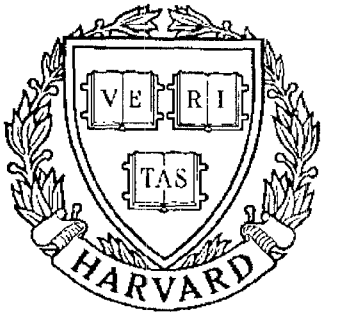
TECHNICAL RESEARCH REPORT

SYSTEMS RESEARCH CENTER



Supported by the National Science Foundation Engineering Research Center Program (NSFD CD 8803012), Industry and the University

Zero-Crossings and Noise Suppression in Auditory Wavelet Transformations

by K. Wang and S..A. Shamma

TR 92-94

Zero-Crossings and Noise Suppression in Auditory Wavelet Transformations*

Kuansan Wang
Shihab A. Shamma
Systems Research Center and Department of Electrical Engineering
A. V. Williams Building
University of Maryland

August 31, 1992

Abstract

A common sequence of operations in the early stages of most biological sensory systems is a wavelet transform followed by a compressive nonlinearity. In this paper, we explore the contribution of these operations to the formation of robust and perceptually significant representations in the auditory system. It is demonstrated that the neural representation of a complex signal such as speech is derived from a highly reduced version of its wavelet transform, specifically, from the distribution of its locally averaged zero-crossing rates along the temporal and scale axes. It is shown analytically that such encoding of the wavelet transform results in mutual suppressive interactions across its different scale representations. Suppression in turn endows the representation with enhanced spectral peaks and superior robustness in noisy environments. Examples using natural speech vowels are presented to illustrate the results. Finally, we discuss the relevance of these findings to conventional subband coding of speech signals.

I Introduction

It has been widely demonstrated that in signal processing and recognition systems dealing with natural inputs such as speech and natural images, combining biologically and perceptually motivated features with conventional techniques provides for deeper insights and achieves better performance. Examples range from image processing [1, 2], to sound segregation [3, 4, 5] to speech recognition [6, 7]. However, in order to gain a true understanding of why such features are helpful, and hence how to develop novel design principles for advanced signal and information processing techniques, it is essential that an analytically tractable and biologically feasible

*This work is supported in part by the Office of Naval Research under contract N00014-91-J-1003, the National Science Foundation's Engineering Research Centers Program: NSFD CDR 8803012, and the Air Force Office of Scientific Research under contract AFOSR-88-0204.

framework be developed. Such framework has to be capable of describing the distinctive functions common to the sensory systems with minimal complexity. Two of the most common transformations that occur in the early stages of sensory processing is a multiresolution representation of the signal (effectively modeled by an affine wavelet transformation), followed by a compressive nonlinearity representing the limited dynamic range of sensory channels due to neural thresholding and saturation properties. In the auditory system, this sequence occurs in the cochlea of the inner ear [8], while in the visual system, it is effectively carried out at the level of the straite cortex [9]. There are many hypotheses as to the advantages afforded by such transformations [8, 9]. One observed but so far difficult to explain benefit is the robustness of the resulting biological representations to noise interference. Undoubtedly, many factors that may account for this robustness have origins that are well beyond the early processing stages discussed here. Nevertheless, it is commonly demonstrated that immediate improvements occur even when only such early transformations are introduced [6, 10].

In an earlier paper [8] we presented a mathematical formulation of these two transformations as expressed in the early stages of auditory processing. We also demonstrate that the neural representation of a complex sound signal such as speech is primarily derived from a highly reduced version of its wavelet transform. Specifically, the perceptually significant spectral features of the signal are encoded in the locally averaged zero-crossing rates along the temporal and scale axes of the transform. In this paper, we shall demonstrate analytically that these operations endow the auditory representations with superior robustness in noisy environments and an enhanced representation of the perceptually significant features of the acoustic spectrum. Finally, we shall discuss briefly the relevance of these results to conventional subband coding of speech signals.

This paper is organized as follows. In section II, we briefly review the mathematical framework of the auditory model discussed earlier in detail in [8]. The frequency analysis stage in the inner ear is approximated by a wavelet transform. With a high gain nonlinearity and the lateral inhibition, it can be shown that in the wavelet domain, only the information at zero crossings is preserved. This procedure can be interpreted as a signal-driven sampling process of which the sampling instants are defined by the zero crossings. The characteristics of the zero crossing rate are further investigated in section III under a stochastic model. In sections IV through VI, we present the detailed analytical theory related to the suppression effects and demonstrate the model representations of speech. Finally, we discuss how model parameters are dictated by engineering performance criteria.

II Mathematical Formulations of the Auditory Transformations

There have been many descriptions of the sequence of operations that a signal undergoes in the early stages of auditory processing, ranging from detailed biological models to approximate computational algorithms [11, 12, 13, 14, 15, 4]. All models, however, can be functionally divided into three stages: analysis, transduction, and reduction (fig. 1). In this section, we briefly review the mathematical formulations that we previously developed to describe these stages. A detailed development of this framework recently appeared in [8].

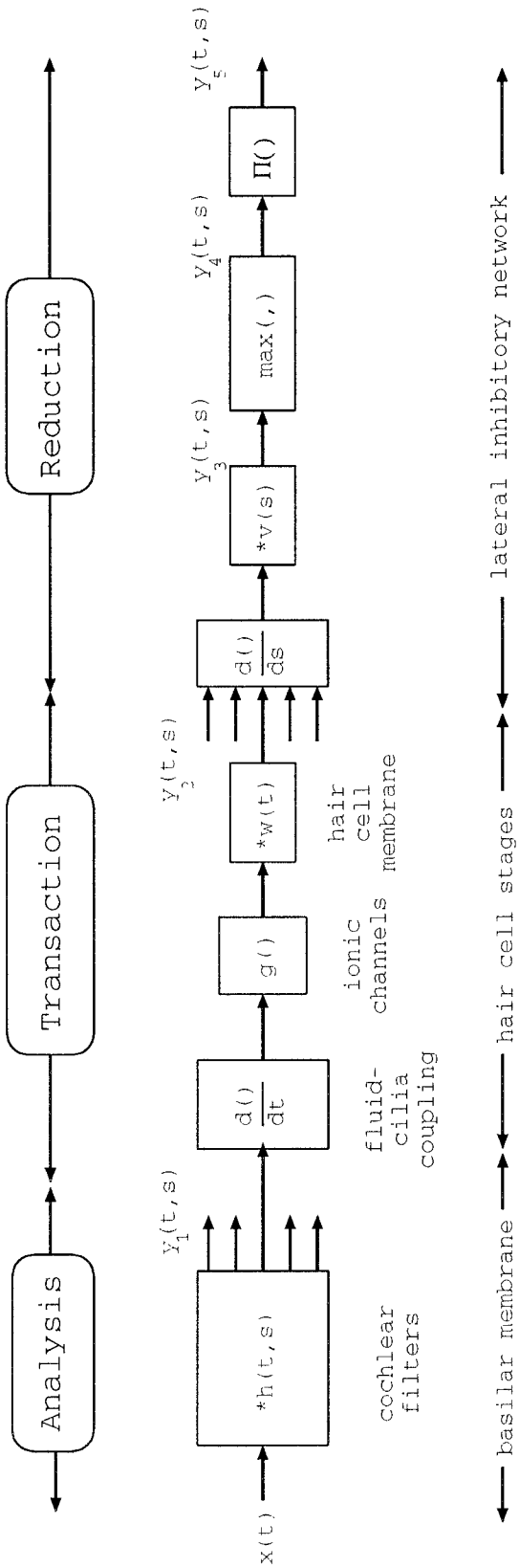


Figure 1: A schematic description of the auditory model

The analysis stage

The first important transformation of sound signals occurs in the membranes of the cochlea of the inner ear. Here the signal, $x(t)$, is approximately analyzed into different frequency bands that are spatially segregated along the length of the cochlea. From a signal processing point of view, one may view the responses at a given location along the cochlea, s, as a linear filtering operation described by:

$$y_1(t,s) = x(t) *_t h(t,s) \tag{1}$$

where an impulse response $h(t,s)$, referred to as a *cochlear filter*, can be associated with each point of the cochlea to describe the response $y_1(t,s)$ caused by the input signal $x(t)$, and $*_t$ denotes the convolution in the time domain. In mammals, the cochlear filters are generally well tuned with the characteristic center frequencies (CF's) decreasing from the base of the cochlea towards the apex. Over a wide frequency range, the impulse responses can be functionally related by a dilation, i.e.,

$$h(t,s) = h_m(a^s t) \tag{2}$$

for some dilation parameter a and seed filter $h_m(t)$. As such, $y_1(t,s)$ can be interpreted as an affine wavelet transform of the stimulus $x(t)$. While the exact shape of these filters is not critical in establishing any of the results here, certain salient features of the filters are. For instance, it is important that the filters have asymmetric shapes with steep high cut-offs. However, for engineering applications of our results, this constraint can be relaxed and the determination of the shape of the filter becomes a design problem driven by the specific application at hand. These issues will be discussed in light of our results later in Appendix B.

The transduction stage

Most compressive nonlinearities in sensory processing occur at this stage. They are typically caused by the limited dynamic range of the output of the transducing cells (and neural cells in general) relative to that of the signal. In the auditory system, the transduction nonlinearity is sandwitched between two linear operations. It is preceded by a derivative since the cells are velocity (rather than amplitude) driven, and is followed by a lowpass filter due to the leaky membranes of the cells. Consequently, the cochlear filter outputs are transformed as:

$$y_2(t,s) = g(\partial_t y_1(t,s)) *_t w(t) \tag{3}$$

where $g(\cdot)$ is a compressive nonlinear function, $w(\cdot)$ is a low pass filter, and $y_2(t,s)$ represents the transducer output at each point along the cochlea.

The reduction stage

This is the stage at which a representation of the sensory signal is generated. There is no unique way to define such an operation. Rather, it is dictated by various considerations, primary among them is the nature of the features to be highlighted by the representation. Thus, in the case of the auditory system, representations serving the perception of pitch, timbre, or localization of sound would necessarily emphasize different aspects of the sensory signal. Here we focus on the acoustic spectrum, a fundamental cue for the perception of timbre and the recognition of speech signals. The specific transformation we choose to model is one common to all sensory systems, called *lateral inhibition.* Functionally, the lateral inhibition network (LIN) operates along the scale axis, i.e., across different cochlear channels in our auditory

model. It serves to reduce the correlated activity across the channels, enhancing instead rapid fluctuations along the spatial axis of the cochlea. A simple model of its operation is given by a leaky spatial derivative of the instantaneous outputs of the cochlea:

$$\begin{aligned} y_3(t,s) &= \partial_s y_2(t,s) *_s v(s) \\ &= [g'(\partial_t y_1(t,s))\partial_s\partial_t y_1(t,s)] *_t w(t) *_s v(s) \end{aligned} \tag{4}$$

where $y_3(t,s)$ is an intermediate output of the LIN, and $v(s)$ is a spatial smoothing window. At each point, the output can now be tracked in one of several ways. The sequence below tracks roughly the envelope of $y_3(t,s)$ in time:

$$y_4(t,s) = \max(y_3(t,s),0) \tag{5}$$

$$y_5(t,s) = y_4(t,s) *_t \Pi(t) \tag{6}$$

where the low pass filter $\Pi(t)$ has a time constant at the order of 10 – 20 ms, and $y_5(t,s)$, which is referred to as the *auditory representation* in the following, conveys the information encoded in the envelope of $y_3(t,s)$.

When the nonlinearity is driven into saturation, the equations above can be significantly simplified. This so-called high gain limiting case is common in the auditory system for speech signals at or above moderate sound levels. The nonlinearity in this case can be approximated by a Heaviside function and $g'(\cdot)$ approaches in distribution to a Dirac delta function [16], i.e.,

$$y_3(t,s) = [\delta(\partial_t y_1(t,s))\partial_s\partial_t y_1(t,s)] *_t w(t) *_s v(s) \tag{7}$$

Since the support of delta function contains only the origin and, as shown in [17],

$$\delta(f(t)) = \sum_{t_i \in Z} \delta(t_i)\partial_t f(t_i)$$

where

$$Z = \{t : f(t) = 0\}$$

denotes the zeroes of function $f(\cdot)$, the term in the brackets can be interpreted as a sampling process of the quantity

$$\begin{aligned} \partial_s\partial_t y_1(t,s) &= \partial_t\partial_s(x(t) *_t h(t,s)) \\ &= (\partial_t x(t)) *_t (\partial_s h(t,s)) \end{aligned} \tag{8}$$

divided by $\partial_t^2 y_1(t,s)$, and the sampling instants are determined by the zero crossings of

$$\partial_t y_1(t,s) = (\partial_t x(t)) *_t h(t,s) \tag{9}$$

An intuitive and graphical interpretation can be obtained by noting that on the spatio-temporal domain, the gradient of $\partial_t y_1(t,s)$ evaluated at some zero crossing, say (t_i,s_i), is $(\partial_t^2 y_1(t_i,s_i), \partial_s\partial_t y_1(t_i,s_i))$, therefore the sampled quantity

$$S(t_i,s_i) = \frac{\partial_s\partial_t y_1(t_i,s_i)}{\partial_t^2 y_1(t_i,s_i)}$$

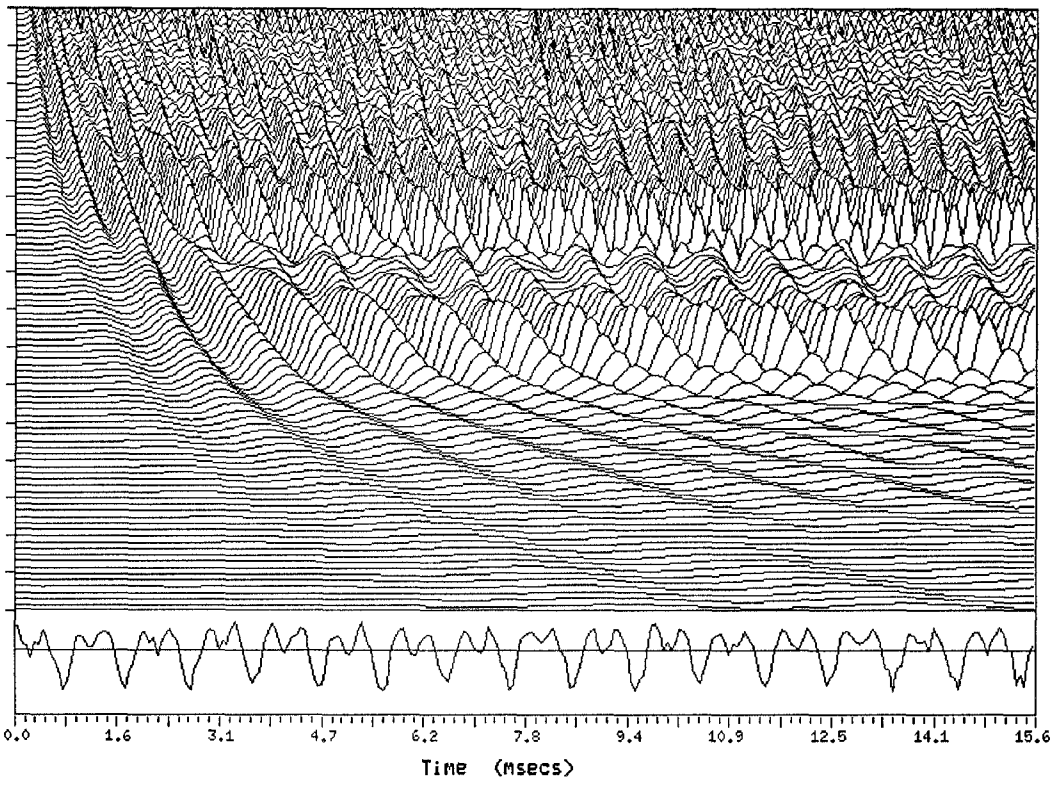


Figure 2: The basilar membrane response of the preemphasized two tone stimulus ($\partial_t y_1(t,s)$).

is the 'slope' of the gradient which is perpendicular to the surface of $\partial_t y_1$ at (t_i, s_i). For example, fig. 2 shows the $\partial_t y_1(t,s)$ of a signal that is composed of two single tones corrupted by a weak white Gaussian noise. Along the zero crossing contour, the slope of the gradient is small at the channels that are coherently responding to a common component until the component is close to the CF of the filter. The spectral contents of the input can be thus analyzed by detecting the 'edge' of the gradient slope. Since $\partial_t y_1(t,s)$ is a bandpass signal, it can be assumed [18] that

$$\partial_t y_1(t,s) = a(t,s)\sin\theta(t,s)$$

where $a(t,s)$ and $\theta(t,s)$ are instantaneous amplitude and phase. In Appendix A we show that, when evaluated at most zero crossings ($\{(t,s) : \theta(t,s) = n\pi\}$),

$$S(t,s) = \frac{\partial_s \theta(t,s)}{\partial_t \theta(t,s)} \tag{10}$$

i.e., $S(t,s)$ is the ratio of the phase variations along the spatial and temporal axes. These phase variations will be associated with a statistic called the *dominant frequency* as discussed in the following section. At relatively much fewer instants, $S(t,s)$ is sampled at the zero crossings of the envelope signal $a(t,s)$. At these locations, it is evaluated as:

$$S(t,s) = \frac{\partial_s a(t,s)}{\partial_t a(t,s)}. \tag{11}$$

III Characteristics of the zero crossings

With minimal assumptions, the stochastic properties of the zero crossings of $y_5(t,s)$ can be concisely described. Although the characterization of zero crossings is possible for continuous time signals [17], the discussion is simpler if $y_5(t,s)$ is properly sampled and turned into a discrete time process. The descretization is plausible since $y_5(t,s)$ is a low passed signal whose bandwidth is no more than that of the low pass filter $\Pi(\cdot)$ in eq. 6. Consequently, all the frequencies referred thereafter are conventionally normalized between $[0,\pi)$ with π being half of the sampling frequency.

General properties of zero crossings

Suppose Y_t is a discrete time random process. When Y_t is passed through the high gain rectifier, the output process will be

$$X_t = \begin{cases} 1, & Y_t \geq 0 \\ 0, & Y_t < 0 \end{cases}$$

Define an indication function as

$$d_t = (X_t - X_{t-1})^2$$

then d_t is either 1 or 0, and whenever $d_t = 1$ corresponds to a zero crossing of Y_t at time t. If Y_t is a weakly stationary Gaussian process with zero mean and autocorrelation function $R(t)$, it can be shown [19] that

$$EX_t = \frac{1}{2} \tag{12}$$

$$EX_tX_{t-1} = \frac{1}{4} + \frac{1}{2\pi}\sin^{-1} r \tag{13}$$

$$Ed_t = \frac{1}{2} - \frac{1}{\pi}\sin^{-1} r \tag{14}$$

where

$$r = \frac{R(1)}{R(0)}$$

is the correlation coefficient between adjacent samples. Under the assumption of stationarity, Ed_t can be interpreted as the zero crossing rate of Y_t since

$$Ed_t = \frac{1}{N}E[\sum_{k=0}^{N-1} d_k]$$

It is also clear from eq. 14 that

$$\cos(\pi Ed_t) = r \tag{15}$$

Therefore, Ed_t also indicates the correlation of the random process: the more the process is positively (negatively) correlated, Ed_t is closer to 0 (1), implying less (more) likely a zero crossing will occur between two adjacent samples.

Now suppose Y_t is composed of a single tone (signal) and noise, i.e.,

$$Y_t = s_t + N_t$$

such a process is generally not stationary and is difficult to analyze. A useful technique to (at least weakly) stationarize the process and thus simplify the analysis without loss of generality is to randomize the phase of the single tone, i.e., let

$$s_t = \sigma \sin(\omega t + \Phi)$$

or

$$s_t = A \cos \omega t + B \sin \omega t$$

in which A and B are uncorrelated (but not independent), zero mean Gaussian random variables with

$$E[A^2] = E[B^2] = \frac{\sigma^2}{2}$$

and σ^2 is the (deterministic) power of the tone. For any non-random signal, its phase can be viewed as being randomly chosen before the beginning of the process. In other words, the non-random signal can be treated as an ensemble of the process and therefore applies to the statistical properties.

Suppose the noise N_t is a zero mean Gaussian noise with a covariance function $\rho(t)$ that is independent of s_t. Then clearly Y_t is a zero mean Gaussian process with correlation function

$$\begin{aligned} R(t,\tau) &= E[A^2]\cos(\omega\tau)\cos(\omega\tau + \omega t) + E[B^2]\sin(\omega\tau)\sin(\omega\tau + \omega t) + \rho(t) \\ &= \frac{\sigma^2}{2}\cos \omega t + \rho(t) \\ &= R(t) \end{aligned}$$

therefore, Y_t is weakly stationary. By eq. 15, it follows that

$$\cos(\pi E d_t) = \frac{\frac{\sigma^2}{2}\cos\omega + \rho(1)}{\frac{\sigma^2}{2} + \rho(0)}$$

If the signal to noise ratio (SNR) $\sigma^2/\rho(0) \gg 1$, since $\rho(1) \leq \rho(0) \ll \sigma^2$, we have

$$\cos(\pi E d_t) \longrightarrow \cos(\omega)$$

Note that ω/π is the zero crossing rate for a single tone with frequency ω. This implies that under a sufficiently large SNR, the frequency of the tone can be unbiasedly estimated by calculating the zero crossing rate and during which process, the effects generated by noise are suppressed.

Zero crossings of the cochlear outputs

Now we can further examine the sampling instants in eq. 7, i.e., the zero crossings of $\partial_t y_1(t,s)$. Suppose the signal $x(t)$ is composed of a single tone with frequency ω_0 and a white Gaussian noise with unit power. It can be shown that the power spectral density of the noise after the filter $H(\omega, s)$ is

$$\omega^2 |H(\omega, s)|^2$$

therefore, the correlation function for the noise

$$\rho(t,s) = \partial_t h(t,s) *_t \partial_t h(-t,s)$$

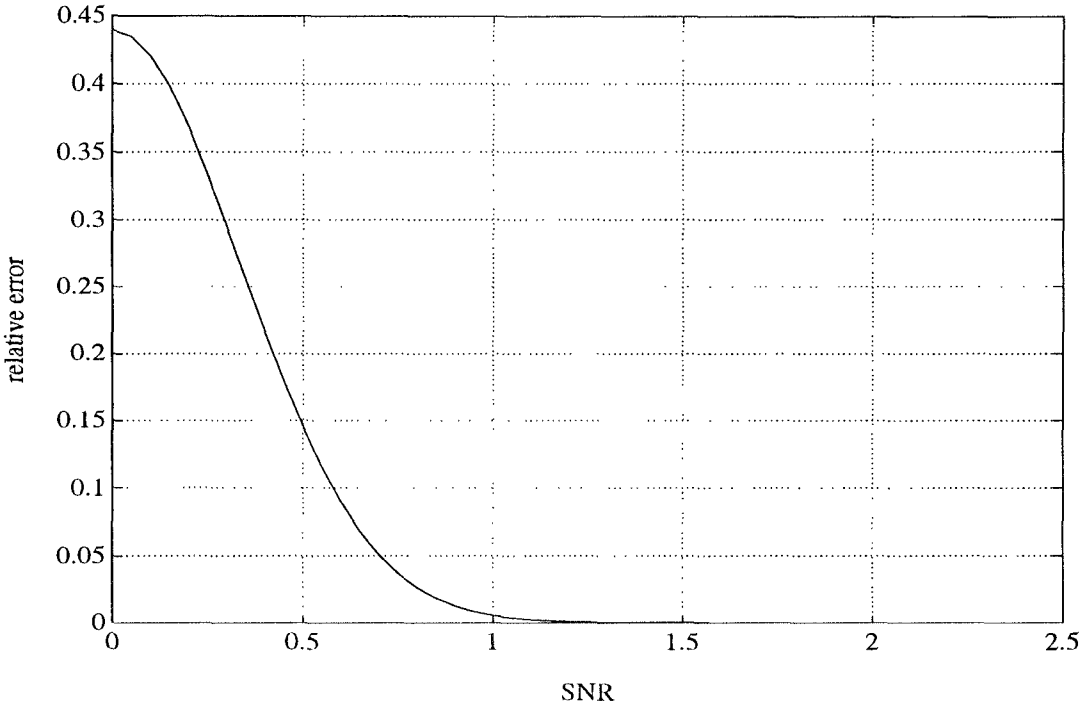


Figure 3: The percentage error of estimating tone frequency with zero crossing rate.

A single tone, on the other hand, preserves its frequency ω_0 after filtering while its power is amplified by $|H(\omega_0, s)|^2$. By applying eq. 14 or 15, the accuracy of detecting the tone frequency by calculating the zero crossing rate at certain s can be evaluated. For example, fig. 3 depicts the percentage error versus the SNR in estimating an 800 Hz tone with the filter whose CF is around 1 kHz. An estimation with 99% accuracy can be obtained when the SNR is as low as 0.42 dB.

The dominant frequency principle

An interesting statistic of the zero crossing rate is known as the dominant frequency [20]. Suppose Y_t has complex components instead of just a single tone, i.e.,

$$Y_t = \sum_i A_i \cos(\omega_i t) + B_i \sin(\omega_i t)$$

By assuming A_i's and B_i's are uncorrelated, a formula analogous to eq. 15 can be obtained:

$$\cos(\pi E d_t) = \frac{\sum_i \sigma_i^2 \cos(\omega_i)}{\sum_i \sigma_i^2} \tag{16}$$

The above equation implies that the resultant zero crossing rate of a compound signal is, through a cosine relationship, the 'center of mass' of its components (fig. 4). In the auditory model, the signal is passed through a filter bank, implying the intensity of each component is weighted by the cochlear filters. When the noise is added, the above equation can be rewritten by applying the rule of 'center of mass':

$$\cos(\pi E[d_t; s]) = \frac{\rho_s(1) + \sum_i |\omega_i H(\omega_i; s)|^2 \sigma_i^2 \cos(\omega_i)}{\rho_s(0) + \sum_i |\omega_i H(\omega_i; s)|^2 \sigma_i^2} \tag{17}$$

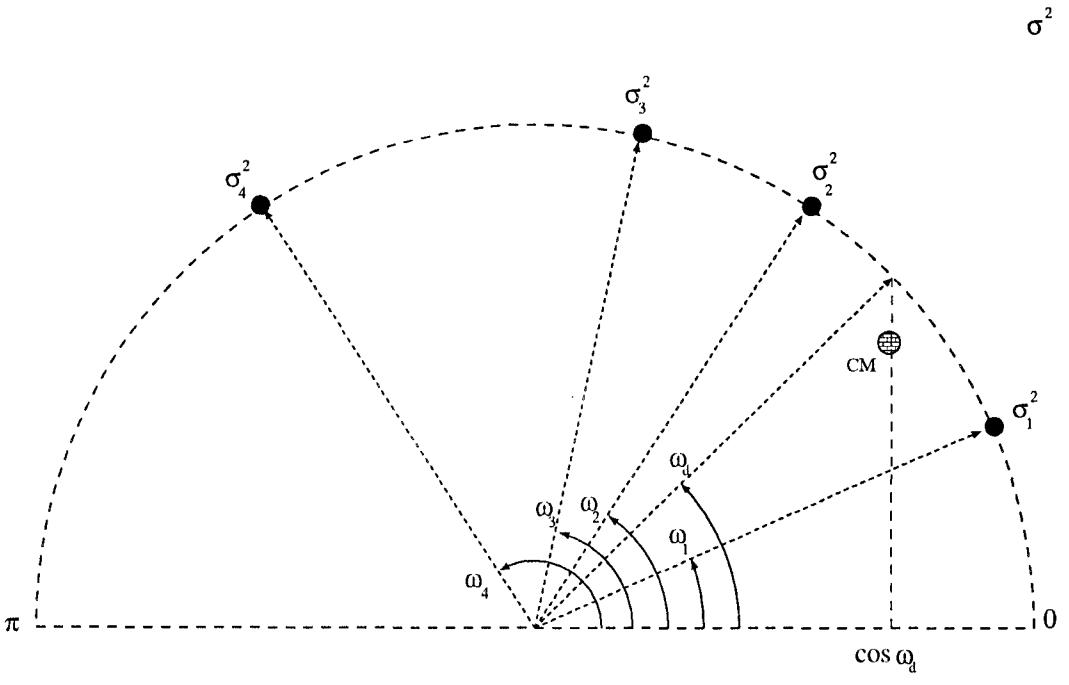


Figure 4: A schematic description of the dominant frequency principle. The components of the signal can be imagined as particles distributed on the unit circle of a plane. The location and the mass of each particle is determined by the frequency and the power of the corresponding component respectively. The resultant dominant frequency is the angle associated with the location of the CM.

If a certain component ω_d is stronger than others after filtering, it is clear that

$$E[d_t; s] \simeq \omega_d/\pi \tag{18}$$

When eq. 18 holds, we say ω_d *dominates* channel s (in the sense of zero crossings), and ω_d is the *dominant* frequency of this channel. Note that the shape of cochlear filters plays an important role in determining the dominant frequency for each channel. As an example, consider again the two-tone stimulus corrupted by a white Gaussian noise of which $\partial_t y_1(t, s)$ is shown in fig. 2. The two tones have the same intensity (σ) and are 6 dB higher than the noise. The dominant frequencies are computed from eq. 17 and depicted in fig. 5. In contrast, the CF's of the cochlear filters are also shown. It can be seen that the tones both dominate the zero crossing rates over a wide range (close to 1/3 octave) which corresponds to the width of the cochlear filters. Beyond the filters, the zero crossing rate steadily increases because of the existence of the noise. As described in the previous section, the zero crossings determine the sampling process of the spectral decomposition of the input stimulus. When a channel is dominated by a frequency other than the CF, the spectral components in this channel are not "properly" sampled and resolved in the auditory representation. This results in the suppression effect, which is further discussed in the next section.

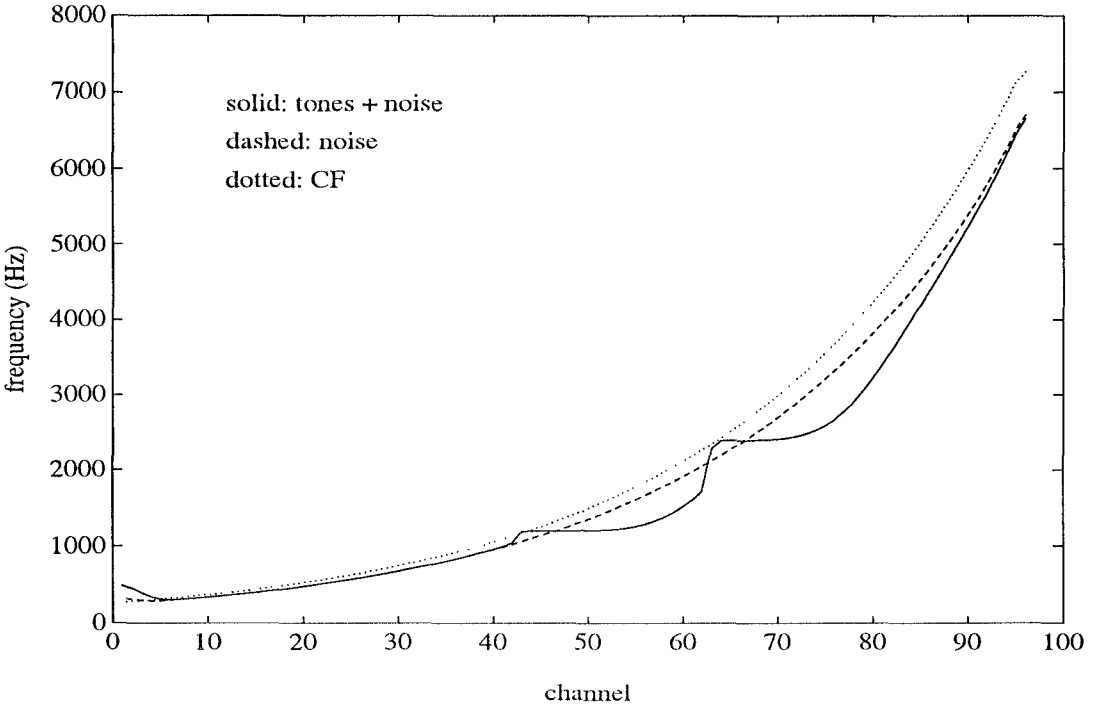


Figure 5: The zero crossing rate for the two tone stimulus in noise. The tones are 1024 Hz and 2048 Hz and are 6 dB higher than the white noise. Also shown are the center frequencies of the differential filters and the zero crossing rates in the absence of the tones.

IV The origin of suppression in the auditory representations

In this section, we discuss the properties and the physical meaning of the auditory representation as expressed by $S(t,s)$, the envelope of $y_3(t,s)$,

$$S(t,s) = \frac{\partial_s \partial_t y_1(t,s)}{\partial_t^2 y_1(t,s)}.$$

First note that, from eq. 8, the numerator of $S(t,s)$ can be interpreted as a wavelet transform of the preemphasized signal $\partial_t x(t)$ induced by the *differential filters* $\partial_s h(t,s)$ instead of the cochlear filters $h(t,s)$. Analogous to eq. 2 in the frequency domain,

$$H(\omega,s) = H_m(\omega/a^s)$$

The differential filter is

$$\partial_s H(\omega,s) = H'_m(\omega/a^s)(-(\log a)/a^s)\omega$$

when compared to

$$\partial_\omega H(\omega,s) = H'_m(\omega/a^s)/a^s$$

we have

$$\partial_s H(\omega,s) = (-\log a)\omega\partial_\omega H(\omega;s) \tag{19}$$

As shown in Appendix B, the cochlear filters are designed to have asymmetric shapes such

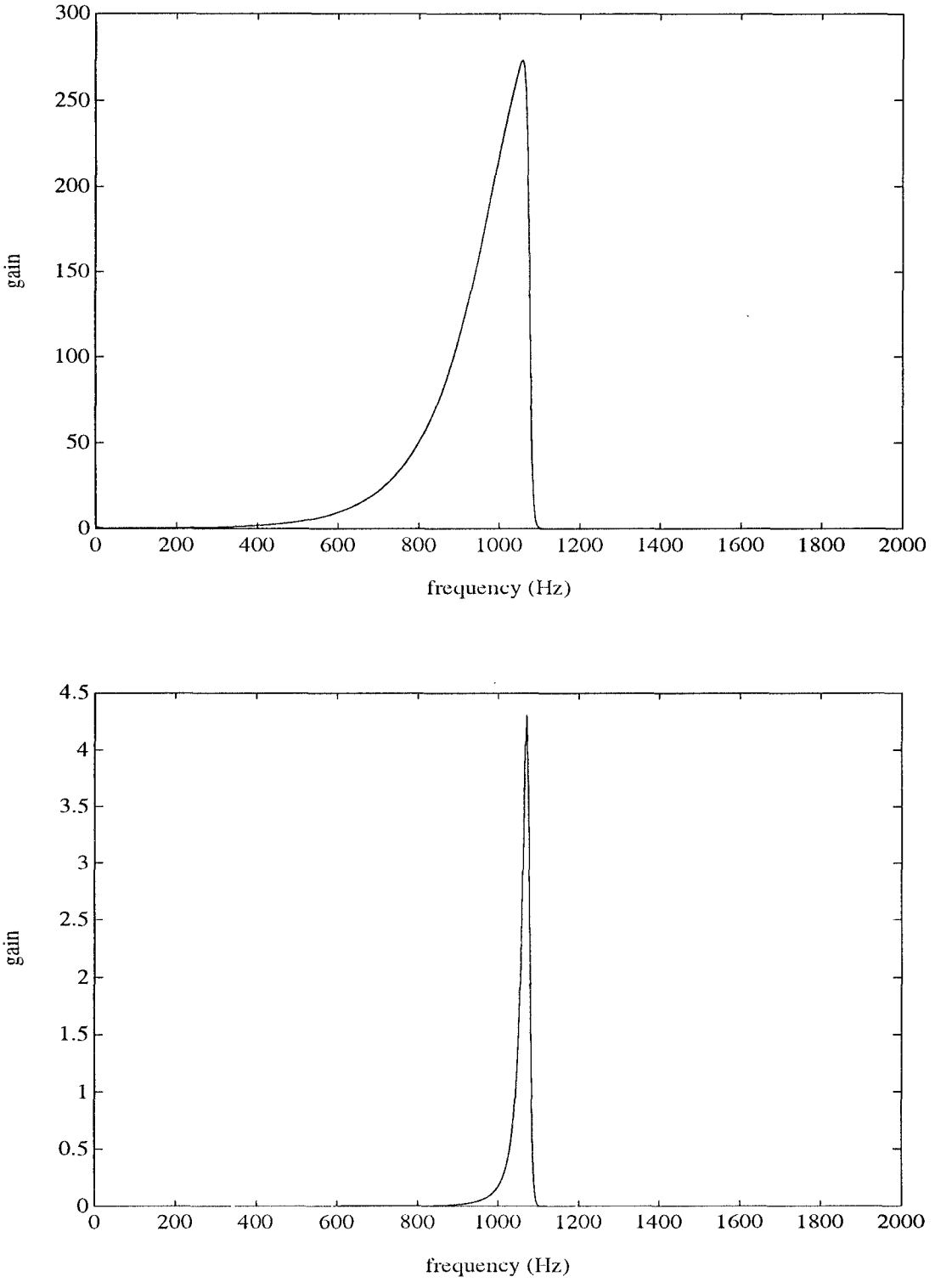


Figure 6: The seed cochlear filter (top) and its corresponding differential filter with center frequency around 1 kHz.

that, according to eq. 19, the differential filter has a very narrow bandwidth (for example, the cochlear and differential filter at $s = 0$ are shown in fig. 6). Therefore, eq. 8 can be interpreted as a narrow-band spectral decomposition of $x(t)$, of which resolution is determined by the bandwidths of the differential filters. For simplicity, assume that the differential filter is so narrow that it can be approximated in the frequency domain by a delta function located at CF. Eq. 8 can be therefore rewritten as

$$\begin{aligned} \partial_s \partial_t y_1(t,s) &= \frac{1}{2\pi} \int j\omega X(\omega) \partial_s H(\omega, s) e^{j\omega t} d\omega \\ &\approx \omega_{cs} |X(\omega_{cs})| \cos(\omega_{cs} t + \phi_{cs}) \end{aligned} \tag{20}$$

where ω_{cs} and ϕ_{cs} are the CF and the phase of the differential filter at s. Note that if the transducer compressive nonlinearity did not exist, i.e., $g(\cdot)$ is equal to the identity function, there would not have been a delta function term in eq. 7, and the auditory representation $y_5(t,s)$, which traces the envelope of $y_3(t,s)$, would have simply become a short-time spectrum of $x(t)$. With the nonlinearity, a sampling process and a normalizing factor are introduced.

Similarly, from eq. 9,

$$\begin{aligned} \partial_t y_1(t,s) &= (\partial_t x(t)) *_t h(t,s) \\ &= a(t,s) \sin\theta(t,s) \end{aligned}$$

can also be interpreted as a spectral analyzing process with the broader full cochlear filters. If we now specify the phase $\theta(t,s) = \omega_{ds} t + \phi_{ds}$ in channel s, then we can think of $a(t,s)$ as the baseband signal that results from modulating down $a(t,s)\sin\theta(t,s)$ by ω_{ds}. In this case,

$$\partial_t y_1(t,s) = a(t,s) \sin(\omega_{ds} t + \phi_{ds}) \tag{21}$$

where ω_{ds} is the dominant frequency of channel s.

Now consider

$$\partial_t^2 y_1(t,s) = \omega_{ds} a(t,s) \cos(\omega_{ds} t + \phi_{ds}) + \partial_t a(t,s) \sin(\omega_{ds} t + \phi_{ds})$$

When evaluated at the zero crossings, the second term vanishes and $\cos(\omega_{ds} t + \phi_{ds}) = \pm 1$. However, if we define

$$\begin{aligned} \Delta\omega_s &= \omega_{cs} - \omega_{ds} \\ \Delta\phi_s &= \phi_{cs} - \phi_{ds} \end{aligned}$$

it is easy to verify that at zero crossings,

$$\frac{\cos(\omega_{cs} t_i + \phi_{cs})}{\cos(\omega_{ds} t_i + \phi_{ds})} = \cos(\Delta\omega_s t_i + \Delta\phi_s)$$

Therefore, the final auditory representation traces the envelope of

$$S(t,s) \approx \left[\left| \frac{\omega_{cs} X(\omega_{cs})}{\omega_{ds} a(t,s)} \right| \cos(\Delta\omega_s t + \Delta\phi_s) \right]. \tag{22}$$

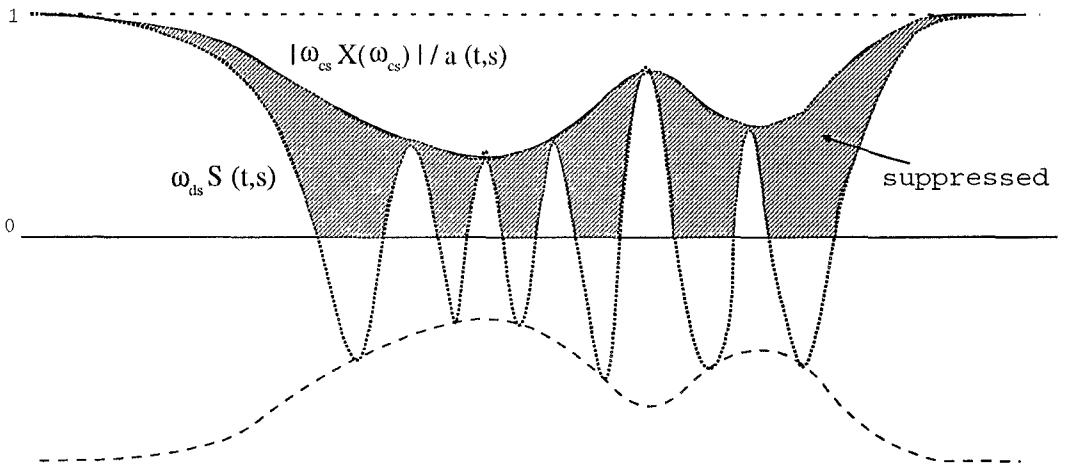


Figure 7: An illustration of the suppression effects in the auditory represenation.

There are two important observations to make about this expression. First is that for channels in which there is only a small contribution from spectral components other than ω_{cs}, i.e., the dominant frequency ω_{ds} is equal to ω_{cs}, or $\Delta\omega_s = \Delta\phi_s = 0$, $a(t,s)$ becomes approximately a constant reflecting the instantaneous energy in the cochlear filter. And therefore the envelope represents a normalized measure of the strength of the signal at $X(w_{cs})$. Second is that this auditory representation is *suppressed* by a cosine factor that results from the discrepancy of the dominant frequency and the CF of the channel. This attenuation effect is illustrated in fig. 7. Note that $a(t,s)$ ceases to be a constant in this case as smaller ac components begin to be added to it. The amount of suppression can be estimated from eq. 22 by applying the center of mass principle (fig. 4) to

$$\begin{aligned} \Delta\omega_s &= \omega_{cs} - \omega_{ds} \\ &= \omega_{cs} - \cos^{-1}\left[\frac{\int_0^\pi |\omega X(\omega) H(\omega,s)|^2 \cos\omega d\omega}{\int_0^\pi |\omega X(\omega) H(\omega,s)|^2 d\omega}\right] \end{aligned} \tag{23}$$

However, since both $a(t)$ and $\Delta\omega_s$ are affected by the change in the dominant frequency, the suppression is not readily computable from eq. 22. Instead, we can use eqs. 10 and 11 directly and noting that we can model each channel as being driven by a quasi-single tone at frequency ω_{ds}. Since in our severely asymmetric cochlear filters, ω_{ds} is always lower than ω_{cs}, this implies that the dominant frequency is approximately constant around the dominated channel, i.e., $\partial_s \omega_{ds} = 0$. Therefore,

$$\partial_s \theta(t,s) = \partial_s(\omega_{ds} t + \phi_{ds}) = \partial_s \phi_{ds}.$$

The phase term is defined as

$$\partial_s \phi_{ds} = \partial_s \arg H(\omega_d, s)$$

where $\arg H(\cdot, s)$ denotes the phase of the filter at s. Since the filters are related by a dilation, the derivative with respect to s can, in the same manner as in eq. 19, be transformed into a derivative with respect to ω. More specifically,

$$\partial_s \arg H(\omega_d, s) = (-\log a)\omega_d \partial_\omega \arg H(\omega_d, s)$$

Therefore, from eq. 10,

$$\begin{aligned} S(t,s) &= \frac{\partial_s \theta(t,s)}{\partial_t \theta(t,s)} \\ &= \frac{(-\log a)\omega_d \partial_\omega \arg H(\omega_d, s)}{\omega_d} \\ &= (-\log a)\partial_\omega \arg H(\omega_{cs} - \Delta\omega_s, s) \\ &= (-\log a)\partial_\omega \arg H_m(\omega_m - \Delta\omega_m) \end{aligned} \tag{24}$$

where ω_m is the CF the seed filter and $\Delta\omega_m = \Delta\omega_s / a^s$ is the *effective* frequency shift of $\Delta\omega_s$ around ω_m. The percentage amount of suppression is therefore

$$1 - \frac{\partial_\omega \arg H_m(\omega_m - \Delta\omega_m)}{\partial_\omega \arg H_m(\omega_m)}$$

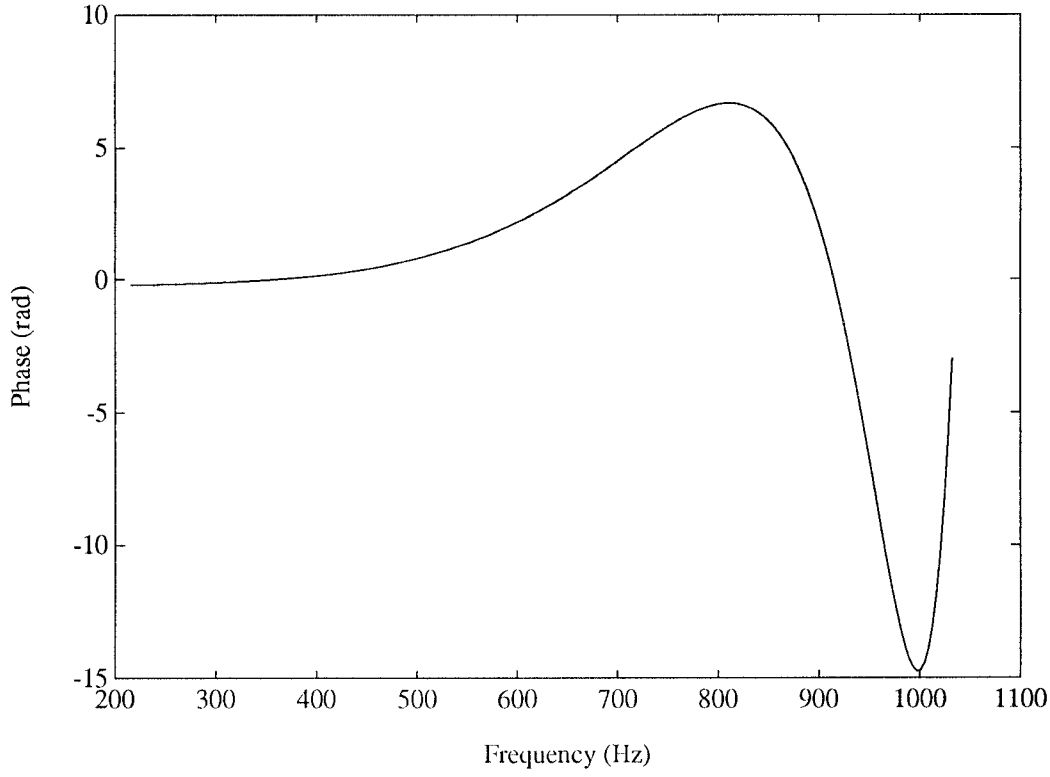
i.e., it traces the curve of the differentiation of the phase in the cochlear filter. The phase function and the differentiation of the seed filter are shown in fig. 8. For this specific filter shown, a 14 Hz down shift in frequency accounts for the first 3 dB attenuation. Note that the ratio of energy that appeared earlier in eq. 22 appears implicitly through the form of the dominant frequency expression in eq. 23. A similar analysis can be applied to the case where the zero crossings are defined by $a(t,s) = 0$ in order to evaluate the ratio $S(t,s) = \frac{\partial_s a(t,s)}{\partial_t a(t,s)}$. Note that $a(.,s)$ has a similar form to that of the phase function in that its derivative is narrow and centered about the CF. Changes in the dominant frequency shift the differential curve in a manner similar to that described above for the phase, causing suppression to appear.

Finally, another more intuitive view of the underlying causes of the suppression can be seen if we consider the sampling process of the $y_3(t,s)$. Roughly, $y_3(t,s)$ can be thought of as the output of a narrow differential filter centered at ω_{cs}. If we think of the zero crossing sampling as a uniform sampling at twice ω_{cs}, then we can assume that half-wave rectifying (in effect halving the sampling rate) will produce the base-band signal near dc reflecting the envelope of $y_3(t,s)$. This in turn is estimated by the final narrow low pass filter (fig. 9). If the sampling rate is now lowered to ω_{ds}, a portion of the base-band signal will shift away from dc, in effect reducing the final output. Thus, suppression can be seen as a direct result of undersampling the signal due to a reduction of the dominant frequency caused by interference from lower frequency components.

V General Discussions

Suppression and the dominant frequency

Suppression as defined and explored so far can be seen as intimately related to the *dominant frequency* principle. As such, it is a direct consequence of using an *average* measure of the rate of zero crossings in representing the signal spectrum. Suppression, thus, should appear in one form or another whenever the dominant frequency measure is used, regardless of the details of the analysis or the representation. For instance, there have been several different algorithms previously proposed to utilize the zero crossing rates on the auditory channels to estimate the



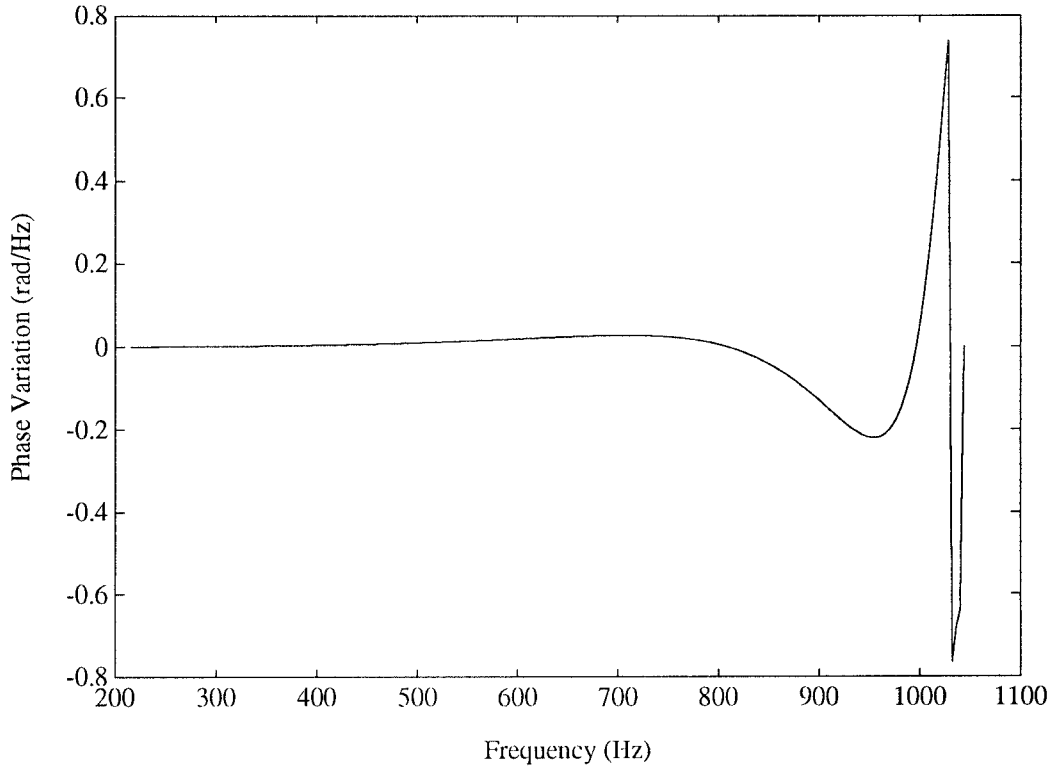


Figure 8: The phase function and its derivative of the seed filter.

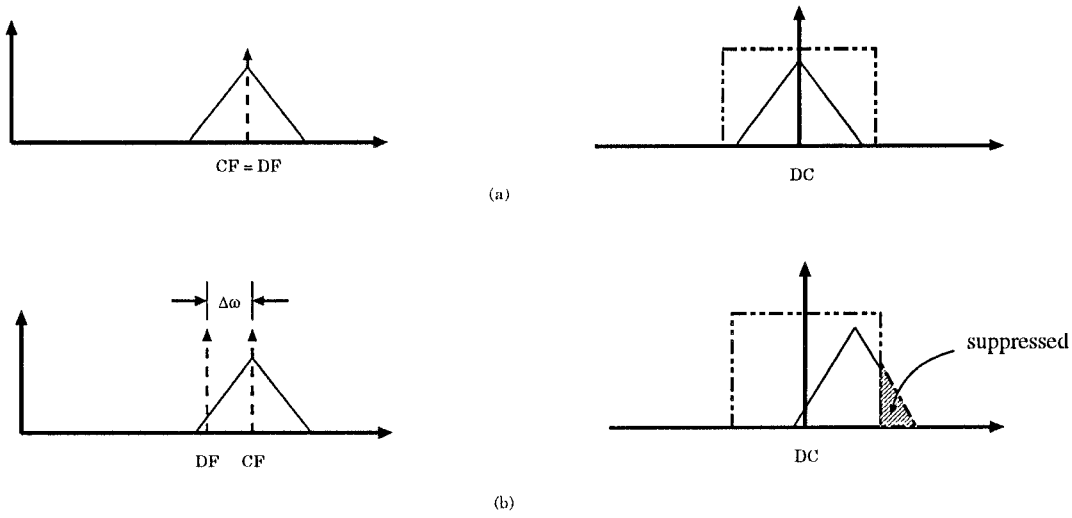


Figure 9: A schematic description of the suppression effects in the signal-driven subband coding method. (a) When the dominant frequency coincides with the center frequency of the anlayzed segment, the spectral contents are preserved after sampling. (b) When the dominant frequency deviates from the center frequency, a portion of the spectral contents will be lost.

signal spectrum [11, 21, 22, 23]. In all these representations, suppression of smaller spectral components by neighboring larger ones is a common finding that can be traced to the use of the *average* rate of zero crossings.

The dominant frequency as a sampling process

In many instants, the dominant frequency is not used directly (e.g., as a count) but rather indirectly, for instance, as the frequency of a sampling process. This is the case in our model where the zero crossings are used to sample the spatial derivative of the phase of the wavelet transform. Assigning these values to the zero crossings allows us to extract a "spectrum-like" pattern relatively easily through a simple average count. Note that this is only possible because of the specific asymmetric design of our filters which had narrow band derivatives. Sampling other measures (e.g., the amplitudes of the peaks as in [24]) or employing different filter shapes would clearly change drastically the extracted pattern and its interpretation. Suppression manifests itself here through the effects of changing the sampling rates in different channels as discussed earlier with reference to (fig. 9).

Zero crossing representations

It has been commonly assumed that encoding signal representations through zero crossings is at best a volatile strategy because of the many instabilities observed in reconstructing these signals when corrupted by noise. This indeed is the case if one attempts to preserve the finest details of the signal through the exact locations of each zero crossing in the trace. In the auditory system, no such thing occurs. Rather, the representations are derived from average measures of the zero crossings. These of course cause the deformity in the representation that we called suppression. However, from a perceptual point of view, such inaccuracies are not only tolerable, but desirable since they are accompanied by enhancements of spectral peaks and noise robustness (as seen in the examples of section VI).

Spectral enhancements and noise robustness through suppression

In the auditory model, the origin of these improvements is the suppression created by the *overlap* in the cochlear channels. For instance, a large spectral component dominates the zero crossings not only of its specific channel, but also of many adjacent channels, thus suppressing the expression of their own smaller spectral components. Intuitively, suppression therefore operates as a highpass filter of the spectral pattern, enhancing the relative expression of nearby peaks while reducing the overall slow variations or tilts in the spectrum. It is in this sense that suppression also maintains the integrity of the representation when high levels of relatively broadband noise is added to the signal. Note that one now can readily appreciate the effects of using cochlear filters with different bandwidths and asymmetries. Thus, the broader the filters, the more overlap occurs across the channels, and hence the stronger are the effects of suppression. The more asymmetric the filters are, the steeper are its edges, and the better is its resolution of the spectral details.

Therefore, the auditory representation encodes the relative, not the absolute, levels of the spectral components. They in fact are available only by virtue of their mutual suppressive interactions. Thus, if spectral components are widely separated compared to the widths of the cochlear filters, their relative levels will, strictly speaking, be lost since no suppressive interactions take place anywhere. This situation, however, is highly pathological for it assumes the absence of any stimulus background or channel noise which may act as a reference activity with which all components may interact.

The role of the nonlinearity

Another important issue concerns the role of the specific shape of the compressive nonlinearity in the coding process. In the model, the nonlinearity was assumed to be centered at the origin, in effect acting as a zero crossing detector. More generally, transducer nonlinearities may be biased, thus acting as level detectors. In our model, this is equivalent to viewing all channels as containing an additional dc bias (a zero-frequency component) which further decreases all the dominant frequencies, each according to the level of its activation. In other words, the bias provides an additional means for estimating the absolute level of each spectral component even when no other across spectral interactions are taking place. This resembles closely the role of the background noise that we discussed earlier.

Another variation on the nonlinearity is the inclusion of a finite dynamic range as opposed to the infinite gain case that we analyzed. Assuming exactly the same processing strategy, the only effect of the finite dynamic range is to convert the sampling process from one using the Dirac delta functions to one using broader, finite amplitude pulses. Even in the extreme case of a half-wave rectifier, i.e.,

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

the model equations and interpretations remain essentially unchanged. For instance, in this case, the derivative $g'(\cdot) = u(\cdot)$, a unit step, and

$$\tilde{y}_3(t,s) = u(\partial_t y_1(t,s))\partial_s\partial_t y_1(t,s) *_t w(t) *_s v(s)$$

The above equation can be interpreted as the spectral decomposition $\partial_s\partial_t y_1(t,s)$ being "sampled" at the positive cycles of $\partial_t y_1(t,s)$. Since the on-off instants are determined by the zero crossings, we may also expect suppression effects in the corresponding model representation.

The inclusion of a finite dynamic range, however, does open up other possibilities for encoding the signal spectrum. For instance, in the example of the half-wave rectifier, one can bypass the zero crossings and the LIN and simply integrate the cochlear outputs over a given time interval. Because of the finite dynamic range, it can be shown that the absolute level of the input spectral components is reflected in the output profile, with a sensitivity that increases with the extent of the dynamic range. Note that the fundamental conclusion here is that provided the average rate of the zero crossing is used in the encoding process, then regardless of the many details, suppression will manifest itself.

Zero-crossings in subband coding

The basic scheme underlying the auditory model and conventional subband coding are similar: the spectrum of the input is divided into segments in the frequency domain, then each segment is sampled at a comparatively low rate. In subband coding, the sampling process for each segment is deterministic and the rate is pre-defined. By carefully manipulating the sampling rate, the original spectrum can be perfectly reconstructed. In the auditory model, on the other hand, the sampling process is contingent upon the dominance of spectral components of the input. Generally speaking, the encoded spectrum is distorted by enhancing the spectral peaks relative to the valleys, and by reducing the overall spectral tilt.

However, since this sort of distortion is perceptually meaningful for acoustic signals, and since it produces robust representations in noise, it may be advantageous to incorporate it in the traditional subband coding scheme. This can be done simply by sampling each frequency segment (the analog of our $\partial_s \partial_t y_1(t,s)$) not at a rate proportional to its bandwidth, but rather at the dominant frequency of zero crossings in a larger segment analogous to $\partial_t y_1(t,s)$. This rate (being approximately half the Nyquist rate of the channel) effectively creates a base-band duplicate of the spectral contents in the narrow frequency segment. For further down sampling, the base-band signal can be extracted with a low pass filter whose bandwidth is equal to the bandwidth of the analyzed frequency segment. Note that if the dominant frequency is equal to the center frequency of the analyzed segment, the above mentioned method is equivalent to conventional approach. However, if the dominant frequency is different, a portion of the spectrum in the analyzed segment will be suppressed, as shown earlier in fig. 9.

In summary, with this proposed coding scheme, the sampling frequency for each segment is determined by the spectral contents of the signal. The amount of suppression and its exact form can be designed through the choice of the larger frequency segments to reflect the specific application at hand. And finally, the encoded representations are robust and perceptually enhanced for acoustic signals.

VI Examples of the enhancement and noise robustness of the auditory representations of speech signals

In this section, we demonstrate the spectral enhancements and noise robustness that result from the suppressive interactions in the auditory model. Naturally spoken vowel segments are used in all computations. We shall first compare the auditory representation against the the power spectrum of the signal viewed with the resolution of the cochlear differential filters. Next, we demonstrate the importance of absolute level independence in the auditory model

when averaging the spectra of a large number of speakers. Finally, we illustrate the superior noise robustness of the auditory representations compared to the power spectrum.

Enhancements of spectral peaks

Fig. 10–11 show waveforms and their corresponding auditory representations and power spectra for vowels /uw/ (as in "two"), /aa/ (as in "what") and /iy/ (as in "tea") spoken by a male speaker The power spectra are obtained by removing the model's transducer nonlinearity in the hair cell, i.e., by setting $g(x) = x$. To facilitate the comparisons, the power spectra are all pre-emphasized and the spatial channels are labeled by the CF's of the differential filters. The first thing to observe about the patterns is that, because of the constant Q cochlear filters, all low frequency harmonics are resolved, whereas at high frequencies, only the formants are preserved. One significant difference between the auditory and power spectrum representations is the enhancement of the peaks relative to the valleys. This is true both for the resolved harmonic peaks and for the overall formant peaks (e.g., note depression of the valley in vowel /iy/). The other difference between the two representations is the reduced spectral tilt in the auditory representation and the normalization of the overall absolute level of the signal. The beneficial effects of this change are demonstrated below when considering the averaged spectra of a large number of speakers.

Encoding of invariant acoustic features

It has been well established [25] that in speaker-independent speech recognition systems, the spectral features to be used must be shared among speakers in various contexts and at different intensities. For speech vowels, most of these features can be associated with the characteristics in the vocal tract in the sound producing process [26]. For example, vowels can be uniquely identified by their place of articulation and degree of constriction (fig. 12), which respectively indicate the point of resonance and the the openness of the vocal tract during the sound production. Generally speaking, as the place of articulation moves towards the glottis (marked "front" in fig. 12), significant responses in the spectrum are expected to move from high to low frequency regions. On the other hand, as the vocal tract becomes more constricted, the relative ratio of the strongest and the second strongest formants becomes larger. To examine how these features are generally encoded in the auditory representations and the power spectra, figs. 13 show the average patterns of the vowels /uw/, /aa/ and /iy/ taken from all the sentences spoken by 25 female and 44 male speakers in the `train/dr5` directory of the TIMIT database. It can be seen that the clue to the place of articulation is well encoded in both representations. For example, front vowel /iy/ has a significant high frequency formant in both representations, while back vowel /uw/ has a low frequency formant and middle vowel /aa/ has one in between. It can also be seen that the high constriction vowels such as /iy/ and /uw/ have larger formant ratio while all the formants of /aa/ are lumped in the mid frequency with equal intensity. Therefore, the overall spectral shape that reveals the salient acoustic features of vowels is encoded in both representations.

However, the relative contribution of the spectral shape to the overall spectrum is much less in the case of the power spectra. For example, it can be seen in fig. 13 that the variation in the spectral shape for the 5 octaves (from 250 Hz to 8kHz) accounts only roughly for (1.222-1.214)/1.22, or less than 0.14% per octave of the power spectra. Such a tiny dynamic range can be easily overwhelmed by a energy change or a spectral tilt such as preemphasis. While

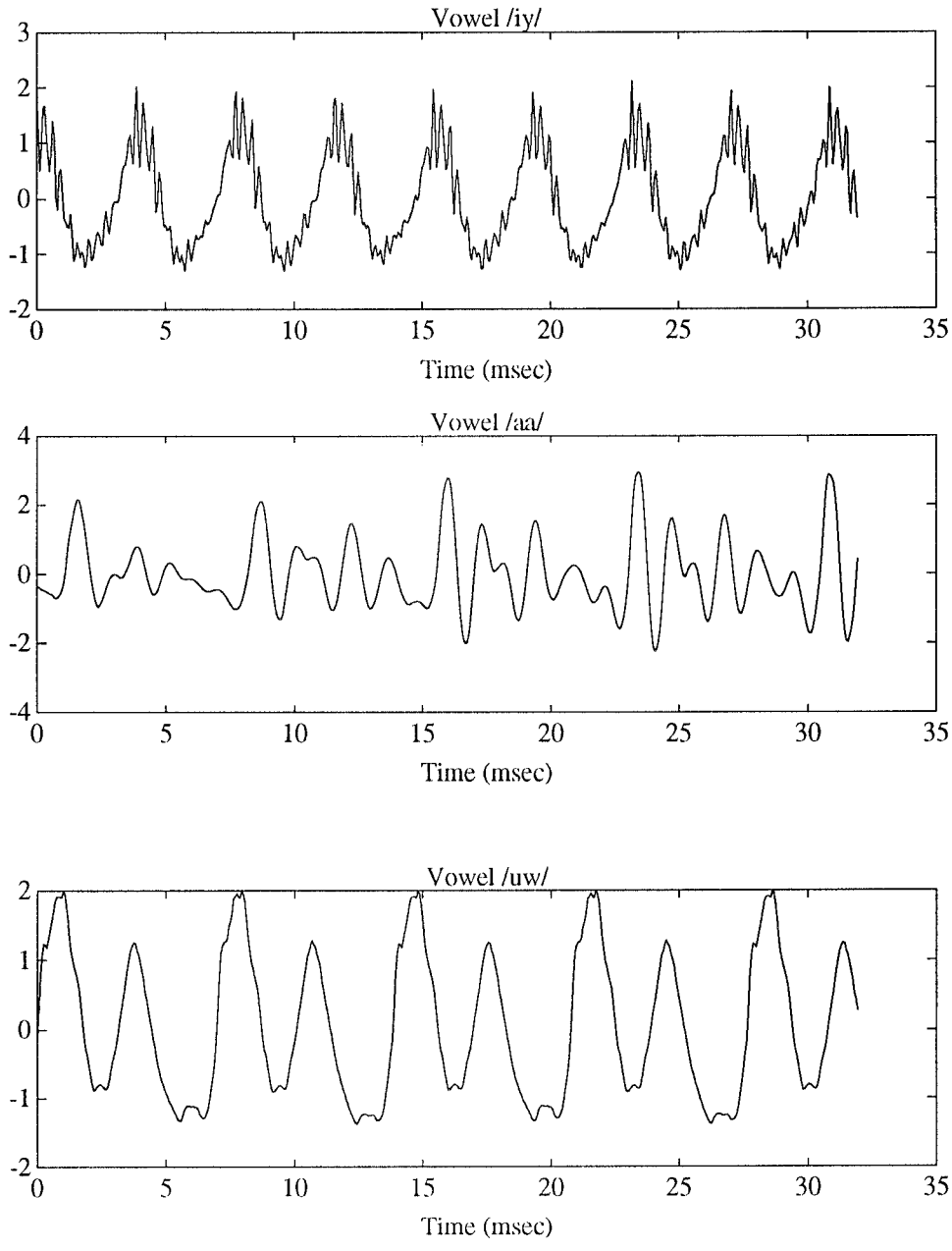


Figure 10: The waveforms of the vowels /iy/,/aa/, and /uw/

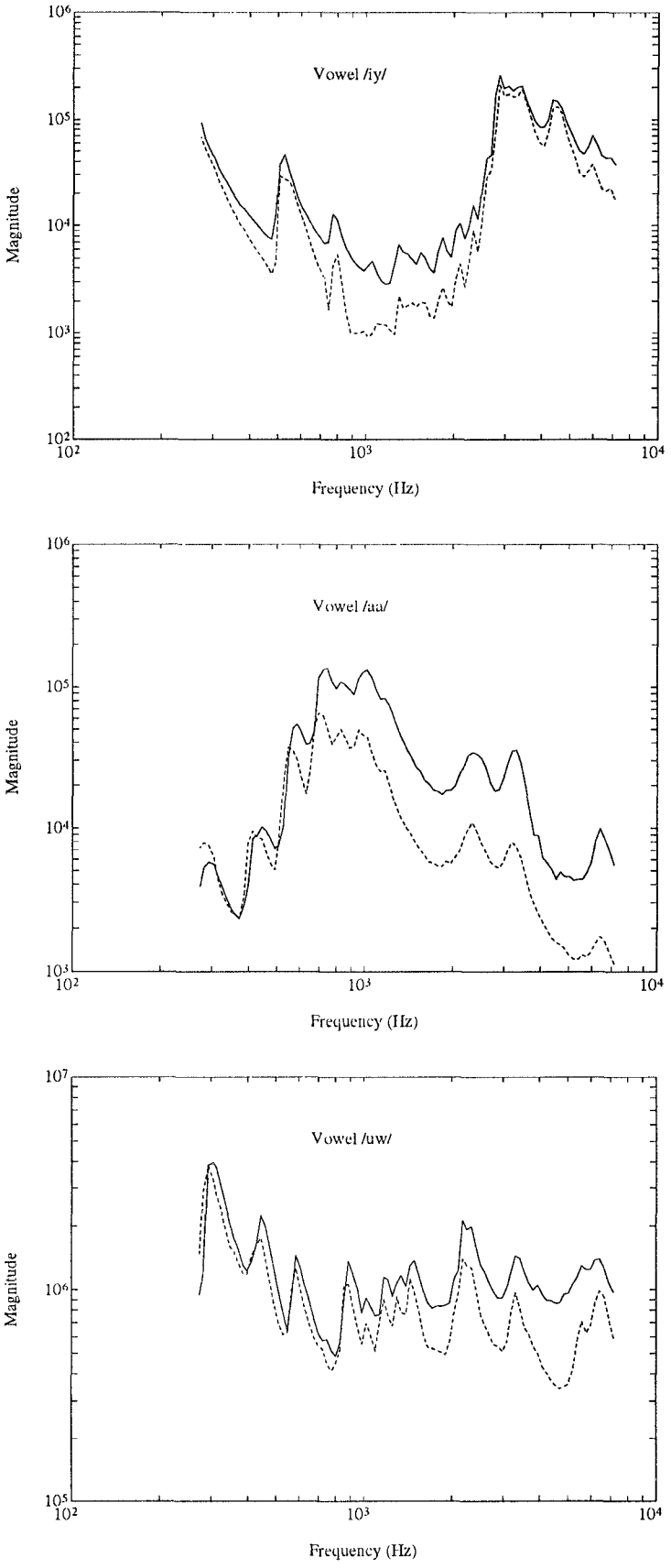


Figure 11: The power spectra (solid) and the auditory representations (dashed) of the vowel /iy/, /aa/, and /uw/ from a male speaker.

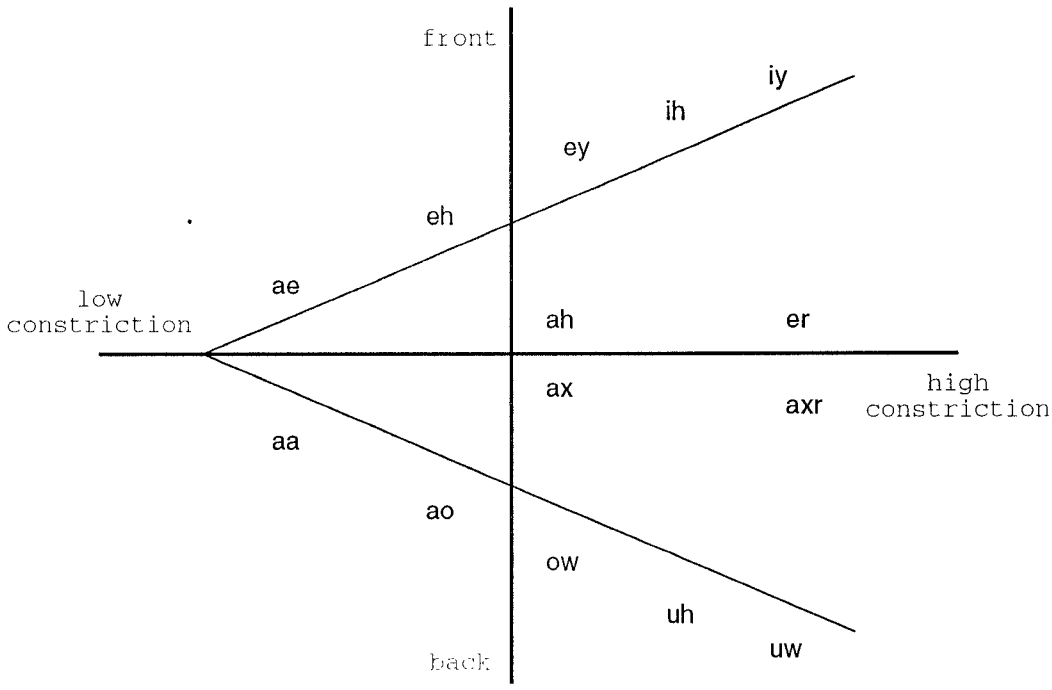


Figure 12: A schematic classification of the vowels in American English. All symbols are defined in the TIMIT database.

the spectral shape information in the power spectra is largely offset by the absolute intensity of the patterns, the auditory representations, on the other hand, enhance the spectral shape by the mutual suppression and discarding of the absolute level.

Noise robustness

To examine noise suppression in the model, we compared the two representations of vowel /uw/ in several levels of white background noise (fig. 14 and 15. In both cases, the SNR is defined as the total signal energy to noise energy over the duration of the vowel. The difference in robustness between the two representations is quite striking. Thus, while the power spectrum representation has already lost all of its higher formant and harmonic peaks, they are all well preserved in the auditory representation even at 0 dB SNR. We are currently in the process of examining the consequences of such differential rates of degradation in the context of a larger speech recognition system.

VII Summary

In summary, the fundamental message of this paper is that encoding a signal by the zero crossing rates of its wavelet transform provides for a robust and enhanced representation. In this process, the wavelet transform furnishes a multiresolution spectral representation of the signal, with significant correlations across the different scales. The extraction of the zero crossing rates imply mutually suppressive interactions among adjacent overlapping scales, which in turn give rise to the spectral enhancements and noise robustness. Since biological sensory systems carry out these operations, our perceptual representations of auditory and visual signals are

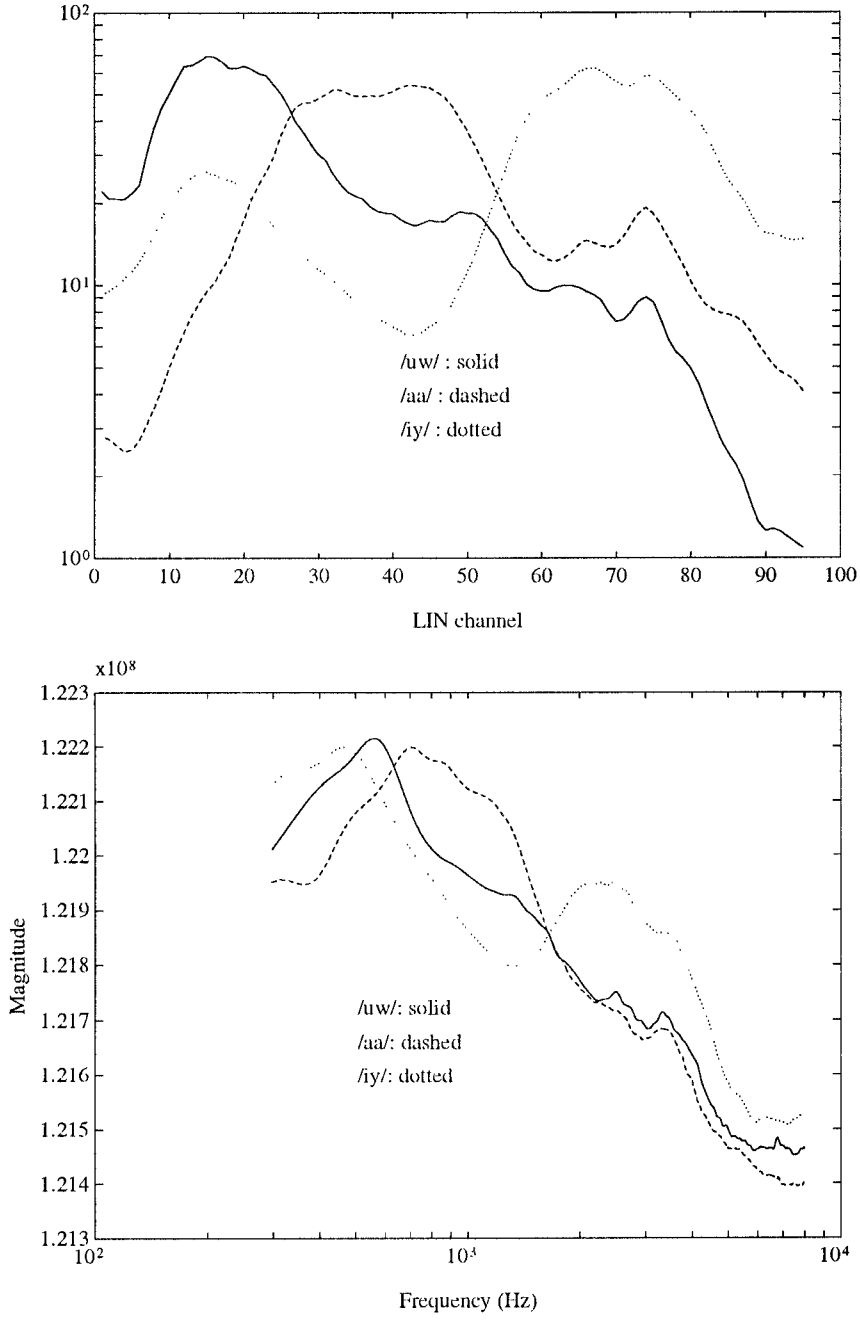


Figure 13: The average patterns of the auditory representations (top) and the power spectra of /iy/, /aa/, and /uw/.

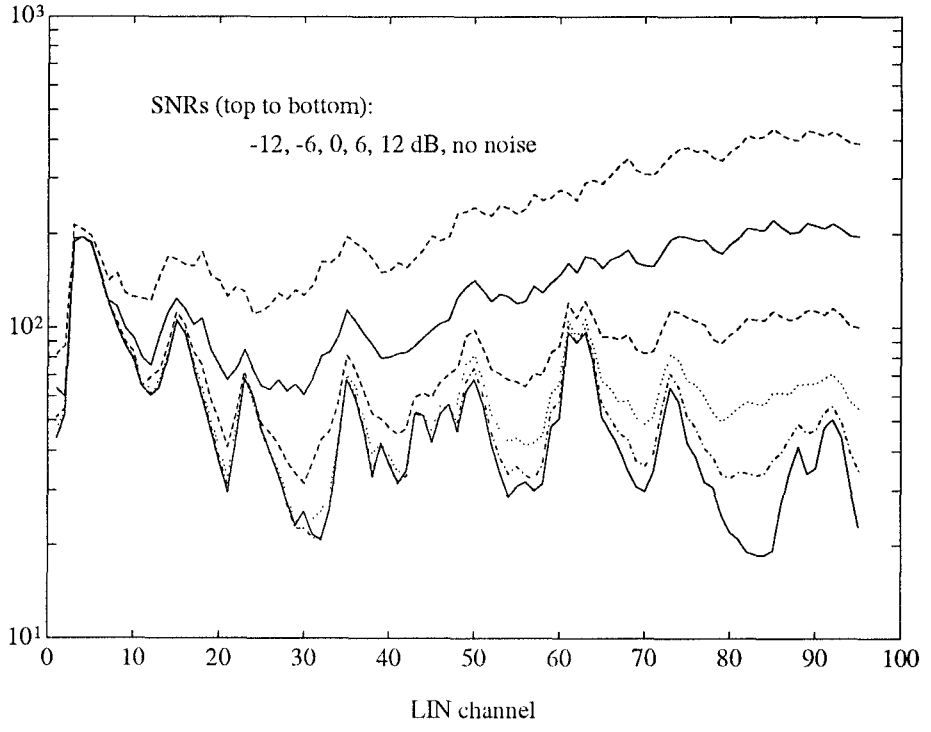


Figure 14: The auditory representations of /uw/ in several levels of background noise.

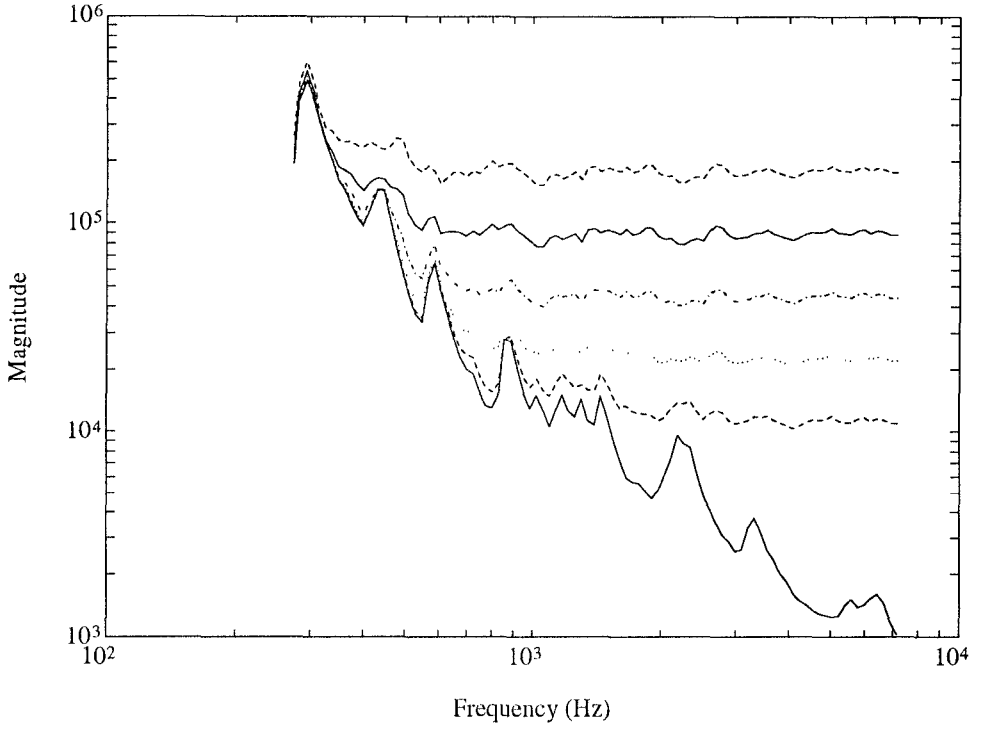


Figure 15: The power spectra of /uw/ in several levels of background noise.

most likely "distorted" similarly. Consequently, it may be advantageous to incorporate such transformations in signal processing systems dealing with such natural signals.

Appendix A Evaluation of $S(t,s)$

In this appendix, we evaluate

$$S(t,s) = \frac{\partial_s \partial_t y_1(t,s)}{\partial_t^2 y_1(t,s)}$$

at the zero crossings of

$$\partial_t y_1(t,s) = a(t,s)\sin\theta(t,s)$$

i.e., at

$$Z = \{(t,s) : \theta(t,s) = n\pi\} \cup \{(t,s) : a(t,s) = 0\}$$

We first show that

$$S(t,s) = \frac{\partial_s a(t,s)}{\partial_t a(t,s)}$$

when evaluated at zero crossings defined by $a(t,s) = 0$, and

$$S(t,s) = \frac{\partial_s \theta(t,s)}{\partial_t \theta(t,s)}$$

when evaluated at the zero crossings defined by $\{(t,s) : \theta(t,s) = n\pi\}$.

It is easy to verify that

$$\begin{aligned} \partial_s \partial_t y_1(t,s) &= \partial_s a(t,s)\sin\theta(t,s) + a(t,s)\cos\theta(t,s)\partial_s\theta(t,s) \\ \partial_t^2 y_1(t,s) &= \partial_t a(t,s)\sin\theta(t,s) + a(t,s)\cos\theta(t,s)\partial_t\theta(t,s) \end{aligned}$$

Assuming non-trivial case, the second terms of both above equations vanish at $a(t,s) = 0$. Therefore, $S(t,s)$ is the ratio of the first terms, namely,

$$S(t,s) = \frac{\partial_s a(t,s)}{\partial_t a(t,s)}$$

Similarly, when $\theta(t,s) = n\pi$, the first terms equal to 0 and

$$S(t,s) = \frac{\partial_s \theta(t,s)}{\partial_t \theta(t,s)}$$

A similar and more geometrical oriented derivation can also be used to evaluate $S(t,s)$ in this case. As explained in section II, $S(t,s)$ is the slope of the gradient along the zero crossing contour, namely,

$$S(t,s) = -1/\left.\frac{ds}{dt}\right|_{(t,s)\in Z}$$

Since the zero crossings are described by the equation

$$\theta(t,s) = n\pi$$

by taking the time-derivative on both sides, we have

$$\partial_t \theta(t,s) + \partial_s \theta(t,s) \frac{ds}{dt} = 0$$

which implies

$$S(t,s) = \frac{\partial_s \theta(t,s)}{\partial_t \theta(t,s)}$$

□

Appendix B Cochlear filter design

As mentioned in the text, the cochlear filters have a direct impact on the performance of the auditory system. Many characteristics of the cochlear filters have been identified through psychophysical experiments as being crucial to the abilities in sound perception, detection and understanding. For the auditory model to exhibit suppression, noise robustness, high resolution, and a fast dynamic response, strict constraints have to be imposed upon the cochlear filters. These issues were discussed in detail in [8], and also partly in the text (sections III-V). The most obvious feature of the filters is their severe asymmetrical shape, with steep roll-offs on the high frequency side, and a slow decay on the low frequency side. The overall effect is that of a relatively broad filter, whose differential with respect to frequency is quite narrow. The broadness of the filter facilitates the detection of rapid transients. while the narrowness of the differential filter provides for high frequency resolution [8].

Our choice of the seed filter, from which all other filters are produced by dilations, is

$$f(x) = x^a e^{-bx} \quad \text{for } x \geq 0 \tag{25}$$

i.e., the magnitude response has a shape similar to gamma function that either side on the peak frequency of the filter assumes an exponential decay, where the x axis here is corresponding to the frequency axis of the filter. This approach is similar to the double exponential function used in [27], and is very close to the semi-tone function popular in cochlear filter approximation [3]. A nice property of the function defined in 25 is that the desired values of the peak frequency, bandwidth, and the symmetry of the filter can all be easily obtained by tuning the two parameters a and b. For instance,

$$f'(x) = x^{a-1} e^{-bx} (a - bx)$$

i.e., a maximum point, for this function, occurs at $x = a/b$. This ratio of a to b determines how close the maximum point is away from the origin, which is corresponding to the roll-off on the high frequency side of the filter. On the other side of the peak frequency, the decay is mostly determined by the term e^{-bx} in 25, i.e., by parameter b alone. Roughly speaking, the bandwidth and the symmetry of the filter are therefore determined by b and a/b respectively. Once a satisfactory filter shape is obtained, it then can be shifted along the x axis so that the its maximum resides on the point corresponding to the desired peak frequency.

In practice, the spatial axis of the model is discretized into finite number of channels. The less channels the model, the better. However, there is a tradeoff among the number of channels,

the frequency range covered by the model, and the frequency resolution of the model. In this paper, we demonstrated a model that covers the frequency band from 250 Hz to 6.7 kHz in 96 channels with a filter density of 20 per octave. By letting the differential filters overlap at 3 dB points, the frequency resolution, which is roughly the ratio of 3-dB bandwidth to the peak frequency of the differential filter, is 1/20, or only 5%. While it may be good enough for many applications, such a resolution is not comparable to the auditory system which has approximately 3000 channels covering the audible band from 150 Hz to 22 kHz with a resolution 0.5% .

References

[1] David Marr. *Vision.* W. H. Freeman and Company, New York, NY, 1982.

[2] Xiaonong Ran. *A three-component image model based on human visual percetption and its applications in image coding and processing.* PhD thesis, Department of Electrical Engineering, Unversity of Maryland, August 1992.

[3] David K. Mellinger. *Event formation and separation in musical sound.* PhD thesis, Department of Music, Stanford University, 1991.

[4] Richard F. Lyon. Computational models of neural auditory processing. In *Proc. IEEE Int. Conf. ASSP*, San Diego, CA, March 1984.

[5] Martin P. Cooke. *Modelling auditory processing and organization.* PhD thesis, Department of Computer Science, University of Sheffield, May 1991.

[6] O. Ghitza. Auditory nerve representation as a front end for speech recognition in a noisy environment. *Computer Speech and Language*, 1:109–130, 1986.

[7] T. K. P. Ngyuen, R. P. Lippmann, B. Gold, and D. B. Paul. A physiologically motivated front end for speech recognition. Technical Report 893, Lincoln Laboratory, MIT, February 1991.

[8] Xiaowei Yang, Kuansan Wang, and Shihab A. Shamma. Auditory representations of acoustic signals. *IEEE Transactions on Information Theory, Special Issue on Wavelet Transforms and Multiresolution Signal Analysis*, 38(2):824–839, March 1992.

[9] Martin S. Silverman, David H. Grosof, and Russell L. DeValois. Spatial-frequency organization in primate striate cortex. *Proceedings of the national academy*, 86(2), January 1989.

[10] W. Byrne, J. Robinson, and S. A. Shamma. The auditory processing and recognition of speech. In *Proceedings of the speech and Natural Language Workshop*, pages 325–331, October 1989.

[11] Jont B. Allen. Cochlear modeling. *IEEE ASSP Magazine*, pages 3–28, January 1985.

[12] Graeme K. Yates. Frequency selectivity in the auditory periphery. In Brian C. J. Moore, editor, *Frequency Selectivity in Hearing*, chapter 1, pages 1–50. Academic Press Inc., 1986.

[13] James O. Pickles. The neurophysiological basis of frequency selectivity. In Brian C. J. Moore, editor, *Frequency Selectivity in Hearing*, chapter 2, pages 51–121. Academic Press Inc., 1986.

[14] S. A. Shamma, R. S. Chadwick, W. J. Wiber, K. A. Morrish, and J. Rinzel. Biophysical model of cochlear processing: Intensity dependence of pure tone responses. *Journal of the Acoustical Society of America*, 80(1):133–145, July 1986.

[15] Shihab A. Shamma. Speech processing in the auditory system I: the representation of speech sounds in the responses of the auditory nerve. *Journal of the Acoustical Society of America*, 78(5):1612–1621, November 1985.

[16] Robert J. Zimmer. *Essential Results of Functional Analysis.* The University of Chicago Press, Chicago, MI, 1990.

[17] Athanasios Papoulis. *Probability, random variables, and Stochastic Processes, 2nd Edition.* McGraw Hill Book Company, 1984.

[18] Athanasios Papoulis. *The Fourier integral and its applications.* McGraw-Hill Book Company Inc., New York, NY, 1962.

[19] J. T. Barnett and B. Kedem. Zero-crossing rates of functions of gaussian processes. *IEEE Transactions on Information Theory*, 37(4):1188–1194, July 1991.

[20] Benjamin Kedem. Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE*, 74(11):1477–1493, November 1986.

[21] O. Ghitza. Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment. *Journal of Phonetics*, 16(1):109–204, 1988.

[22] M. B. Sachs and E. D. Young. Encoding of steady state vowels in the auditory-nerve: Representation in terms of discharge rate. *Journal of the Acoustic Society of America*, 66:470–479, 1979.

[23] L. Deng, C. D. Geisler, and S. Greenberg. A composite model of the auditory periphery for the processing of speech. *Journal of Phonetics*, 16(1), 1988.

[24] Stephane G. Mallat. Multifrequency channel decompositions of images and wavelet models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(12):2091–2110, December 1989.

[25] P. W. Jusczyk. Toward a model of the development of speech perception. In J. S. Perkell and D. H. Klatt, editors, *Invariance and variability in speech process.* LEA Publisher, 1986.

[26] James L. Flanagan. *Speech Analysis Synthesis and perception, 2nd edition.* Springer-Verlag, New York, NY, 1972.

[27] B. R. Glasberg and B. C. J. Moore. Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. *Journal of Acoustical Society of America*, 79(4):1020–1033, April 1986.