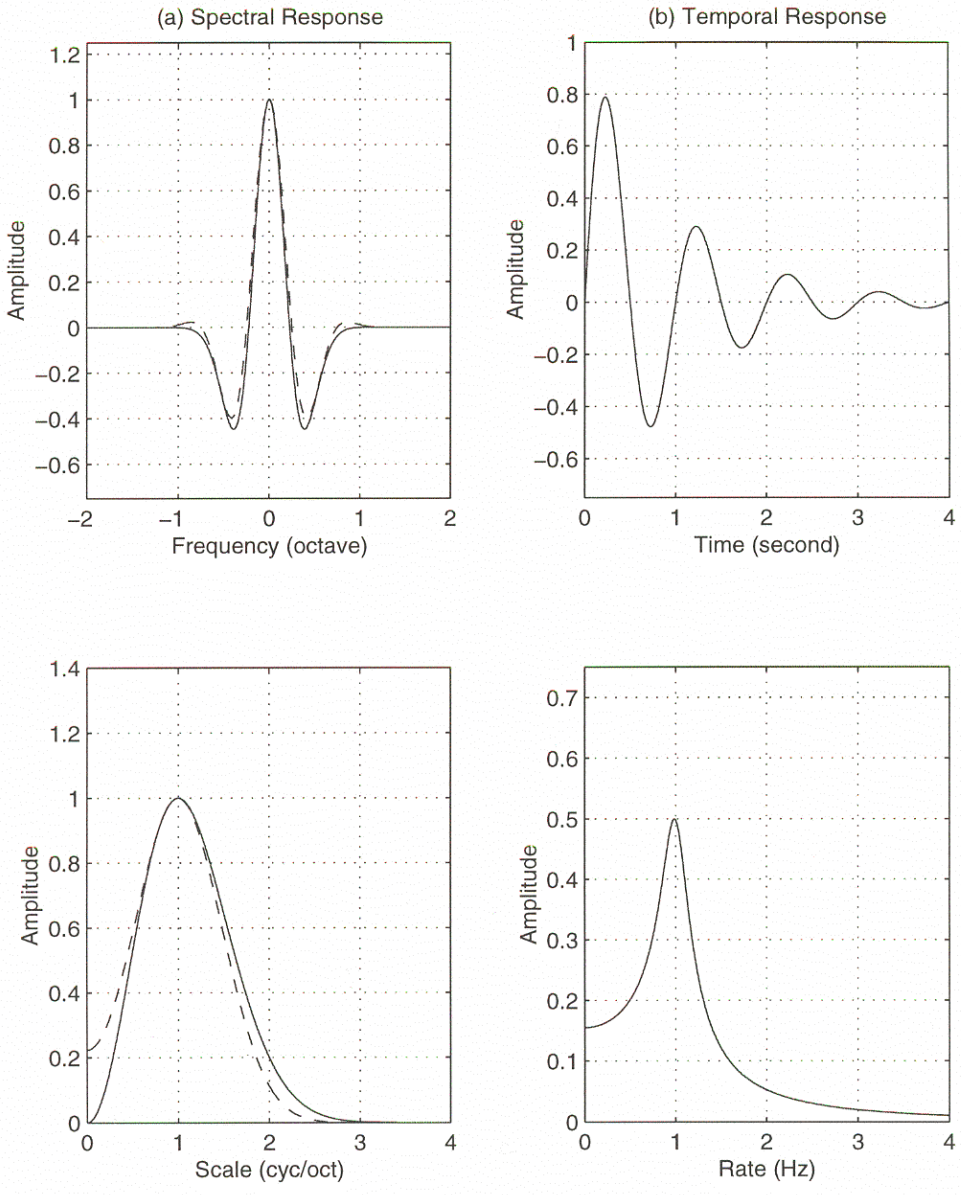

Figure C.1: Cortical response seed functions: (a) Spectral response for $\Omega = 1$ cyc/oct: the negative second derivative of a Gaussian function (solid) or the Gabor function (dashed); (b) Temporal response for $R = 1$ Hz.