TECHNICAL RESEARCH REPORT

Institute for Systems Research

Self-Normalization and Noise-Robustness in Early Auditory Representations

by K. Wang and S. Shamma

The Institute for Systems Research is supported by the National Science Foundation Engineering Research Center Program (NSFD CD 8803012), Industry and the University

TR 93-47 r1

Self-Normalization and Noise-Robustness in Early Auditory Representations[1]

Kuansan Wang and Shihab Shamma
Institute for Systems Research and Department of Electrical Engineering
University of Maryland, College Park, MD 20742
Phone: (301)405-6596, E-mail: kuansan@src.umd.edu

Abstract

A common sequence of operations in the early stages of most sensory systems is a multiscale transform followed by a compressive nonlinearity. In this paper, we explore the contribution of these operations to the formation of robust and perceptually significant representation in the early auditory system. It is shown that auditory representation of the acoustic spectrum is effectively a self-normalized spectral analysis, i.e., the auditory system computes a spectrum that is divided by a smoothed version of itself. Such a self-normalization induces significant effects such as spectral shape enhancement and robustness against scaling and noise corruption. Examples using synthesized signals and a natural speech vowel are presented to illustrate these results. Furthermore, the characteristics of auditory representation are discussed in the context of several psychoacoustical findings, together with the possible benefits of this model for various engineering applications.

1 Introduction

Sound signals undergo a series of complex transformations in the early auditory system. These transformations convert the acoustic spectrum of the stimulus into an internal representation which we shall call the *auditory spectrum*. Higher central auditory stages analyze further the auditory spectrum into more elaborate representations, interpret them, and eventually form corresponding sound percepts. Therefore, determining the characteristics of the auditory spectrum and the underlying auditory processing that give rise to it is critical for a deeper understanding of the basic perceptual elements of sound. This in turn would inspire novel signal processing algorithms that are perceptually oriented, and suggest new representations of sound signals that might prove useful in such applications as automatic speech recognition (ASR) systems and underwater acoustics.

It is with these hopes that many efforts were launched over the years to develop auditory-based front-ends for ASR systems [8, 18, 26, 13]. It is fair to say, however, that despite repeated demonstrations of enhanced performance (ranging from increased noise-robustness [8], better generalizations across speakers [4], to improved phonemic representations [28]), auditory front-ends did not progress beyond the research laboratory. Instead, most ASR systems continue for various reasons to use more traditional representations such as LPC-based cepstral coefficients, or perhaps slightly modified versions of them. This is partly due to the apparent complexity and nonuniformity of the proposed auditory models, and the heavy computational loads they demand compared to the benefits gained.

From a signal processing perspective, the succession of linear, nonlinear, adaptive, and cross channel processing stages that are known to occur in the auditory system makes it difficult to come up

[1]This article is scheduled to appear in *IEEE Transactions on Speech and Audio Processing* in July, 1994

with a clear coherent statement of the exact nature of the auditory spectral representations. Instead, it is often necessary to resort to simulations and large speech recognition experiments to demonstrate overall system improvements. In such an exercise, one often loses sight of the exact mechanisms or stages that give rise to the improvements and ways to further enhance the performance.

With these problems in mind, we proposed earlier a "minimal" auditory model [38] in which a few of the most important stages were preserved and certain simplifying analyses were introduced. The goals of the work were two folds: (1) to show what the auditory spectrum looked like, and (2) to demonstrate that, despite the nonlinearities and apparent distortions in the auditory spectrum compared to the acoustic spectrum, no information was lost since fairly accurate reconstructions of the original signal were still possible.

There were two shortcomings of that work that we seek to tackle in this paper. The first is that the *deterministic* analytical framework of the earlier model is still too difficult to manipulate. For instance, it is difficult to use it to predict quantitatively such things as the enhanced peak-to-trough ratios in the auditory spectrum, the expected amount of noise suppression, or simply what the auditory spectrum is in relation to the acoustic spectrum. The second problem is the simplifying omission in the earlier paper of a normalizing factor in the auditory spectrum, which made it look less distorted when compared to the original acoustic spectrum.

In this paper, we shall continue to use the earlier "minimal" model of the early auditory system (critically reviewed in section 2 below). This work, however, proposes a new stochastic framework for the analysis of the early auditory system which employs a *random* source model. Our goal is to explain and predict the general trends and properties of early auditory processing, rather than to pursue an exact and deterministic description on the model output for any specific signal. This approach is motivated by several factors. First, many signals are random in nature and can usually be well described by some stochastic quantities, e.g., the autocorrelation and power spectrum[24, 2]. Second, the analysis task of a signal processing system is usually significantly simplified if the system performance is described in terms of a more general underlying signal space that is associated with certain probability measure (i.e., a probability space). The signal, as argued in [10], can always be viewed as a sample from a signal space of which parameters are randomly chosen before the signal starts. The performance of system, described in terms of the responses to the signal space as a whole, may roughly be applied for the sample in a statistical sense. For example, any single tone of frequency ω belongs to a space

$$\{R\sin(\omega t + \Theta)\}$$

in which the amplitude R and phase Θ are random variables. If R has a Rayleigh distribution and Θ is uniformly distributed over $[0, 2\pi]$, the signal is indeed a zero mean, stationary Gaussian process [20]. While it is difficult to obtain a detailed closed-form description for the system responses to a specific single tone, a concise expression is usually possible for a stationary Gaussian process. The usefulness of this approach is well demonstrated, for example, by the characterization of zero crossings and spectral analysis in [16, 3].

The organization of this paper is as follows. First, we shall review in section 2 the basic outlines of the auditory model. In section 3, we present the stochastic framework for the analysis and discuss how

the auditory spectrum is interperted under various conditions and simplifications. Next we illustrate the salient properties of the auditory spectrum and compare them to the analytical predictions. Four specific examples of acoustic signals are next considered in section 4: a single tone in broadband noise, a harmonic series in noise, a pre-emphasized speech vowel, and a speech vowel in broadband noise. Finally, in section 5, we discuss the implications of the auditory processing, especially its effects on the representation of acoustic power spectrum, and elaborate briefly on its relationship to several psychoacoustical measures and to commonly used representations and signal processing techniques in ASR systems.

2 Brief Review of the Mathematical Formulations of the Auditory Model

There have been numerous descriptions of the early auditory system, ranging from detailed biophysical models to schematic computational algorithms [8, 7, 12, 1]. Despite this diversity, all models are generally composed of three major stages: analysis, transduction, and reduction (fig. 1). In the following, we briefly review a *minimal* model that is both mathematically tractable and biophysically defensible, at least for the case of broadband signals (such as speech) and at moderate to high levels of intensity. This model was previously described in more detail in [38].

2.1 The analysis stage

Sound pressure waves impinging upon the ear drum cause vibrations that are transmitted to the fluids of the cochlea via the oscicles of the middle ear. These vibrations induce pressure differences across the basilar membrane. They in turn produce mechanical displacements in the form of traveling waves whose amplitudes peak at specific locations along the cochlea in an ordered manner depending on the frequency of the stimulus. Thus, for high frequencies the maximum response occurs near the base of the cochlea, while for lower frequencies it occurs near the apex. In this way, the spatial axis of the cochlea may be associated with a tonotopically ordered (i.e., a frequency) axis. One simple way to describe the response characteristics of the basilar membrane is to associate each point on it with a transfer function, i.e., to model the basilar membrane as a bank of filters. Suppose the sound signal is described by $x(t)$, then the basilar membrane response is given by

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

where $h(t,s)$ denotes the cochlear filter at a specific location s on the basilar membrane and $*_t$ denotes the convolution in the time domain.

In describing the basilar membrane responses in linear terms as above, we have chosen to ignore a host of additional simple and complex details that may be critical in some applications. For instance, we have ignored the spectral filtering induced by the outer ear which is important for auditory localization tasks. We have also ignored the relatively slow adaptive (AGC-like) action and lowpass filtering of the middle ear muscles and bones which are useful in protecting against very loud sounds.

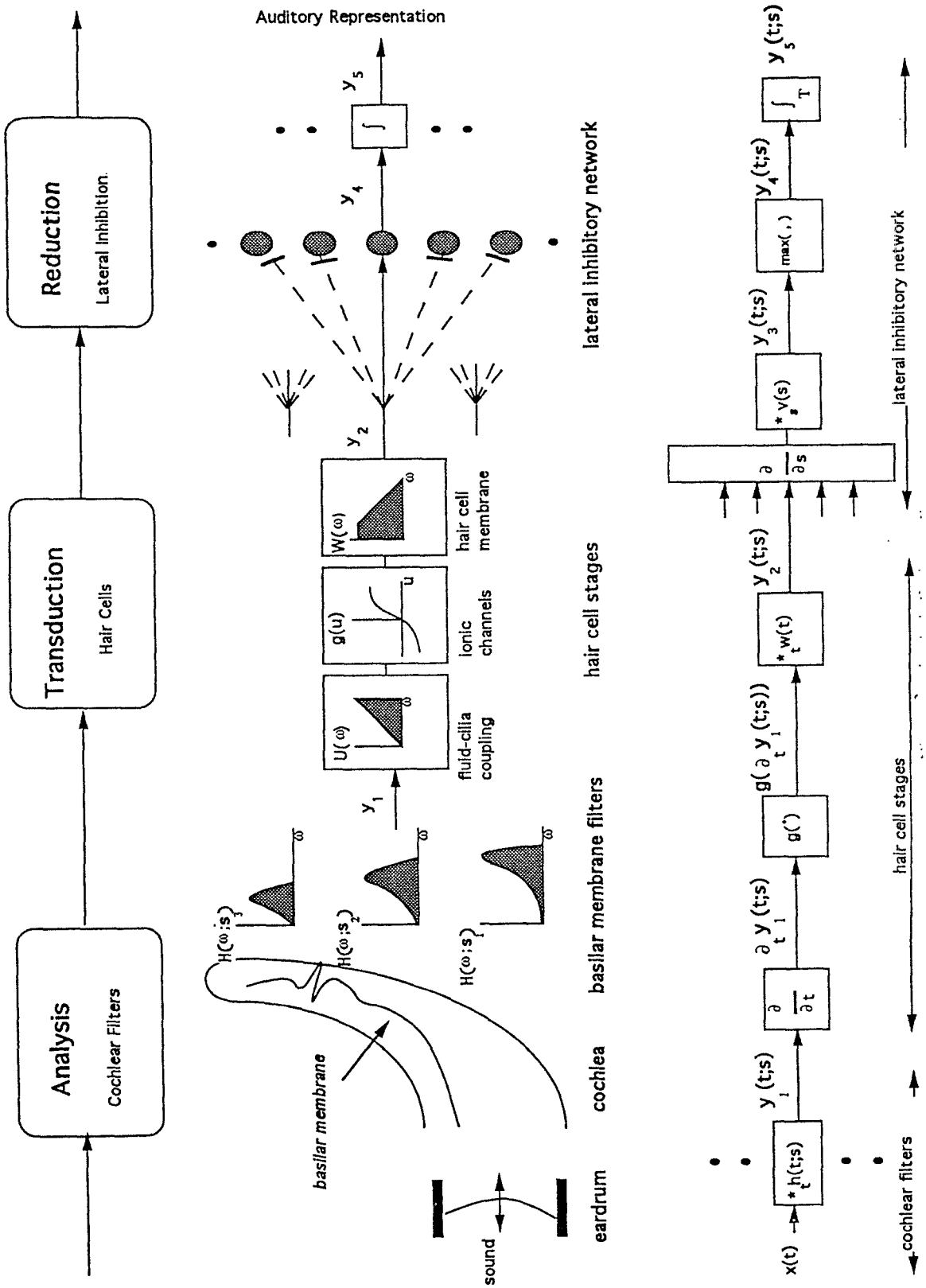


Figure 1: Schematic description of the auditory model (adapted from [38]). (a) Block diagram of the three basic stages in the early auditory system. (b) Quasi-anatomic sketches of the auditory stages. (c) Mathematical model of the different stages.

Finally, we have ignored all the nonlinearities and adaptive mechanisms in the basilar membrane that are attributable to the motion of the outer hair cells and the feedback communicated via the efferent system. These mechanisms are particularly important in enhancing the sensitivity of the basilar membrane at very low sound levels. All these simplifications, however, can be justified if we restrict the domain of applicability of this model to the processing of broadband signals (such as speech) at moderate to high levels of intensity. In this case, it has been repeatedly shown that the cochlear filters behave largely as linear and relatively broadly tuned filters [7].

For the analysis presented in this paper, the exact shape of the cochlear filters is unimportant. Rather, it is understood that the filters satisfy certain gross characteristic features typical of the shapes inferred from available physiological and psychoacoustical data. For instance, cochlear filters are generally relatively broadly tuned and significantly asymmetric in shape, with a steep roll-off on their high frequency sides. The center frequency of each filter is called the characteristic frequency (CF). For filters with CF's larger than about 800 Hz, the bandwidth is linearly proportional to the CF, i.e., the filters have constant Q's; for lower CF's, the filter's become gradually less tuned. The constant tuning over most of the cochlear length implies that the CF's of the filters are logarithmically mapped. As such, the cochlear filters can be related by a simple dilation and the basilar membrane response $y_1(t,s)$ in eq. 1 is a *wavelet transform* of the sound signal $x(t)$. For the analysis presented here, this property of the filters is not essential. Instead, filter shapes can be designed to satisfy arbitrary application-dependent criteria.

2.2 The transduction stage

The mechanical vibrations on the basilar membrane are tranduced into electrical activity along a dense, topographically ordered array of auditory nerve fibers. At each point on the basilar membrane, the motion of the membrane causes a local fluid flow which bends small filiments (*cilia*) that are attached to the so-called (inner) hair cells. The bending of the cilia controls the ionic currents through a nonlinear channel into the hair cells. This ionic flow, in turn, generates electrical potentials across the hair cell membranes and these potentials are finally conveyed by the auditory nerve fibers in the form of neural spikes to the cochlear nucleus.

The above biophysical process transforming the motion on the basilar membrane into the neural spikes in the auditory nerves can be modeled by the following three steps [30]. First, since the ionic current is *velocity* driven, a temporal derivative is employed to convert instantaneous membrane displacement into velocity (fig. 1). The nonlinear channel through the hair cell is then modeled by a sigmoid-like function, and the leakage of the cell membranes is accounted for by a low-pass filter. The hair cell responses can therefore be described by

$$y_2(t,s) = g(\partial_t y_1(t,s)) *_t w(t) \tag{2}$$

where $g(\cdot)$ denotes the compressive nonlinear function and $w(\cdot)$ is a temporal smoothing window. The lowpass filter $w(\cdot)$ in effect filters out all response frequencies $\geq$4 kHz. Consequently, for the very high frequency cells, the intracellular potential response exhibits only a steady dc increase in level with

little or no fluctuations. Much of the theory we discuss here is only relevant for the lower frequency ranges where much of speech signal energy is and phase-locked activity is still present. Thus, the $w(t)$ plays a minor role and will be usually ignored.

The hair cell nonlinearity $g(\cdot)$ is a compressive function described by a monotonically nondecreasing sigmoid function. Its response may be divided into a dynamic (linear) region, a cutoff region, and a saturated region. The so-called *dynamic range* usually refers to the area where the *gain* of the nonlinear function, $g'(\cdot)$, is large. The dynamic range of the hair cell nonlinearity is about 30-40 dB. Beyond this level, the hair cell is driven into saturation and the fine detail of the signal is heavily compressed and poorly represented.

The receptor potentials (or enhanced versions thereof) generated at the end of these stages are conveyed via the auditory-nerve fibers to the cochlear nucleus, the first station of the central auditory system. This is achieved through a series of transformations in which the receptor potentials are converted into stochastic trains of electrical impulses (firings) on the auditory nerve. Detailed biophysical models of these transformations can be found in [7, 12, 37]. More abstractly, the stochastic firings can be modeled as nonstationary point processes with instantaneous rates that approximately reflect the underlying receptor potentials [31]. Recipient neurons in the cochlear nucleus may then reconstruct estimates of the receptor potentials by effectively computing the ensemble averages of activity in locally adjacent fibers [29].

From an information processing point of view, these complex transformations merely convey the receptor potentials to the cochlear nucleus. Consequently, in a functional model, they can all be bypassed. Such a simplified view ignores two types of effects that have figured prominently in several classical models of auditory processing. The first is the adaptive mechanisms operative at the hair cell/auditory-nerve junctions which might be important in describing the responses to the onset of sound [21]. To a limited extent, the enhancing action of these mechanisms can be modeled by a linear stage and incorporated into the form of the filter in eq. 2 above.

The second simplification concerns the range of thresholds and spontaneous rates of firings observed in the responses of the auditory nerve. In most fibers ($> 85\%$), thresholds are lower than typical conversational sound levels, while their spontaneous rates are relatively high. These fibers exhibit a limited dynamic range between threshold and saturation (approximately 30 dB), and are consistently found to be almost totally saturated when driven by broadband sounds at moderate levels (60–70 dB SPL). The remaining fibers exhibit low spontaneous rates, a wider range of thresholds, and sloping saturations. Consequently, they are less likely to be saturated at normal sound levels. To model these two populations, it is possible to use two or more nonlinearities with appropriately weighted outputs.

2.3 The reduction stage

The auditory nerve transmits the hair cell responses $y_2(t, s)$ to the central auditory system where the information about various attributes of the sound, such as timber, pitch, and other spatial and temporal characteristics are extracted and processed. In this report, we focus on the extraction and representation of the acoustic spectrum, a fundamental cue for the perception of timber and the

recognition of speech signals.

There are many ways by which the spectral contents of the stimulus may be estimated from the patterns of auditory nerve responses. A detailed review can be found in [29]. The scheme discussed here is a particularly simple one which is found in all sensory systems and is implemented by a neural network commonly known as the *lateral inhibitory network* (LIN). In early vision, such network exists in the retina and functions like an edge detector/magnifier [19]. In audition, the LIN is presumed to exist in the cochlear nucleus which receives direct inputs from the auditory nerve and exhibits physiological and anatomical characteristics consistent with the structure and functions of LIN [29]. The simplest model of the LIN consists of a single layer of nonlinear neurons that are mutually inhibited either in a feed forward or feedback manner [27]. From a mathematical point of view, the LIN operations can be effectively divided into three steps as depicted in fig. 1. The first two steps are a derivative with respect to the tonotopic axis that mimics the lateral interaction among LIN neurons, followed by a half-wave rectifier modeling the nonlinearity of the LIN neurons. More realistically, the derivative is leaky in nature, i.e., is accompanied by a local smoothing due to the finite spatial extent of the lateral interactions [29]. Generally speaking, this operation essentially detects spatial discontinuities in the $y_2(t,s)$ responses along the cochlear axis s. The last step of LIN models primarily the fact that the central auditory neurons, unlike the auditory nerve fibers, are unable to follow rapid temporal modulations higher than, for example, a few hundred hertz. Rather, they signal a temporally integrated version of the outputs. The LIN operations then can be described by

$$\begin{aligned} y_3(t,s) &= \partial_s y_2(t,s) *_s \nu(s) \\ &= [g'(\partial_t y_1(t,s))\partial_s\partial_t y_1(t,s) *_t w(t)] *_s \nu(s) \qquad (3) \\ y_4(t,s) &= \max(y_3(t,s), 0) \qquad (4) \\ y_5(t,s) &= y_4(t,s) *_t \Pi(t) \qquad (5) \end{aligned}$$

where $\Pi(\cdot)$ is a temporal integration window and $\nu(\cdot)$ is a spatial smoothing function. At the final output of the LIN we thus obtain a representation of the sound that, as we shall elaborate, approximately reflect a short-time spectral profile of the signal. This pattern will be referred to as the *auditory spectrum* of the signal.

The characteristics of this auditory spectrum and its application for speech signal representation have already been empirically studied. For example, it was shown in [38, 28, 34] that the formant structure of various phonemes are well preserved and that simple speech recognition systems can be designed to distinguish most phonemes based on these features. In another series of experiments [6], it was demonstrated that the auditory representation has a significant advantage over conventional representations in noise robustness when employed as a front-end for ASR systems. The issue of noise robustness was also studied in [38], where it was observed that the speech signal reconstructed from the auditory spectrum exhibit noise suppression. The exact reasons for this phenomenon were not determined then because of the difficult nature of the deterministic analysis of the model. In the next section, we present an alternative stochastic analysis of the model to overcome these difficulties.

3 Stochastic Analysis of the Early Auditory Model

In this section, a stochastic interpretation of the auditory processing described above is developed. The fundamental goal of this analysis is to describe accurately the nature of the spectral analysis performed by the auditory system, and to explain the enhancements and noise suppression effects previously observed in the auditory spectrum[38]. For simplicity, we shall ignore the temporal and spatial smoothings $w(t)$ and $\nu(s)$ as in [38].

Suppose the sound signal $x(t)$ can be modeled by a zero mean random process. Accordingly, so are the signals $y_1(t,s)\cdots y_5(t,s)$ random processes (not necessarily zero mean). Since the auditory spectrum $y_5(t,s)$ is a filtered process, the variance of the process $y_5(t,s)$ is limited by the smoothness, or the bandwidth of the lowpass filter $\Pi(t)$ [10]. If this bandwidth is narrow enough, the variance of $y_5(t,s)$ would be very small for any instant t. According to Chebyshev inequality [10], this implies that $y_5(t,s)$ can be approximated by its expectancy

$$\begin{aligned} E[y_5(t,s)] &= E[y_4(t,s) *_t \Pi(t)] \\ &= E[y_4(t,s)] *_t \Pi(t) \end{aligned} \tag{6}$$

where

$$\begin{aligned} E[y_4(t,s)] &= E[\max(y_3(t,s),0)] \\ &= E[\max(g'(\partial_t y_1(t,s))\partial_s\partial_t y_1(t,s),0)] \end{aligned} \tag{7}$$

Given $\Pi(\cdot)$, eq. 6 suggests that the auditory representation can be characterized by examining $E[y_4(t,s)]$. If $y_4(t,s)$ satisfies an ergodic theorem within a duration comparable to the time constant of $\Pi(t)$, then $y_5(t,s)$ will indeed converge to a time-invariant constant although the signal and $y_4(t,s)$ may well be non-stationary [10]. A stochastic process whose sample mean converges (along with the number of samples) is said to satisfy an ergodic theorem. Such condition is not contingent upon whether the process is stationary or not [10]. In this report, we refer to the ergodic mean of $y_4(t,s)$, towards which $y_5(t,s)$ converges, as the auditory spectrum. With such a definition, the auditory spectrum is a first order statistic of $y_5(t,s)$. In contrast, conventional definition of power spectrum requires the *second* order statistic (autocorrelation or autocovariance function) to be a Teoplitz function, i.e., $R(t,\tau) = R(t-\tau)$, which in turn has to base on a (weakly) stationary assumption on the stochastic process.

For notational simplicity, let $U = \partial_t y_1(t,s)$ and $V = \partial_s\partial_t y_1(t,s)$. Assume the hair cell nonlinearity is monotonically non-decreasing, i.e., $g'(\cdot) \geq 0$. Eq. 7 can then be rewritten as

$$\begin{aligned} E[y_4(t,s)] &= \int\int \max(g'(u)v,0) f_{uv}(u,v)dudv \\ &= \int\int g'(u)\max(v,0) f_{uv}(u,v)dudv \\ &= \int g'(u)E[\max(V,0)|U=u]f_u(u)du \\ &= E[g'(U)E[\max(V,0)|U]] \end{aligned} \tag{8}$$

where $f_u(\cdot)$ and $f_{uv}(\cdot,\cdot)$ denote the probability density function (pdf) of U and joint density function of U and V, respectively.

3.1 The "linear" auditory spectrum

In the case that $g'(x) = 1$, i.e., there is no hair cell nonlinearity, eq. 8 is reduced to

$$E[y_4(t,s)] = E[E[\max(V,0)|U]] = E[\max(V,0)]$$

implying $E[y_4(t,s)]$ is simply the dc component of the half-wave rectified V. Recall that

$$\begin{aligned} V &= \partial_s \partial_t y_1(t,s) \\ &= (\partial_t x(t)) *_t \partial_s h(t,s) \end{aligned} \tag{9}$$

i.e., V is the components of the preemphasized signal $\partial_t x(t)$ passing through the *differential filter* $\partial_s h(t,s)$. Since the signal $x(t)$ is zero mean, so is V. Hence the quantity $E[\max(V,0)]$ is proportional (though not necessarily linearly) to the standard deviation σ, or the instantaneous *energy* of V.[2] Thus in this linear case, the signal is analyzed into various frequency bands defined by the differential cochlear filters, and the scheme is no different from a filter-bank based frequency analysis framework, using the differential filters (as opposed to the cochlear filters) as analyzing filters. While the cochlear filters are broad and highly asymmetric, the differential filters are narrowly tuned and centered around the same frequencies (Fig. 2). As such, $E[y_4(t,s)] = E[\max(V,0)]$ is simply the *spectral energy profile* of the sound signal $x(t)$ across the channels indexed by s.[3]

3.2 The effects of the hair cell nonlinearity

When the compressive hair cell nonlinearity is taken into consideration, it transforms the estimate of the spectral energy profile into a *conditional* measure. Thus, while still being estimated through the dc component of the half-wave rectified waveform, the spectral energy becomes modulated by the instantaneous gain of the hair cell nonlinearity, as described by eq. 8. In the following, we elaborate on the implications of this nonlinear effect.

First consider the nonlinear gain $g'(U)$. As mentioned earlier in section 2, the hair cell nonlinearity has a finite dynamic range so that it is occasionally driven into saturation. In practice, this occurs when the sound intensity is at or above a moderate level. Since the nonlinearity is characterized by a negligible gain outside the dynamic range, this implies that the estimation of the energy resolved by the differential filters is not made equally at all instants as in the linear case. Instead, the estimate now takes place only when $U = \partial_t y_1(t,s)$ lies within the dynamic range of the nonlinearity. Therefore, the estimate is also affected by the "rate" of U being within the dynamic range. When U has a large

[2]The physical term "energy" usually refers to the variance σ^2 of a random process rather than the standard deviation σ. However, due to the one-to-one correspondence between the two quantities, their minor difference is not distinguished throughout this paper.

[3]Note that this narrowband filtering leading to the LIN auditory spectrum is fundamentally identical to the operations invoked to obtain the so-called ALSR spectrum in[25].

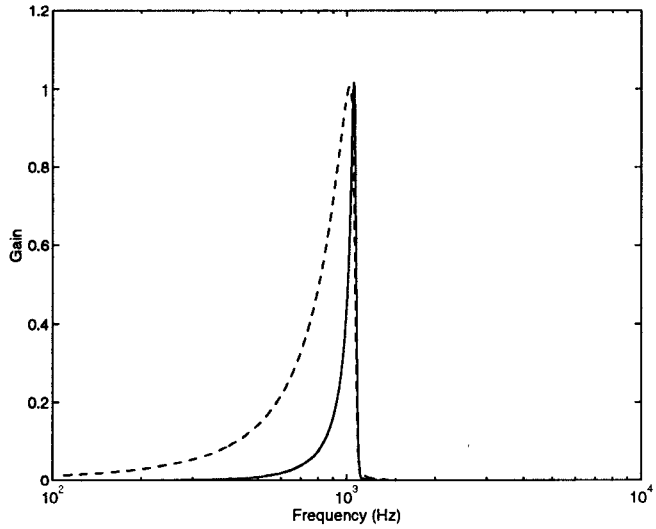


Figure 2: **The cochlear filter $H(\omega, s)$ centered around 1 kHz (dashed) and its corresponding differential filter, $\partial_s H(\omega, s)$. The ordinate is labeled by arbitrary linear units.**

variance, it is less probable to restrict its value within a certain range. This "rate" decreases as the energy of U increases, and vice versa. Therefore, conceptually, $E[y_4(t, s)]$ contains a term that is *inversely proportional* to the energy of U.

The second term in eq. 8, the conditional expectation $E[\max(V, 0)|U]$, suggests an additional property of the nonlinearity. As is in the linear case, the dc component of the half-wave rectified waveform of V is evaluated. However, the evaluation now only takes place when U is in the dynamic range of the nonlinearity. Effectively, the compressive nonlinearity introduces a sampling process imposed on V, the output waveform from the differential filter. The energy profile estimation is then carried out based on these samples instead of the whole waveform. In the case that U and V are independent, $E[\max(V, 0)|U] = E[\max(V, 0)]$, i.e., V is unbiasedly sampled and therefore the estimation of its energy is left unaffected. In general, however, the dependency between U and V affects the energy estimation, and hence the auditory spectrum (see later for an example). Since the two processes U and V are generated by passing the same source, $\partial_t x(t)$, through the cochlear and differential filters, respectively, the dependency between U and V is basically dependent on the design of the filters.

In summary, $E[y_4(t, s)]$ is a quantity that is proportional to the energy of V, and inversely proportional to the energy of U and their dependency. As established earlier, the auditory spectrum is an ergodic average of $y_4(t, s)$. Thus, by the definitions of U and V, this suggests that $y_5(t, s)$ is an averaged *ratio* of the signal energy passing through the differential and cochlear filters. That is, the auditory spectrum is a *self-normalized* spectral profile. Before elaborating on the significance of such self-normalization, we consider the following special cases to exemplify our arguments.

3.3 The special case of a high gain nonlinearity

The arguments above were based on the premise that the compressive hair cell nonlinearity has a finite dynamic range. However, the above analytical expressions can be dramatically simplified by pushing the nonlinearity to the extreme where it becomes a Heaviside (step) function. The dynamic range consequently is reduced to a singleton at the origin and the gain is infinite, described by $g'(x) = \delta(x)$, the Dirac delta function. In this case, eq. 8 can be rewritten as

$$E[y_4(t,s)] = E[\max(V,0)|U=0]f_u(0). \tag{10}$$

The first term in the above equation illustrates that V is sampled at the *zero crossings* of U. The second term $f_u(0)$, the probability density of $U = 0$, is generally inversely proportional to the standard deviation of U since U is assumed to have zero mean.

3.4 The high gain auditory spectrum for a Gaussian source

In order to obtain an explicit expression for the joint distribution of U and V, let us assume the source $x(t)$ is a zero mean Gaussian process. Since U and V are obtained by linear filtering, U, V are also zero mean Gaussian. Let r, σ_u, σ_v denote the correlation coefficient and the standard deviations of U and V, respectively. Note that all these quantities, like U and V themselves, are indexed by (t,s), which we drop throughout for notational simplicity. As shown in [20], a special property of the Gaussian distribution is that the conditional distribution $f_{v|u}(v|u)$ is also Gaussian with mean u and variance $\sigma^2 = \sigma_v^2(1-r^2)$. Hence,

$$\begin{aligned} E[\max(V,0)|U=u] &= \int_0^\infty v f_{v|u}(v|u)dv \\ &= \frac{\sigma}{\sqrt{2\pi}} e^{-u^2/2\sigma^2} + u[1-\Phi(-u/\sigma)] \end{aligned} \tag{11}$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of a Gaussian random variable with zero mean and unit variance. By eqs. 10 and 11, $E[y_4(t,s)]$ for the high gain case becomes

$$E[y_4(t,s)] = \frac{1}{2\pi}\frac{\sigma_v}{\sigma_u}\sqrt{1-r^2}. \tag{12}$$

This estimate can be decomposed into three terms corresponding to the nonlinear effects described earlier: $\sigma_v/\sqrt{2\pi}$ is the energy profile of the linear case (i.e., $E[\max(V,0)]$), $1/\sqrt{2\pi}\sigma_u$ is the rate of U entering the dynamic range of the nonlinearity (i.e., $f_u(0)$), and $\sqrt{1-r^2}$ accounts for the decrease in energy estimation due to the conditioning. Note that in this expression, σ_v is normalized by σ_u, which demonstrates the self-normalizing nature of the auditory spectrum.

3.5 The high gain case for a weakly stationary source

Now we consider the additional assumption that the source is weakly stationary and has a spectral density function (s.d.f.) $S_x(\omega)$. For a Gaussian source, weak stationarity also implies strict stationarity.

The statistics mentioned above all become time-invariant constants and can be obtained from the following:

$$\sigma_u^2 = \frac{1}{2\pi}\int |\omega H(\omega,s)|^2 S_x(\omega)d\omega \quad (13)$$

$$\sigma_v^2 = \frac{1}{2\pi}\int |\omega \partial_s H(\omega,s)|^2 S_x(\omega)d\omega \quad (14)$$

$$r = \sigma_{uv}/\sigma_u\sigma_v \quad (15)$$

$$\sigma_{uv} = \frac{1}{2\pi}\int \omega^2 H^*(\omega,s)\partial_s H(\omega,s) S_x(\omega)d\omega \quad (16)$$

where H^* denotes the complex conjugate of H. Let the cochlear band $\Omega_c(s)$ and the differential band $\Omega_d(s)$ be the passbands of the (preemphasized) cochlear filter $|\omega H(\omega,s)|$ and differential filter $|\omega\partial_s H(\omega,s)|$ respectively. σ_u^2 and σ_v^2 can be restated as the spectral energy within the cochlear band and the differential band, respectively. If the differential band is very narrow, for example, $\omega\partial_s H(\omega,s) = \delta(\omega \pm \omega_s)$, $\sigma_v^2 = S_x(\pm\omega_s)$. That is, σ_v^2 is indeed the power of the signal at frequency ω_s.

The dependency of U and V is described by r in eq. 15. Since $S_x(\omega)$ is real and positive, eq. 15 can be rewritten as

$$r = \frac{< \omega\partial_s H(\omega,s)\sqrt{S_x(\omega)}, \omega H(\omega,s)\sqrt{S_x(\omega)} >}{||\omega H(\omega,s)\sqrt{S_x(\omega)}|| \cdot ||\omega\partial_s H(\omega,s)\sqrt{S_x(\omega)}||} \quad (17)$$

where $< \cdot,\cdot >$ denotes the inner product in the ω domain. The parameter, r, is therefore a quantity measuring the similarity of the two functions, $\omega\partial_s H(\omega,s)\sqrt{S_x(\omega)}$ and $\omega H(\omega,s)\sqrt{S_x(\omega)}$. According to Cauchy-Schwartz inequality, $r \leq 1$ and the equality holds if and only if the two functions are equal. If the differential and cochlear filters are designed to be dissimilar, r becomes small for broadband $S_x(\omega)$ and $\sqrt{1-r^2} \approx 1$. For the filters shown in fig. 2 (which are used in the following section), $\sqrt{1-r^2}$ usually ranges between 0.76 to 0.96, and is equal to 0.84 for white noise.

3.6 Summary

To sum up, $E[y_4(t,s)]$ is a normalized power spectrum of the sound stimulus (eq.12), and so is the auditory spectrum (eq. 6). The normalization here is neither uniform nor predetermined. Rather, it is driven by the energy distribution of the signal. For each channel s, the output reflects approximately the *ratio* of the energy of its differential filter to that of its cochlear filter. The spectral components of the sound signal $S_x(\omega)$ therefore receive unproportional scaling: a spectral peak resolved by the differential filter receives a relatively small normalization factor since the cochlear filter in its vicinity integrates energy from the valleys surrounding the peak. The opposite is the case for a spectral valley. Effectively, this difference in the normalization further enlarges the spectral peak to valley ratio, a phenomenon referred to here as *spectral enhancement* or *noise-suppression* depending on what is being suppressed in the spectral valley. Generally speaking, the broader the cochlear band is relative to the differential band, the more enhancement effects can be expected.

In the following section, we shall illustrate these nonlinear effects with a series of examples, and then in section 5 further elaborate their analytical significances and psychoacoustical implications.

4 Examples of Spectral Enhancement and Noise Suppression

We illustrate in this section the properties of the auditory spectrum in the high gain case for four signals: tone in noise, a harmonic series in noise, a natural vowel, and a natural vowel in noise. Our aim here is to highlight some of the characteristic features of the auditory spectrum that result from self-normalization property discussed earlier. Throughout this section, we assume that $y_5(\cdot, s)$ is well behaved so that the auditory spectrum $E[y_5(t,s)]$ can be approximated by an ensemble of $y_5(t,s)$. We start with a description of the implementational details of the model used in this section.

4.1 Implementation of the auditory model

Although the theory developed in the previous section does not depend upon specific $H(\omega, s)$, the design of the cochlear filter plays an important role in the application of the auditory model. This issue has been discussed in some detail earlier in [36] (see also section 2). Motivated partly by biophysical and psychophysical considerations, the filters chosen here are related by a dilation, i.e.,

$$H(\omega, s) = H_m(a^s \omega) \tag{18}$$

for some constant a and seed filter $H_m(\omega)$. Since

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

the differential filters are also related by a dilation; So are $\Omega_d(s)$ and $\Omega_c(s)$. If the differential band $\Omega_d(s)$ is reasonably narrow, the spatial axis s can be thought of as being labeled by logarithmic frequencies, which is roughly equivalent to the tonotopic structure of the auditory system. One convenient property of the filter dilations is that, on a logarithmic frequency axis, the filters have the same shape and width as the seed filter. They are merely translations of each other. The filter design problems are thus simplified to a single seed filter rather than the whole filter bank. Fig. 2 shows an example of a cochlear filter and its corresponding differential filter used in the computations discussed below. They are obtained following the criteria and design principles outlined in [36]. Generally speaking, the cochlear filter is designed to be relatively broad and asymmetric so that its differential version is reasonably narrow. In this implementation, the 3-dB bandwidth of the differential filter is roughly 5% of the peak frequency, and $\Omega_d(s) \subset \Omega_c(s)$. The seed filter is then dilated 2 octaves down and 2.8 octaves up to cover the frequency range from 250 Hz to 7 kHz. The spatial channel is discretized at a resolution of 20 channels/octave, causing the differential filters to overlap at their 3-dB points. Finally, the smoothing window $\Pi(t)$ is implemented by a leaky integrator with a time constant $\tau = 20$ msec, i.e., $\Pi(t)$ is a single-pole low pass filter with the corner frequency at 50 rad/sec. With these, if the source is stationary,

$$E[y_5(t,s)] = E[y_4(t,s)] \int_0^\infty e^{-t/\tau} dt = \tau E[y_4(t,s)].$$

Furthermore, if the source is Gaussian, by eq. 10, we have

$$E[y_5(\cdot, s)] = \frac{\tau}{2\pi} \frac{\sigma_v}{\sigma_u} \sqrt{1 - r^2} \tag{20}$$

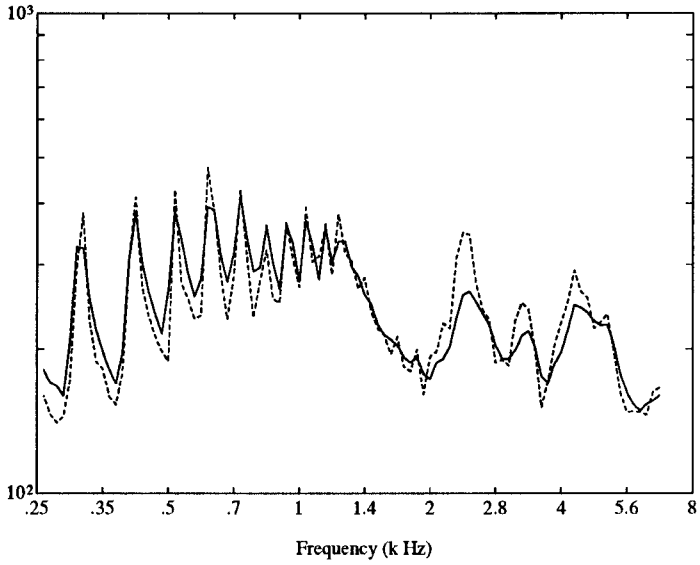


Figure 3: The predicted (solid) and computed (dotted) auditory spectrum for a naturally spoken vowel /aa/. The computed spectrum is scaled before being superimposed upon the predicted pattern. The predicted pattern is obtained by using eq.20. The channels are labeled with the CF of the corresponding differential filters. Roughly, the relationship between the channel number and the CF is described by $s = 20\log_2(f/250)$. The logarithmic ordinate is labeled by arbitrary units.

where all the parameters can be obtained from eqs. 13–15. These equations, which give a stochastic description of the auditory spectrum, are not difficult to evaluate in computer simulations. However, a simpler and a more direct approach from a biological and hardware implementation point of view is to approximate the spatial derivative ∂_s (fig. 1) by a difference and simply subtract the heavily saturated outputs of adjacent cochlear filters, and then half-wave rectify the result and integrate it over $\Pi(t)$. Unless stated otherwise, the auditory spectra discussed below are obtained in this manner. Fig. 3 demonstrates the closeness of the auditory spectra for a vowel /aa/ computed by the two methods.

4.2 Noise suppression in the auditory spectrum

Consider first the auditory spectrum for white Gaussian noise. Regardless of the level of the noise, the filters in eqs. 13–15 can be replaced by those in eqs. 18–19 to obtain

$$r_n^2 = \frac{(\int \omega^3 H'_m(\omega) H^*_m(\omega) d\omega)^2}{\int |\omega H_m(\omega)|^2 d\omega \int |\omega^2 H'_m(\omega)|^2 d\omega} \tag{21}$$

$$\frac{\sigma^2_{v|n}}{\sigma^2_{u|n}} = \frac{\int |\omega H'_m(\omega)|^2 d\omega}{\int |\omega H_m(\omega)|^2 d\omega}. \tag{22}$$

The fact that both of these two quantities are independent of s suggests that the auditory spectrum $E[y_5(\cdot, s)]$ (eq. 20) of a white noise is constant (over s). According to the above equations, this constant

$$\kappa = \frac{\tau}{2\pi} \frac{\sigma_{v|n}}{\sigma_{u|n}} \sqrt{1 - r_n^2} \tag{23}$$

does not depend upon the level of the noise and is determined solely by the shape of the seed filter, a direct reflection of the normalized nature of the auditory spectrum. Since κ does not vary with the noise power, it will be refered to in the following as the baseline of the auditory spectrum, and the noise will be assumed to have unit power.

When a tone of frequency ω_0 (in the sense defined in section 1) is added to the noise, the s.d.f. of the signal $x(t)$ is

$$S_x(\omega) = \sigma^2 \pi \delta(\omega \pm \omega_0) + 1$$

where σ^2 here is the signal to noise ratio (SNR). For s such that $\omega_0 \notin \Omega_c(s)$, $E[y_5(\cdot, s)]$ is still equal to the baseline. However, for channels whose cochlear band covers the tone frequency but the differential band does not, i.e., $\omega_0 \in \Omega_c(s) \cap \Omega_d(s)^c$, substituting the s.d.f. into eqs. 13–15 gives

$$\begin{aligned} \sigma_v^2 &= \sigma_{v|n}^2 \\ \sigma_u^2 &= \sigma_{u|n}^2 + \sigma^2 |\omega_0 H(\omega_0, s)|^2 = \sigma_{u|n}^2 / \beta^2 \\ r &= r_n \beta \end{aligned}$$

where

$$\begin{aligned} \beta &= \sigma_{u|n} / \sigma_u \\ &= \left(1 + \left| \frac{\sigma \omega_0 H(\omega_0, s)}{\sigma_{u|n}} \right|^2 \right)^{-1/2} . \end{aligned} \tag{24}$$

Substituting the above into eq. 20, the auditory spectrum of the tone in white noise becomes

$$E[y_5(\cdot, s)] = \kappa \alpha \tag{25}$$

where

$$\alpha = \frac{\beta \sqrt{1 - r_n^2 \beta^2}}{\sqrt{1 - r_n^2}} \tag{26}$$

Note that eq. 24 indicates that $\beta \leq 1$, which implies that $\alpha \leq 1$ and the equality holds only if $\beta = 1$ (SNR or $\sigma = 0$). Also note that α is roughly proportional to β when β is small. Eq. 25 thus suggests that in the auditory spectrum, the noise level at the channels close to the tone frequency ($\omega_0 \in \Omega_c(s)$) is *suppressed* by a factor α. As the SNR increases, so does the suppression.

This suppression effect is demonstrated in fig. 4 which compares the computed (dashed) and predicted (solid) auditory spectra of a 1 kHz single tone corrupted by a white Gaussian noise at various SNR's. Since the suppression occurs at s where $\omega_0 \in \Omega_c(s) \cap \Omega_d(s)^c$, the suppression effect in this implementation takes place mostly at channels higher in frequency than ω_0 due to the asymmetric shape of the cochlear filters (fig. 2). All these features closely resemble findings from psychoacoustical experiments employing tone in noise stimuli [22, 35].

4.3 Enhancement of spectral peaks

To demonstrate the spectral enhancement in the peak to valley ratio, we compare the auditory spectrum to the *linear* power spectrum of a harmonic series in white Gaussian noise. The linear power

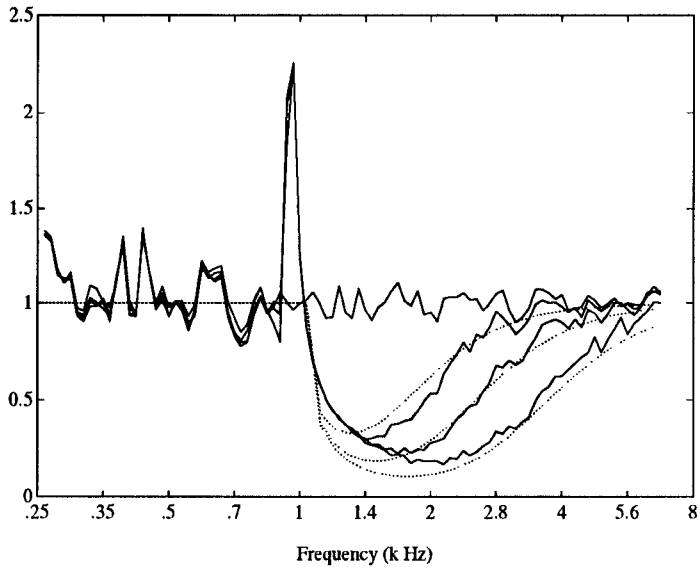


Figure 4: Example of noise suppression in the auditory spectrum. The solid lines are the computed auditory spectrum for the single tone in noise stimulus. The SNR's (see text for definition) are $-\infty$, -6, 0, and 6 dB from top to bottom. The dotted lines are the predicted amount of suppression obtained from eq. 26. The computed auditory spectra are scaled so that the baseline κ is 1.

spectrum is obtained by merely removing the hair cell nonlinearity (i.e., by letting $g(x) = x$). Note that the last two stages in the LIN, a halfwave rectifier followed by a leaky integration, is an envelope detector. Such a scheme can therefore be viewed as estimating the energy resolved by the differential filters alone without self-normalization. The harmonic series has a fundamental frequency 150 Hz, and the phase for each component is randomly chosen. It can be generalized from the argument for a single tone (section 1) that such a harmonic series, with the phase of each component independently randomized, is a weakly stationary signal. In this demonstration, each component in this series is set to be 6 dB stronger than the white Gaussian noise (within the band of analysis). The two computed spectra are compared in fig. 5. As predicted, the valleys in the auditory spectrum are deeper than in the linear power spectrum. Note that in both cases, the low frequency harmonics are well resolved compared to the higher ones since they are more separated on a logarithmic frequency axis. The enhancement effect within this region is similar to the single tone in noise case and can be theoretically predicted in the same fashion.

4.4 Robustness against scaling effects

Another property of a self-normalized spectrum is its relative stability with respect to an overall scaling. This follows directly from the fact that the auditory spectrum is an energy ratio, and hence the scaling factor will appear in both the numerator and denominator in, for example, eq. 20. More generally, the auditory spectrum is relatively insensitive to broadband changes in the spectral shape of the signal as long as the responses from the differential and cochlear filters are affected similarly.

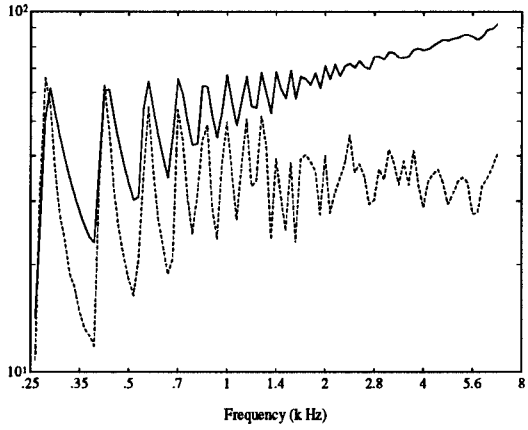


Figure 5: **Spectral peak enhancement in the auditory spectrum. The solid (dotted) trace is the linear power (auditory) spectrum for the stimulus consisting of a harmonic series and white noise. The computed auditory spectrum is scaled before being superimposed on the power spectrum. Since the two patterns are shown in log scale, this adjustment does not affect the reading of the peak to valley ratio. As predicted, the peak to valley ratio is larger in the auditory spectrum due to the spectral enhancement effect discussed in text. The logarithmic ordinate is labeled by arbitrary units.**

To quantify these effects, consider an overall distortion of the acoustic spectrum ($A(\omega)$), or equivalently in our implementation, a distortion in the overall gain of the auditory filters ($A(s)$) as follows:

$$\tilde{H}(\omega, s) = A(s)H(\omega, s)$$

The differential filter is therefore

$$\begin{aligned} \partial_s \tilde{H}(\omega, s) &= A(s)\partial_s H(\omega, s) + A'(s)H(\omega, s) \\ &\approx A(s)\partial_s H(\omega, s) \end{aligned}$$

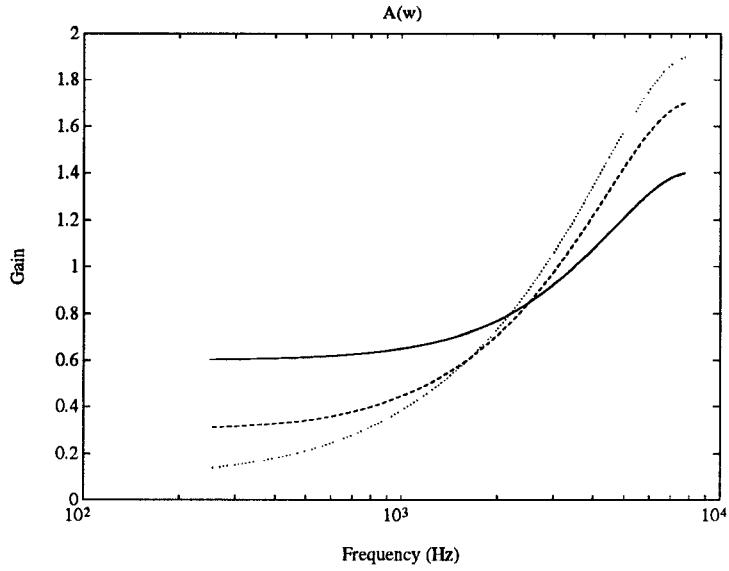
when $A'(s)$ is small relative to $A(s)$, i.e., $A(s)$ is a broadband adjustment. In this case, the gain term $A(s)$ scales the cochlear and differential filter responses approximately equally, and hence the net auditory spectrum remains unchanged.

Depending on the nature of the signal and the application, $A(s)$ may represent intentional adjustments (e.g., pre-emphasis in speech processing) or an undesirable unpredictable distortion (e.g., spectral tilt reflecting inter-speaker variability, or a distortion in a signal caused by a noisy transmission channel). In all cases, it is anticipated that the auditory spectrum should exhibit more stability against the distortion so long as it is broadband in the sense defined above.

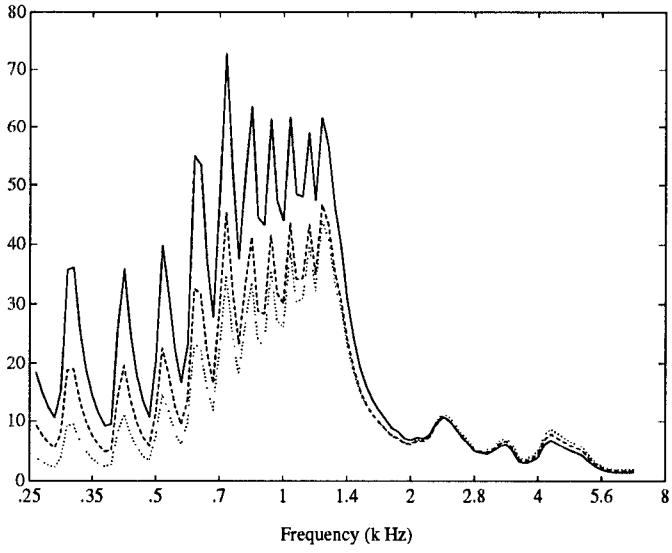
As an example, consider the effects of applying preemphasis. This is normally achieved by prefiltering the signal with a transfer function equivalent to $1 - cz^{-1}$, where c is the preemphasis factor. The effective distortion is

$$A(\omega) = \sqrt{1 + c^2 - 2c\cos\omega}$$

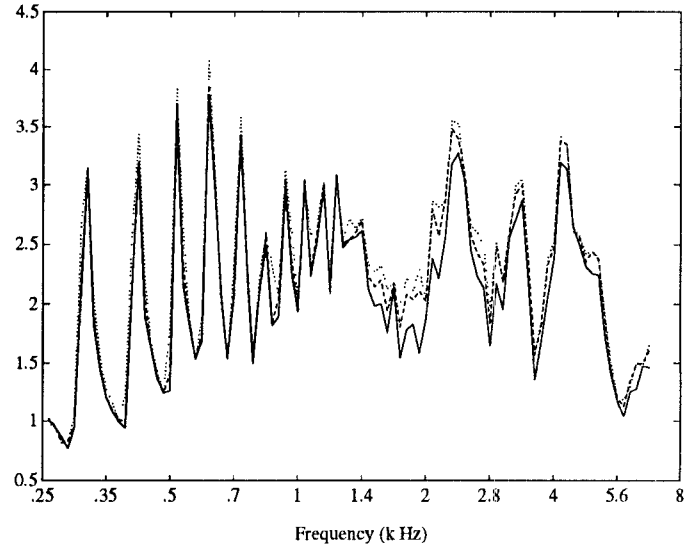
where ω denotes the normalized frequency (with respect to the sampling rate) and is logarithmically mapped by s in our model.



(a)



(b)



(c)

Figure 6: Stability of the auditory spectrum against broadband spectral distortion. In (a), the spectral distortion $A(\omega)$ is shown for $c = 0.4$ (solid), 0.7 (dashed) and 0.9 (dotted). (b) and (c) illustrate the corresponding distorted linear power spectrum and auditory spectrum of the vowel /aa/, respectively. Most distortion in the auditory spectrum is located at the high frequency channels where $A'(s)$ is not small. See text for detailed discussion. The ordinates are labeled by arbitrary units.

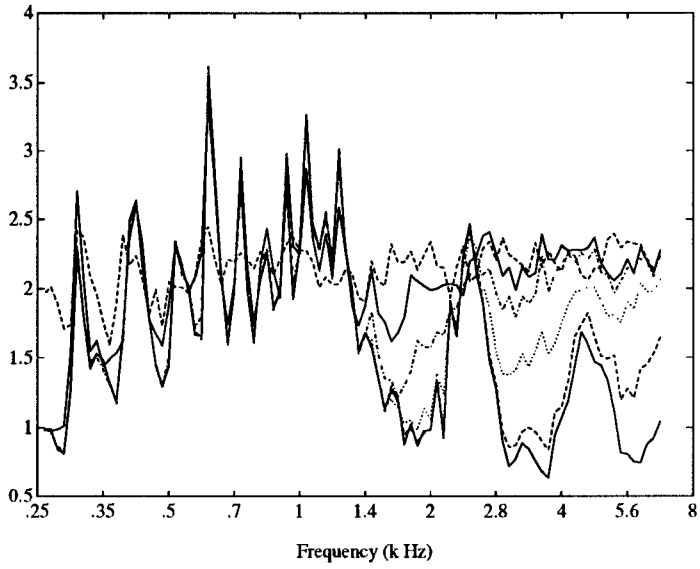


Figure 7: The auditory spectra for vowel /aa/ corrupted by white noise at various SNR's (see text for definition). The SNR's (from top to bottom) are -24, -12, -6, 0, 6 and ∞ dB. The ordinate is labeled by arbitrary units.

Fig. 6(a) shows the distortion weighting $A(\omega)$ for various preemphasis factors. The preemphasized linear power spectrum of a naturally spoken vowel /aa/ are shown in fig. 6(b) . For comparison, the preemphasized auditory spectra are shown in fig. 6(c). As can be seen in the previous example, the harmonics of this vowel are clearly resolved in the low frequency channels, while for the high frequency channels, only the formant structures are visible. Despite the large modification on the input spectrum (and the linear spectrum) shapes, the auditory spectrum remains relatively stable, with the pre-emphasis effects mostly concentrated in the high frequency channels. These occur because of relative steepness of the $A(s)$ function there , i.e., the distortion is not as broadband, and $A'(s)$ is not small.

4.5 Noise robustness of a natural vowel stimulus

One way to demonstrate the noise robustness of the auditory spectrum compared to the linear spectrum is to examine how fast it deteriorates with decreasing SNR. The SNR here refers to the ratio of the total energy of the clean speech to the energy of the additive noise. Fig. 7 shows the auditory spectrum of the vowel /aa/ corrupted by various levels of white background noise. Note that the auditory spectrum approaches the baseline κ as the noise gradually overwhelms the signal. Consequently, it can be viewed that the acoustic features in the auditory spectrum of the signal are represented by $F(s) = E[y_5(t,s)] - \kappa$ and $||F|| \longrightarrow 0$ as noise level increases. The spectral features conveyed in $F(s)$ will be further discussed in the next section. To quantify the effects of noise, we define a distortion

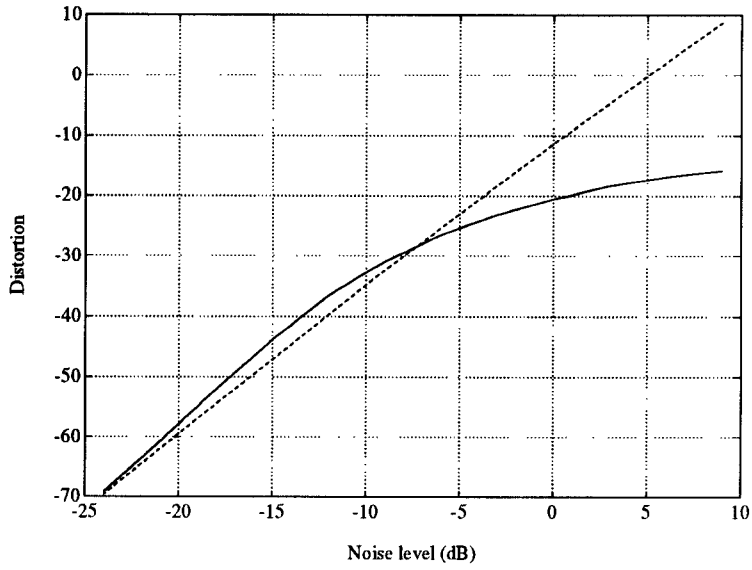


Figure 8: The distortion incurred by noise for both the auditory spectrum (solid) and the inear power spectrum (dotted trace) of the vowel /aa/. Note that the absicca is labeled according to the relative noise to signal level. At low noise level, the two representations have comparable performance. However, the distortion in the auditory spectrum does not increase as fast as the linear power spectrum.

measure based on the Euclidean distance as

$$D = 20 \log ||F(s, SNR) - F(s, \infty)|| \tag{27}$$

where $F(s, x)$ is the pattern when SNR $= x$ and $F(s, \infty)$ is the clean speech pattern. Fig. 8 shows a comparison in the amount of distortion between the auditory and the linear power spectra. As in this example, the two representations have the comparable performance for low noise level. At high noise level (SNR < 7 dB), however, the linear power spectrum degrades at a faster pace than the auditory spectrum.

5 Relating the auditory representation to the perception of the acoustic spectrum

In the previous section, we illustrated with several isolated examples how the local self-normalization of the auditory spectrum manifests itself as feature enhancement and noise robustness. In this section, we attempt to integrate these phenomena within a more specific extension of the theoretical framework developed earlier in section 3. This will allow us to demonstrate how the transformation from an acoustic (power) spectrum into an auditory pattern in the early auditory system may possibly be linked to several important psychoacoustical phenomena.

In the following, the framework for the analysis is first developed (section 5.1) and then several of its perceptual implications are discussed. These include: invariance to broad scale adjustments

(5.2) , the representation of spectral slopes (5.3), the transfer characteristics of rippled spectra (5.3), linearity and superposition in the auditory representation (5.4), and the dominance principle and noise robustness (5.5).

5.1 Transforming the acoustic power spectrum into an auditory pattern

Consider a source power spectrum $S_x(\omega)$ applied to the auditory model. In general, this pattern can be linearly decomposed into

$$S_x(z) = \sum a_n z^n = a_0 + a_1 z + a_2 z^2 + \cdots \tag{28}$$

where $\{1, z, z^2, \cdots\}$ is a collection of orthogonal bases and $\{a_0, a_1, a_2, \cdots\}$ are the associated coefficients.

There are many ways in which this decomposition can be useful. For instance, if we let $z = e^{j\Omega_o \omega}$, the above analysis is a Fourier transformation on the power spectrum, and a_n denotes the intensity for component having a "quefrency" $n\Omega_o$. Such $\{a\}_n$ is very similar to the well known *cepstral* coefficients of the source signal[4].

A slightly different and more useful decomposition of the power spectrum results if we let $z = e^{j\sigma_o s}$ in eq. 28, where $s = \log\omega$. This is essentially a Fourier transformation of the log-frequency power spectrum $S_x(s) = S_x(\log\omega)$ since frequency in the auditory model is logarithmically mapped on the s-axis (see section 4). Such a decomposition simply analyzes $S_x(s)$ into *ripples*, or sinusoidally-modulated (against $\log\omega$) power spectra, of different frequencies[14]. The coefficient a_n denotes the intensity of the component that has a ripple frequency $n\sigma_o$.

Another useful decomposition, usually known as *orthogonal polynomial decomposition*, is to let z_n denote an n^{th} degree polynomial (with the leading coefficient being 1) from an orthogonal bases [17]. a_n then denotes the n^{th} order characteristics of the power spectrum, e.g., a_1, a_2 are the overall slope and quadratic curvature of the spectrum, respectively.

In all cases, the transformation of $S_x(\omega)$ into $S_x(z)$ can be viewed as a mapping similar to a 'change of coordinate' that merely provides an alternative description for the power spectrum of the source signal. Note that for all the decomposition methods discussed above, the coefficient a_0 represents the overall level of the sound signal.

5.1.1 *Decomposing the auditory spectrum*

Now consider the corresponding description for the auditory power spectrum in the high gain case. From eq. 20 we have

$$\begin{aligned} E[y_5(\cdot, s)]^2 &\propto \frac{\int |\omega \partial_s H(\omega, s)|^2 S_x(\omega) d\omega}{\int |\omega H(\omega, s)|^2 S_x(\omega) d\omega} \\ &= \frac{\int_{\Omega_d(s)} v(\sigma + s) S_x(\sigma) d\sigma}{\int_{\Omega_c(s)} u(\sigma + s) S_x(\sigma) d\sigma} \\ &\equiv A(s) \end{aligned} \tag{29}$$

[4]Generally, the cepstrum coefficients refer to a Fourier analysis on the *logarithmic* power spectrum. Though not exactly the same as conventional definition, the term "cepstral analysis" is used here to emphasize the significance of analyzing the power spectrum in its Fourier domain.

in which we refer to the fact that the spatial index s is interchangeable with logarithmic frequency due to the dilation relationship. From eqs. 18 and 19, $v(s)$ and $u(s)$ can also be shown to be $|a^{2s}H_m(a^s\omega_m)|^2$ and $|a^{2s}\partial_s H_m(a^s\omega_m)|^2$, respectively. They can also be decomposed into

$$u(z) = u_0 + u_1 z + u_2 z^2 + \cdots \tag{30}$$

$$v(z) = v_0 + v_1 z + v_2 z^2 + \cdots \tag{31}$$

By rewriting eq. 29 into a z-domain, it directly follows that

$$\begin{aligned} A(z) &= \frac{v_0 a_0 + v_1 a_1 z + \cdots}{u_0 a_0 + u_1 a_1 z + \cdots} \\ &= a'_0 + a'_1 z + a'_2 z^2 + \cdots \end{aligned} \tag{32}$$

Note that the cross terms all vanish since z^n and z^m for $n \neq m$ are orthogonal to each other. For ripple bases $z = e^{j\sigma_o s}$, $\{a'\}_n$ are the Fourier coefficients for the auditory power spectrum, which are analogous to the cepstral coefficients. It should be re-emphasized here that the analysis of the auditory pattern here is a local operation. This is highlighted by the integral ranges $\Omega_d(s)$ and $\Omega_c(s)$, and the translation of $v(\cdot + s)$ and $u(\cdot + s)$ in eq. 29. Therefore, the decomposition coefficients $\{u\}_n$ and $\{v\}_n$ are actually indexed by s implicitly.

5.1.2 Relating the $\{a'\}_n$ coefficients to the $\{a\}_n$'s

With the above formulation, the information conveyed in the auditory spectrum and the effects of auditory processing on source spectral characteristics can be further studied by examing $\{a'\}_n$. By applying long division to eq. 32, we obtain the first two coefficients as

$$a'_0 = v_0/u_0 \tag{33}$$

$$a'_1 = (v_1/u_0 - v_0 u_1/u_0^2)(a_1/a_0) \tag{34}$$

For high order coefficients, there is a trade-off between the generality and simplicity of the analysis by approximating the self-normalization, namely, reducing the degree of the denominator in eq. 32. Recall that this denominator is simply the acoustic power spectrum smoothed by the relatively broad cochlear filter (eq. 29). The justification for such an approximation depends on the bases into which the source or auditory power spectrum are analyzed. For instance, in the ripple decomposition, one may analyze locally the auditory spectrum in terms of the harmonics of a fundamental ripple σ_0 (i.e., $z = e^{j\sigma_o}$) whose period is commensurate (in a mean square sense) with the width of the cochlear filter (fig. 9). In this case, $u(z) \approx u_0 + u_1 z$ and the high order terms $u_2, u_3, \cdots$ are relatively insignificant in comparison with u_0. Using the first order approximation for the denominator,

$$A(z) \approx \frac{v_0 a_0 + v_1 a_1 z + v_2 a_2 z^2 + \cdots}{u_0 a_0 + u_1 a_1 z} \tag{35}$$

and hence the n^{th}-order coefficients are recursively defined as

$$a'_n = (v_n/u_0 a_0)a_n - (u_1 a_1/u_0 a_0)a'_{n-1} \tag{36}$$

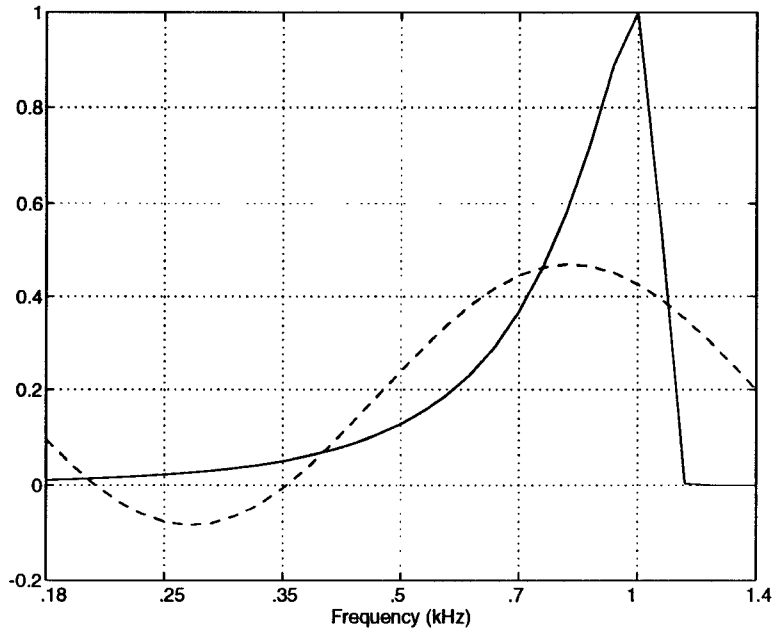


Figure 9: An illustration of first order approximation. The $u(s) = |a^{2s}H_m(a^s\omega_m)|^2$ is shown in solid trace and, in this example, is approximated by a ripple with 1/3 cycles/octave (dashed line).

Furthermore, when $|u_1|^n \ll |u_0|$ for $n > 1$, the feed-back type of recursion (eq. 36) can be rewritten into a feed-forward fashion

$$a'_n = (v_n/u_0)(a_n/a_0) - (v_{n-1}u_1/u_0^2)(a_1/a_0)(a_{n-1}/a_0) \tag{37}$$

Given the above relationships between the coefficients of the acoustic and the auditory power spectral decompositions, we can now examine in more detail how various spectral features are manifested in the early auditory system.

5.2 Level-tolerance and robustness to wide-band adjustment

As pointed out in the previous section, level tolerance is an immediate consequence of self-normalization. This is also evident in the context of the present analysis since dividing both the numerator and denominator in eq. 32 by a_0 (the overall level of the sound signal) does not affect any of the coefficients $\{a'\}_n$. Therefore, the auditory spectrum represents *only* the level-independent attributes of the acoustic features.

Level-independent (or scale-invariant) features have been the focus of many psychoacoustical experiments, such as those known as *profile analysis*[11]. In other studies, it has been established that the absolute sound level plays only a minor role in many aspects of sound perception. For example, it is shown in [17] that when the temporal variation of sound level is artificially limited or "compressed", the perception of speech for human subjects is not only unimpeded, but improved. Similarly, experiments based on noise masking paradigms suggest that it is the relative (not absolute) intensity of

the signal and noise that dominates the perception of the signal. [22]. Finally, in many applications where speech and other acoustic signals are employed, level-independent information is often implicitly extracted or emphasized by the use of logarithmic or decibel signal spectra.

In the auditory spectrum, overall level-tolerance is only a special case of a general insensitivity to broad scale spectral adjustments (as illustrated in section 4.4). The limits of this insensitivity can be deduced from eq. 29 by noting that spectral components that can be resolved by both the cochlear and differential filters are subject to cancellation. Hence, all spectral fluctuations that are of a similar or a broader scale than the bandwidth of the cochlear filters will be attenuated in the auditory spectrum. Equivalently, in terms of the decompositions described above (eq. 35), spectral adjustments characterized by coefficients a_n for $n \geq 2$ will be undistortedly represented in $A(z)$; those characterized by a_1 will be attenuated and the overall level described by a_0 will be eliminated. The perceptual implications of these statements are discussed in more detail below.

5.3 Characteristics of low-order coefficients a'_0 and a'_1

The two low-order coefficients of the auditory spectrum a'_0 and a'_1 are exactly computable from eqs. 33 and 34.

5.3.1 Interpretation of the a'_0

From eq. 33, a'_0 is independent of the sound signal and only reflects the relative gains of the differential and cochlear filters. The auditory spectrum, therefore, provides a common dc reference for all signals. For example, in fig. 7, the auditory spectra for the vowel in various levels of noise maintain roughly the same level despite the increase of total energy in the stimulus. The auditory spectrum barely looks flattened with the increasing noise power. This property is useful for many engineering applications especially when combined with the limited dynamic range in the auditory spectrum. Note that for the white noise, the power spectrum has only the constant term a_0 and therefore $A(s) = v_0/u_0$, which is equal to κ^2 other than a constant scaling factor (κ is defined in sec. 4).

5.3.2 Interpretations of the a'_1

Unlike the a'_0, the exact meaning of the a'_1 coefficient depends on the type of decomposition employed. For instance, in terms of a polynomial decomposition, a'_1 denotes the spectral tilt in the auditory spectrum. This tilt, however, is proportional to the *normalized* spectral slope in the source power spectrum, i.e., it represents a level-independent spectral tilt (see eq. 34). This is a desirable attribute since without the normalization, the spectral tilt would vary with the overall sound level, and hence would not serve as a distinguishing cue for phoneme classifications, such as among stop consonants[32, 5]. Another observation from eq. 34 is that the auditory spectral tilt is attenuated (by the factor $v_1/u_0 - v_0u_1/u_0^2$) relative to that of the power spectrum, a property for many speech recognition applications [15].

In the case of ripple decomposition, a'_1 represents the *relative* intensity of the fundamental ripple component in the auditory power spectrum. A perceptually important special case of this decompo-

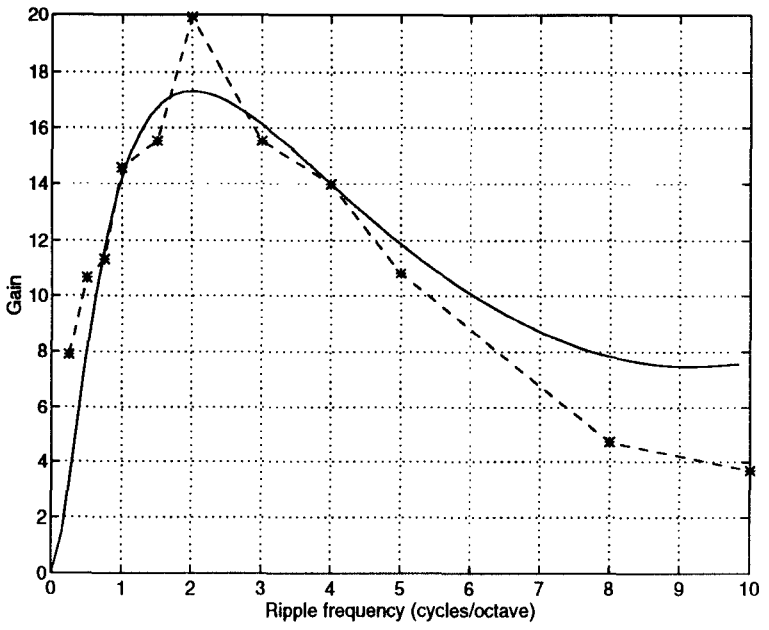


Figure 10: The contrast sensitivity function obtained from psychoacoustic experiments (dashed trace, adapted from [14], fig. 3.26) and model prediction (solid). The psychoacoustic experiments used single ripple stimuli and the thresholds for detecting the ripple, in peak-to-valley ratio, were recorded for various ripple frequencies. The measured thresholds are then converted from dB to linear scale and inverted to yield the 'gain' depicted in this figure. The model prediction, $|v_1/u_0 - v_0 u_1/u_0^2|$ is scaled to fit the psychoacoustic curve.

sition occurs when the source signal contains only one ripple, i.e., $S_x(z) = a_0 + a_1 z$. The auditory power spectrum becomes a harmonic series, with the n^{th} component weighted by a'_n (see eq. 32). The largest term, $|v_1/u_0 - v_0 u_1/u_0^2|$ from eq. 34, can therefore roughly be interpreted as the 'gain' of the ripple since a'_1 is linearly proportional to a_1/a_0. This term (as a function of ripple frequency) is closely analogous to the *contrast sensitivity function* commonly measured in psychophysical studies of the visual system [33]. It has recently been psychoacoustically measured for a wide range of ripple frequencies[14]. Fig. 10 shows the experimental results and the prediction from the auditory model. Both curves display the bandpass nature of the ripple sensitivity, namely, subjects are most sensitive to intermediate ripple frequencies around 2 cycles/octave. In the auditory model, the high ripple frequency roll-off is caused by the finite bandwidth of the differential filters, whereas the low ripple frequency roll-off is due to the self-normalization, which in turn reflects the broad bandwidth of the cochlear filter[5]. Note that, if the coefficients of the ripple decomposition here are viewed as analogous to the commonly used cepstral coefficients, then the transfer function of Fig. 10 is very similar to the

[5]Since the cochlear filter is much broader than the differential filter, $|u_1| \ll |v_1|$ at high ripple frequencies. Hence the gain term $(v_1/u_0 - v_0 u_1/u_0^2)$ is dominated by the characteristics of the differential filter, which is itself a low pass filter for ripples. Therefore the gain drops at the same rate as the ripple transform of the differential filter v_1. On the other hand, when the ripple frequency is low, $v_1 \rightarrow v_0$ and $u_1 \rightarrow u_1$. Accordingly, the gain decreases with decreasing ripple frequency being subject to self-normalization.

bandpass liftering often employed in ASR systems [15]. As yet, there have been no neurophysiological experiments carried out in the early auditory system with rippled spectral stimuli to test these hypotheses directly.

5.4 Characteristics of the high-order coefficients

Higher-order coefficients capture the rapidly varying features of the power spectral shape. Under the first degree approximation on the self-normalization, eqs. 37 and 36 indicate that, if a_1/a_0 is regarded as constant, a high-order coefficient $\{a'\}_n$ is also *linearly* related to the coefficients a_n/a_0 of the source power spectrum. This important finding is intuitively surprising given the many nonlinearities of the model. It implies that linearly filtering (i.e., weighting and/or linearly combining the coefficients of) an arbitrary (normalized) source power spectrum is a justified and easily computable procedure that has well understood effects on the auditory representation. For instance, this is the most important reason why the contrast sensitivity function (fig. 10), which is measured with a single ripple at a time, is still relevant for the analysis of complex spectral shapes (e.g., of a vowel) composed of many ripples.

The linearity of the high-order coefficients manifests itself as a *superposition principle*. More specifically, suppose

$$S_x(z) = a_0 + a_1 z + \sum_{n>1} \alpha_n z^n + \sum_{n>1} \beta_n z^n \tag{38}$$

where the α and β coefficients belong to two superimposed spectra (e.g., of two vowels), and a_0 and a_1 are the low order coefficients of the composite. Then it directly follows from eq. 37 that

$$\begin{aligned} a'_n &= \frac{v_n}{u_0}\left(\frac{\alpha_n + \beta_n}{a_0}\right) - \frac{v_{n-1}u_1 a_1}{u_0^2 a_0}\left(\frac{\alpha_{n-1} + \beta_{n-1}}{a_0}\right) \\ &= \left[\frac{v_n \alpha_n}{u_0 a_0} - K a'_1 \left(\frac{v_{n-1}}{u_0}\right)\left(\frac{\alpha_{n-1}}{a_0}\right)\right] \\ &\quad + \left[\frac{v_n \beta_n}{u_0 a_0} - K a'_1 \left(\frac{v_{n-1}}{u_0}\right)\left(\frac{\beta_{n-1}}{a_0}\right)\right] \\ &= \alpha'_n + \beta'_n \end{aligned} \tag{39}$$

where α'_n (or β'_n) is the n^{th} coefficient for the auditory spectrum due to one signal without the presence of the other.

Therefore, the model demonstrates that auditory processing possesses a quasi-linear property on the level-independent acoustic features. This conclusion is consistent with the fundamental assumptions underlying many psychoacoustical models and experiments which treat the auditory periphery as a bank of linear filters applied to a logarithmic spectrum (i.e., with level-independent features) [23, 22, 9].

5.5 Dominance principle and noise robustness

The previous superposition principle is obtained by fixing the composite a'_1 during the derivation. It can actually be shown that the first order components also abide by a superposition principle, though

not a strictly linear one. Suppose the source signal consists of two (unrelated) spectra that can be described by $\{\alpha\}_n$ and $\{\beta\}_n$ respectively, i.e.,

$$S_x(z) = \sum \alpha_n z^n + \sum \beta_n z^n$$

Then, according to eq. 34, the first-order coefficients are:

$$\begin{aligned} a'_1 &= (v_1/u_0 - v_0 u_1/u_0^2)(a_1/a_0) \\ &= (v_1/u_0 - v_0 u_1/u_0^2)(\alpha_1 + \beta_1)/(\alpha_0 + \beta_0) \\ &= \frac{\alpha_0}{\alpha_0 + \beta_0}\alpha'_1 + \frac{\beta_0}{\alpha_0 + \beta_0}\beta'_1 \end{aligned} \tag{40}$$

i.e., they satisfy a *weighted* superposition principle. When the absolute level of one component is overwhelmed by the other, say $\beta_0 \ll \alpha_0$, a'_1 approaches the coefficient of the *dominant* component (α'_1),. Such a *dominance principle* can be generalized for signals with several individual spectra, and the resultant a'_1 would be the centroid of the corresponding normalized a_1's.

The dominance principle described above can now be employed to re- interpret the noise robustness property which was explained as a suppression phenomena in the previous section. For instance, consider the example of a signal ($\{\alpha\}_n$) in white noise (with level β_0) as in section 4.5, i.e.,

$$S_x(z) = \sum \alpha_n z^n + \beta_0$$

and the corresponding auditory coefficients for $A_x(z)$ are given by

$$\begin{aligned} a'_1 &= \left(\frac{\alpha_0}{\alpha_0 + \beta_0}\right)\alpha'_1 \\ a'_n &= \left(\frac{\alpha_0}{\alpha_0 + \beta_0}\right)\alpha'_n + \left(\frac{\alpha_0}{\alpha_0 + \beta_0}\right)^2 \alpha'_1 \alpha'_{n-1} \end{aligned}$$

where $\{\alpha'\}_n$ denote the auditory coefficients for the clean speech $A_c(z)$. Note that $\alpha_0/(\alpha_0 + \beta_0) = 1/(1+\beta_0/\alpha_0)$, in which β_0/α_0 is the reciprocal of SNR. When SNR is large, $a'_n \approx \alpha'_n$, the coefficient for clean speech. However, when SNR is small, the distortion, measured by Euclidean distance between $A_x(z)$ and $A_c(z)$, roughly increases by the order of $1 - \alpha_0/(\alpha_0 + \beta_0) = (\beta_0/\alpha_0)/(1 + \beta_0/\alpha_0)$. In comparison, the Euclidean distortion for the linear power spectrum increases by the order of β_0/α_0. Therefore, the distortion of linear power spectrum increases at a rate $(1 + \beta_0/\alpha_0)$ faster than the auditory spectrum. This trend was previously demonstrated in fig. 8.

Finally, note that for general non-white noise, the effect of noise on the source power spectrum can be viewed as adding a jitter upon the spectral shape. Therefore, interpreting eq. 37 (or 36) as a feedforward (feedback) one-step estimation on a_n from a_{n-1} means that the auditory spectrum a'_n simply encodes the scaled prediction error. Consequently, the auditory spectrum smooths out the added noise through the prediction process. In addition, it will have the nice properties common to prediction errors, such as having a smaller dynamic range (which in this case is roughly bounded by the ratio of the gains of the differential and cochlear filters). This suggests that auditory model may be a suitable front-end for a variety fields of engineering applications such as recognition and coding.

6 Summary and Conclusions

In summary, we have analyzed and interpreted the properties of a simple analytic model of the early auditory system. The article had three specific objectives: (1) To provide a tractable theoretical framework for this and future analysis of early auditory processing; (2) To interpret the results in the context of basic known psychoacoustical findings; (3) To help justify the advantages of auditory-like and other representations as front-ends for speech recognition systems. Within each of these objectives, the following issues were discussed:

- Theoretical analysis:
  - Stochastic analysis of the model (section 3).
  - Specific extension of the analysis through spectral decomposition to highlight the acoustic-to-auditory transformation of power spectral shape (section 5.1).
- Relation to psychoacoustics:
  - The origin of level-tolerance in the auditory spectrum, and its psychoacoustical correlates (section 5.2).
  - The representation of normalized spectral slopes in the auditory pattern, and its relation to speech perception (section 5.3).
  - The auditory representation of rippled spectra, and its relation to the *contrast sensitivity function* (section 5.3).
  - The linearity of the auditory representation, and its justification of basic assumptions of the psychoacoustical models of the auditory periphery (section 5.4).
- Relevance to engineering applications:
  - For auditory-like models, the analysis explains and quantifies the origins of their experimentally observed noise-robustness as suppression (sections 4.2 and 4.5), as a dominance principle (section 5.5), or as a linear prediction process (section 5.5).
  - The auditory spectrum exhibits (and hence justifies) very similar deformations of the spectrum as those implied by commonly used algorithms in ASR systems. They include the attenuation of spectral slopes (section 5.3), bandpass liftering of cepstral coefficients (section 5.3).

Acknowledgement

The authors would like to thank Dr. L. Lee and E. T. Gan for their enlighting discussions, and three anonymous reviewers for their constructive suggestions regarding the manuscript. Portions of this article have been presented at ICASSP-93 in April, 1993 at Minneoplis, MN. This work is funded by grants from the Office of Naval Research, the Air Force Office of Scientific Research, and by NSF's Engineering Research Centers Program NSFD CD-8803012.

References

[1] J. B. Allen. Cochlear modeling. *IEEE ASSP Magazine*, January 1985.

[2] P. F. Assmann and D. D. Paschall. Autocorrelogram models of the segregation of competing voices. 15th Winter ARO meeting, February 1992.

[3] 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.

[4] A. Bladon. Using auditory models for speaker normalization in speech recognition. In *Proc. Symp. Speech Recognition*, Montreal, Canada, 1986.

[5] S. E. Blumstein and K. N. Stevens. Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America*, 67(2):648–662, February 1980.

[6] 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.

[7] 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.

[8] 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.

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

[10] R. M. Gray and L. D. Davisson. *Random processes.* Prentice-Hall Inc., Eaglewood Cliffs, New Jersey, 1986.

[11] D. M. Green. *Profile Analysis: Auditory Intensity Discrimination.* Oxford Science Publications, New York, NY, 1988.

[12] S. Greenberg. Acoustic transduction in the auditory periphery. *Journal of Phonetics*, 16(1):3–18, 1988.

[13] H. Hermansky, K. Tsuga, S. Makino, and H. Wakita. Perceptually based processing in automatic speech recognition. In *Proc. ICASSP-86*, Tokyo, Japan, 1986.

[14] D. A. Hillier. *Auditory processing of sinusoidal spectral envelopes.* PhD thesis, Washington University, Sever Institute of Technology, 1991.

[15] B. H. Juang, L. R. Rabiner, and J. G. Wilpon. On the use of bandpass liftering in speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(7):947–954, July 1987.

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

[17] H. Levitt and A. Neuman. Evaluation of orthognal polynomial compression. *Journal of Acoustical Society of America*, 90(1):241–252, July 1991.

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

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

[20] P. L. Meyer. *Introductory Probability and statistical applications.* Addison-Wesley Publishing Co., Reading, MA, 1970.

[21] M. Miller and M. B. Sachs. Representation of stop consonants in discharge patterns of auditory nerve fibers. *Journal of the Acoustical Society of America*, 74:502–517, 1983.

[22] B. C. J. Moore and B. J. O'Loughlin. The use of nonsimultaneous masking to measure frequency selectivity and suppression. In B. C. J. Moore, editor, *Frequency Selectivity in Hearing*, chapter 4, pages 179–250. Academic Press Inc., 1986.

[23] R. D. Patterson and B. C. J. Moore. Auditory filters and excitation patterns as representations of frequency resolution. In B. C. J. Moore, editor, *Frequency Selectivity in Hearing*, chapter 3, pages 123–177. Academic Press Inc., 1986.

[24] L. R. Rabiner and R. W. Schafer. *Digital processing of speech signals.* Prentice-Hall Inc., 1978.

[25] M. B. Sachs and E. D. Young. Effects of nonlinearities on speech encoding in the auditory nerve. *Journal of the Acoustical Society of America*, 68:858–875, 1980.

[26] S. Seneff. A joint synchrony/mean-rate model of auditory processing. *Journal of phonetics*, 85(1):55–76, 1988.

[27] S. A. Shamma. Speech processing in the auditory system II: lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *Journal of the Acoustical Society of America*, 78(5):1622–1632, November 1985.

[28] S. A. Shamma. The acoustic features of speech sounds in a model of auditory processing: vowels and voiceless fricatives. *Journal of Phonetics*, 16:77–91, 1988.

[29] S. A. Shamma. Spatial and temporal processing in central auditory networks. In C. Koch and I. Segev, editors, *Methods in Neural Modelling.* MIT Press, Cambridge, MA, 1989.

[30] 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), July 1986.

[31] W. M. Siebert. Frequency discrimination in the auditory system: place or periodicity mechanisms? *Proc. IEEE*, 58:723–730, 1970.

[32] K. N. Stevens and S. E. Blumstein. Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 64(5):1358–1368, November 1978.

[33] R. L. D. Valois and K. K. D. Valois. *Spatial Vision.* Oxford Science Publications, New York, NY, 1990.

[34] K. Wang. Neural networks that recognize phonemes by their acoustic features. Master's thesis, University of Maryland, December 1989.

[35] K. Wang and S. A. Shamma. A functional model of the early auditory system. In *Proc. International Symposium on Time-Frequency and Time-Scale Analysis*, Canada, October 1992.

[36] K. Wang and S. A. Shamma. Zero-crossings and noise suppression in auditory wavelet transformations. Technical Report TR 92-94, Institute for Systems Research, University of Maryland, 1992.

[37] L. A. Westerman and R. L. Smith. Rapid and short term adaptation in auditory nerve responses. *Hearing Research*, 15:249–260, 1984.

[38] X. Yang, K. Wang, and S. 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.