TECHNICAL RESEARCH REPORT

The Case of the Missing Pitch Templates: How Harmonic Templates Emerge in the Early Auditory System

by Shihab Shamma, David Klein

T.R. 99-27

ISR

INSTITUTE FOR SYSTEMS RESEARCH

The Case of the Missing Pitch Templates:
How Harmonic Templates Emerge
in the Early Auditory System

Shihab Shamma   David Klein

Center for Auditory and Acoustics Research, Institute for Systems Research Electrical Engineering Department, University of Maryland, College Park, MD

Abstract

Periodicity pitch is the most salient and important of all pitch percepts. Psycho-acoustical models of this percept have long postulated the existence of internalized harmonic templates against which incoming resolved spectra can be compared, and pitch determined according to the best matching templates (Goldstein, 1973a). However, it has been a mystery where and how such harmonic templates can come about. Here we present a biologically plausible model for how such templates can form in the early stages of the auditory system. The model demonstrates that *any* broadband stimulus such as noise or random click trains, suffices for generating the templates, and that there is no need for any delay-lines, oscillators, or other neural temporal structures. The model consists of two key stages: cochlear filtering followed by coincidence detection. The cochlear stage provides responses analogous to those seen on the auditory-nerve and cochlear nucleus. Specifically, it performs moderately sharp frequency analysis via a filter-bank with tono-topically ordered center frequencies (CFs); the rectified and phase-locked filter responses are further enhanced temporally to resemble the synchronized responses of cells in the cochlear nucleus. The second stage is a matrix of coincidence detectors that compute the average pair-wise instantaneous correlation (or product) between responses from all CFs across the channels. Model simulations show that for any broadband stimulus, high coincidences occur between cochlear channels that are exactly harmonic distances apart. Accumulating coincidences over time results in the formation of harmonic templates for all fundamental frequencies in the phase-locking frequency range. The model explains the critical role played by three subtle but important factors in cochlear function: the nonlinear transformations following the filtering stage; the rapid phase-shifts of the traveling wave near its resonance; and the spectral resolution of the cochlear filters. Finally, we discuss the physiological correlates and location of such a process and its resulting templates.

Contents

1   Introduction and Background

More than any other auditory percept in the last century, pitch has been a potent source of inspiration and controversy in auditory research. Its importance stems from its role in perceiving the prosody of speech, melody of music, and in organizing the acoustic environment into different sources (Summerfield and Assmann, 1990; de Cheveingne et al., 1995). It is generally appreciated that the term "pitch" refers to many distinct percepts (de Cheveingne, 1998; Moore, 1989): They include "spectral pitch" evoked by single tones, "repetition pitch" associated with very slow click trains, or the envelope of amplitude modulated noise and tones, and "periodicity pitch" (also known as virtual, residue, and missing fundamental pitch) evoked by low order, spectrally resolved harmonic tone complexes. The focus of this paper is on "periodicity pitch", the pitch usually associated with musical intervals and melodies, and with speakers voices and speech prosody.

There is general agreement on the perceptual properties and acoustic parameters that give rise to periodicity pitch in humans, and presumably in other mammals and birds (Langner, 1992; Moore, 1989; Plomp, 1976). For instance, the most salient pitch is evoked by harmonically related tone complexes that are spectrally (at least partially) resolved; the pitch heard is that of the fundamental frequency of these harmonics regardless of the energy in that fundamental component; pitch values heard are roughly in the range 50-2000 Hz; The most effective (or dominant) harmonics are the low order harmonics (the 2-5th harmonics); The saliency of the pitch increases with the number of resolved harmonics; And multiple pitches are usually perceived if there are a few harmonics in the complex, or if the tones form an inharmonic sequence.

Numerous theories have been proposed to account for periodicity pitch percepts. Most successful among them are the so-called "spectral pitch theories", best exemplified by the "pattern recognition" theories (Goldstein, 1973b; Terhardt, 1979; Wightman, 1973), and the variations and implementations proposed since then (Duifhuis et al., 1982; Cohen et al., 1995). The two operations common to all are: (1) the pitch value is derived (centrally) from a spectral profile defined along the tonotopic axis of the cochlea (regardless of how this profile is computed); and (2) the input spectrum is compared to internally stored spectral templates, consisting of the harmonic series of all possible fundamentals. These theories have been enormously successful in explaining and predicting the pitches of complex tones, and consequently have provided the dominant view of pitch perception.

Spectral pitch theories, however, suffer two criticisms. The first is the lack so far of convincing biological evidence for the existence of these templates or for how they might be generated. "Learning" the harmonic templates has usually been assumed to be a straight-forward consequence of frequent exposure during early development to speech or natural sounds which tend to be rich in harmonic structure (Terhardt, 1979). However, there are several difficulties with this scenario. Infants are thought now to be born with an innate well developed sense of musical pitch (Clarkson and Rogers, 1995; Montgomery and Clarkson, 1997), presumably long before any serious exposure to speech (sounds in the womb are predominantly noise-like and due to the heart and other internal organs). Another difficulty is that voiced speech usually has a weak or absent fundamental component, raising the question of why learned templates consisting of prominent higher harmonics are perceived at (or are linked to) the pitch of the fundamental and

not any other arbitrary frequency. A second criticism of spectral pitch theories is their inability to account for other weaker pitch percepts such as "repetition pitch", which apparently operate in different parameter ranges, and require different mechanisms.

To address these criticisms, alternative theories have been proposed to explain how the pitch percept might be computed without need for explicit stored harmonic templates. These theories can be described as "temporal" in that they postulate mechanisms that extract a pitch value from the temporal response waveforms in each auditory channel (independent of other channels), and then combine the results from all channels to get the final estimate. As such, "temporal" theories unlike "spectral" theories, make no use of an ordered tonotopic axis, i.e., their computations are unaffected by a shuffling of the tonotopic axis (Lyon and Shamma, 1996).

"Temporal" models vary enormously in the nature of the cues they utilize from each channel, e.g., first or higher order intervals, autocorrelations of the responses, or synchronization measures; they also differ in the mechanisms to measure them, e.g., delay lines and coincidence detectors, or intrinsic oscillators (de Cheveingne, 1998; Slaney and Lyon, 1993; Cariani and Delgutte, 1996; Lickleider, 1951; Meddis and Hewitt, 1991; Patterson and Holdsworth, 1991; Langner and Schreiner, 1988). One often stated advantage of these theories is that most can account for both repetition pitch as well as periodicity pitch with the same mechanisms.

There are several problems with these temporal models. For instance, recent psycho-acoustical evidence does not support the unitary model of pitch (Carlyon, 1998). Furthermore, physiological support for these theories (just as with the harmonic templates) is spotty at best. Thus, while many central auditory responses can be interpreted loosely as exhibiting delays or appropriate oscillatory patterns, the anatomical and physiological data do not yet coalesce as a whole into a compelling picture (Langner, 1992). Most physiological data for pitch tend to be in frequency ranges and from units with best frequencies that are relevant for repetition pitch rather than periodicity pitch (Schreiner and Langner, 1988; Schreiner and Urbas, 1988; Schwartz and Tomlinson, 1990). Finally, most purely temporal theories are descriptive and intuitively appealing, rather than truly predictive as with the spectral pitch theories; hence, it is often difficult to compare their predictions to psycho-acoustical results.

In summary, it is fair to say that spectral pitch theories would be more palatable to many if it were not for the obvious lack of a biologically compelling mechanism for how harmonic templates might come about, and of evidence for their existence. This paper attempts to address the first problem, and provide hints at what to look for, and where to search for physiological and anatomical evidence.

Organization of the paper:

The model we describe here explains how harmonic templates could emerge as a simple consequence of coincidence detection among channels representing the outputs of a cochlear-like filter bank. Once formed, the templates can presumably be used to estimate the pitch as in the many variants of the spectral-matching pitch algorithms. Our focus in this paper is on the template-formation phase. Our goal is to illustrate how biologically plausible processes, response patterns, and connectivity in the early auditory nuclei can give rise to ordered harmonic templates without the need for any specially tailored inputs (such as clean harmonic complex tones), or supervised constraints (such as labeled and ordered inputs and outputs).

In the following, we shall first illustrate the essential mathematical structure of the model (section 2) and then discuss why the templates emerge (section 3). Next we will discuss the potential biological structures and pathways that underlie the model (section 4). We finally will relate the harmonic templates to those hypothesized in various spectral- pitch estimation models, and discuss the wider implications of our findings to models of auditory processing, and to neural processing strategies in general (section 5).

2 A Mathematical Model for Harmonic Template Generation

The two basic stages of the model are illustrated in Figure 1: An analysis stage consists of filter bank followed by temporal and spectral sharpening analogous to the processing seen in the cochlea and cochlear nucleus. The second stage is a matrix of coincidence detectors that computes the pair-wise instantaneous correlation between all filter outputs.

2.1 The analysis stage

This stage consists of a simplified minimal model of early auditory processing. It consists of a cochlear filter bank, followed by hair cell rectification and central spectro-temporal sharpening. These operations are depicted in Figure 1, and described below in detail.

2.1.1 Cochlear filter bank

We employ a bank of 128 bandpass filters, equally spaced along a logarithmic frequency axis x with center frequencies (CF) spanning 5.3 octaves. The filters are moderately tuned and significantly asymmetric, with a steep roll-off on the high frequency sides, as illustrated in Figure 2A (Wang and Shamma, 1994; Yang et al., 1992). They have constant Q's, and hence their bandwidths gradually broaden towards the higher CFs. They are also related to each other by a simple dilation of their impulse responses. Given a discrete-time signal $s(t)$, and cochlear filter impulse responses $h(t; x)$, $x = 1...128$ and $t = 0, 1, ..., n$, any filter's response is computed as:

$$u(t; x) = s(t) * h(t; x), \tag{1}$$

where $*$ denotes convolution with respect to time.

2.1.2 Hair cell filtering and rectification

Hair cells convert the filter outputs into electrical activity along the tonotopically ordered auditory-nerve array. This biophysical process is usually modeled by a three step process (Shamma et al., 1986; Shamma and Morrish, 1986): a high-pass filter accounting for the velocity coupling of the hair cell cilia; a sigmoid function that describes nonlinear hair cell transducer channels; and a low-pass filter representing the leakage in hair cell currents that gradually attenuates phase-locked responses beyond 1-2 kHz.
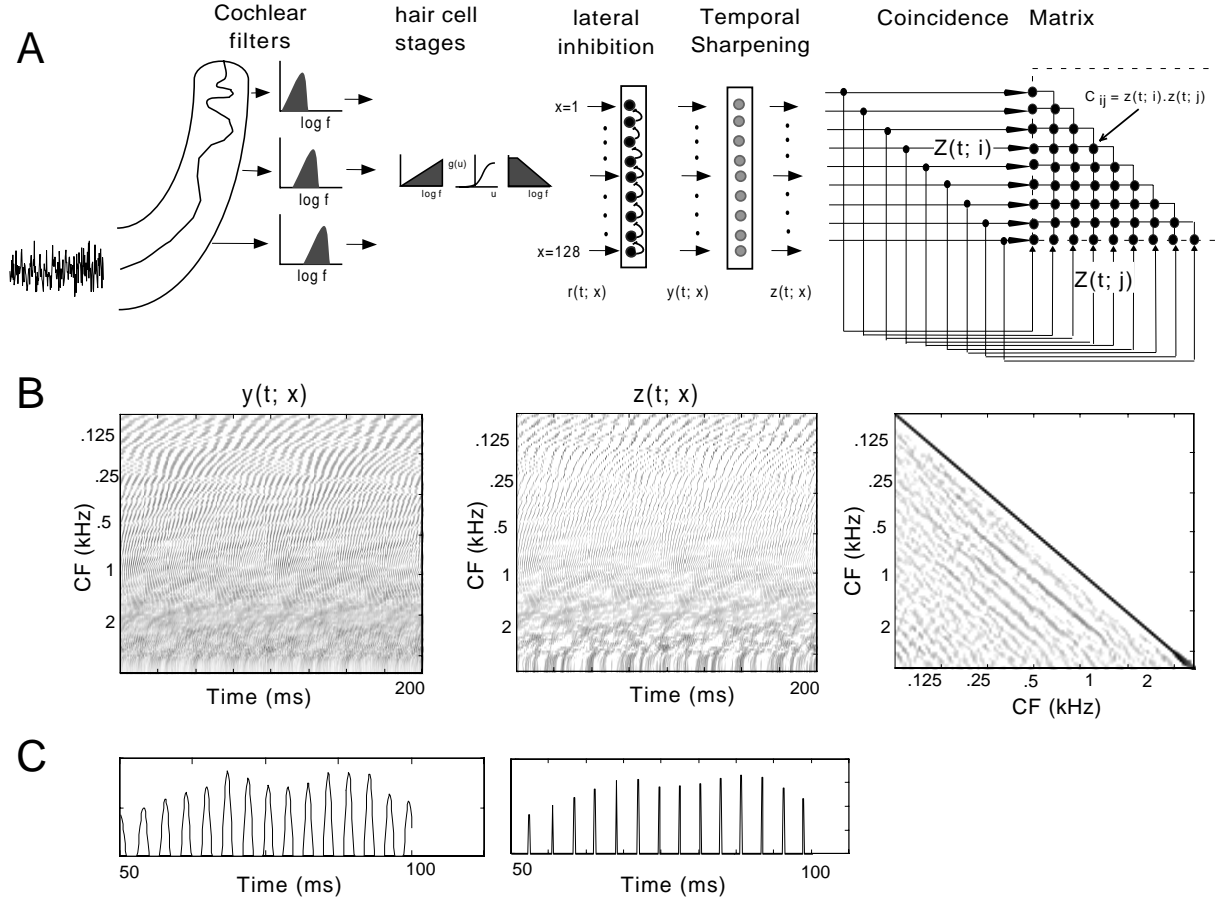
Figure 1: *Schematic model of early auditory stages.* **A:** Sound is analyzed by a bank of 128 tonotopically ordered cochlear filters spanning CFs between 100-4000 Hz. The output waveform from each filter is passed through a hair cell model $(r(t; x))$, followed by a first difference across the channel array simulating the action of a lateral inhibitory network (LIN) $(y(t; x))$. The responses are then temporally sharpened, becoming more synchronized within each channel $(z(t; x))$. The final stage is a matrix of coincidence detectors that compares the responses from all pairs of channels across the array. **B:** The spatiotemporal responses of the channel array at different stages of the model: *(left-to-right)* - The responses at the LIN output $(y(t; x))$; The synchronized responses $(z(t; x))$; the output of the coincidence matrix (C) after one iteration. **C:** The waveform transformation at the synchronization stage. (*Left*) - the waveform at CF$\approx$140 Hz $(y(t; x = 12)$. (*Right*) - the waveform after temporal sharpening.

Here we shall simplify the analysis by incorporating the first temporal derivative into the cochlear filters. Next, the hair cell nonlinearity $g(\cdot)$ is modeled as a simple half-wave rectifier:

$$r(t; x) = g(u(t; x)) = g(s(t) * h(t; x)), \qquad (2)$$

where $g(u) = 0$ for $u < 0$, and $g(u) = u$ otherwise. Note that $g(\cdot)$ can be redefined as a sigmodal function to account for more complex nonlinear effects such as saturation or wider dynamic ranges. The effects of these added modifications is small for reasons discussed later. The hair cell low-pass filter is bundled into the following stage as we describe next. The model outputs at this stage are depicted in Fig.1B-C for a broadband noise stimulus.
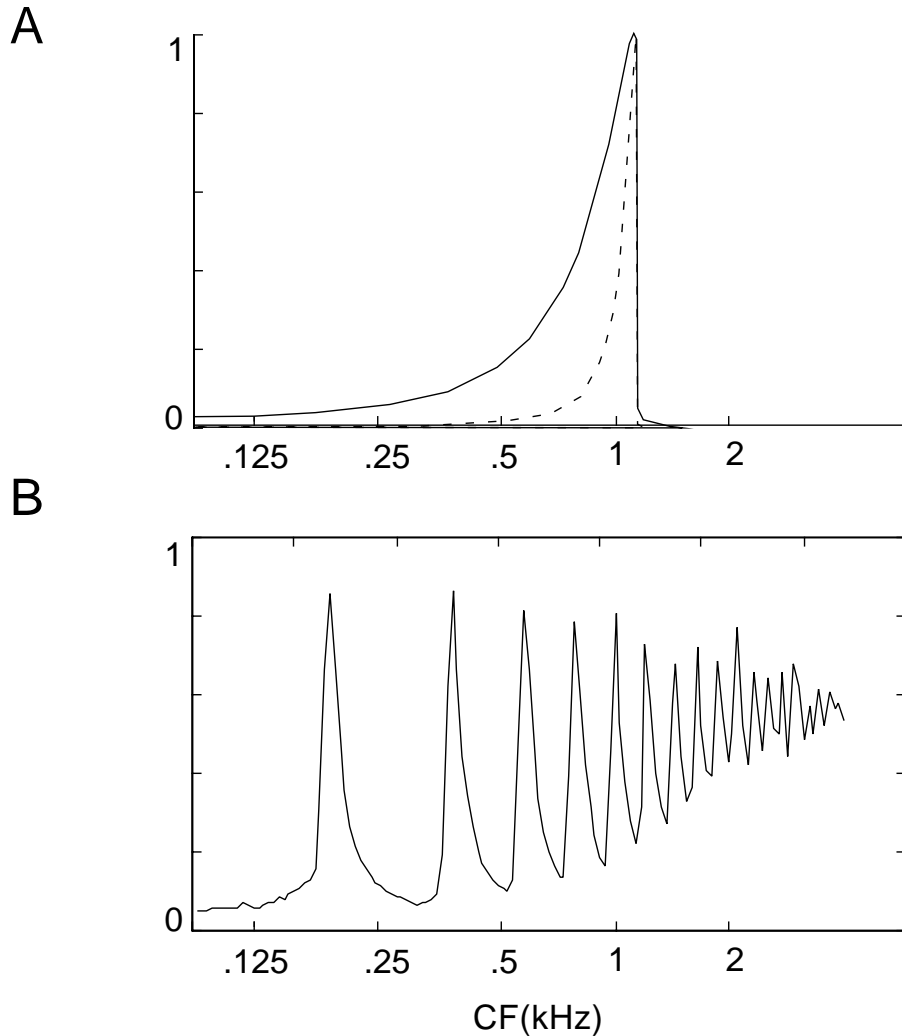
Figure 2: *Details of the model filters.* **A:** (solid line) Magnitude transfer function of the filter at CF=1 kHz; (dotted line) The effective magnitude transfer function *after* the LIN stage (see text). **B:** The integrated output of the LIN ($\sum_t |y(t;x)|$) reflecting the spectrum of a harmonic series stimulus consisting of 20 harmonics of a 200 Hz fundamental (see (Wang and Shamma, 1994) for details).

2.1.3 Spectral and temporal sharpening of the filter outputs

This stage is helpful in enhancing the representation of the harmonics in the templates as we shall discuss later. Spectral sharpening mimics the effect of lateral inhibition (Shamma, 1985a; Shamma, 1985b), and is modeled by a simple derivative across the channel array (or a first-difference operation between the filter outputs)(Wang and Shamma, 1994):

$$y(t;x) = r(t;x) - r(t;x-1), \tag{3}$$

for $x = 2..128$, and y(t ; 1) = 0. It can be shown that this step effectively sharpens the cochlear filters (Wang and Shamma, 1994; Lyon and Shamma, 1996), and is in principle unnecessary if the cochlear filters used are sharp enough to resolve approximately 12-15 harmonics. Figure 2B

illustrates that our frequency analysis at this stage can resolve approximately 14-15 harmonics of a 20-harmonic series stimulus (with a fundamental at 200 Hz).

The next stage performs temporal sharpening which enhances the synchrony of the phase-locked responses. This process mimics transformations such as those seen between the auditory-nerve and the onset units of the cochlear nucleus (Oertel et al., 1990; Palmer et al., 1995; Rhode, 1995). It is approximated by sampling the *positive peaks* of $y(t; x)$:

$$z(t; x) = \sum_{t_p} \delta(t - t_p) \cdot y(t_p; x), \tag{4}$$

where $t_p =$ locations of the positive peaks in time, and $\delta(\cdot)$ is the discrete Dirac delta-function ($\delta(0) = 1$, and $\delta(\cdot) = 0$ otherwise). $z(t; x)$ then becomes a spectrally sharpened and highly temporally synchronized version of the filter responses as illustrated in Fig.1B-C. A simple way to include the effects of the diminishing phase-locked responses with increasing frequency is to replace $\delta(\cdot)$ with a pulse of variable width $\Pi_m(\cdot)$, starting at the zero-crossing point, i.e., $\Pi_m(k) = 1, 0 \leq k \leq m$; the larger m is, the smaller is the frequency range of phase-locking and synchrony (Fig.1C):

$$z(t; x) = \sum_{t_p} \Pi_m(t - t_p) \cdot y(t_p). \tag{5}$$

2.2 The coincidence matching stage

This stage performs an instantaneous match between the responses of all pairs of channels in the array, and integrates all results over time to produce its final output. From a mathematical perspective, the network is a matrix of coincidence detectors, each multiplying the responses from a pair of channels as depicted in Fig.1A:

$$C_{ij}(t) = z(t; i) \cdot z(t; j), \tag{6}$$

for all $i, j = 1...128$ such that $j < i$; For $i = j$, $C_{ii}(t) = z(t; i)$. The absolute values of $C_{ij}(t)$ are then accumulated over time until an adequately smoothed output T_{ij} is obtained:

$$T_{ij} = \sum_{t,N} |C_{ij}(t)|, \tag{7}$$

for N realizations of a random stimulus. Note there are no neural delays anywhere in this model. Instead, coincidences are computed from simultaneous outputs of the filter bank, and the results are then integrated over time.

2.3 Model Simulations

The coincidence network above is capable of producing the harmonic templates as its final averaged output regardless of the exact nature of its input signal, provided it is broadband conveying energy at all frequencies < 3 kHz. We illustrate in Figure 3 the templates generated with broadband noise and random click train input signals $s(t)$. In Fig.3A, the 200 msec stimulus consists of equally-spaced, random-phase, tones (with 10 Hz separation, in the range
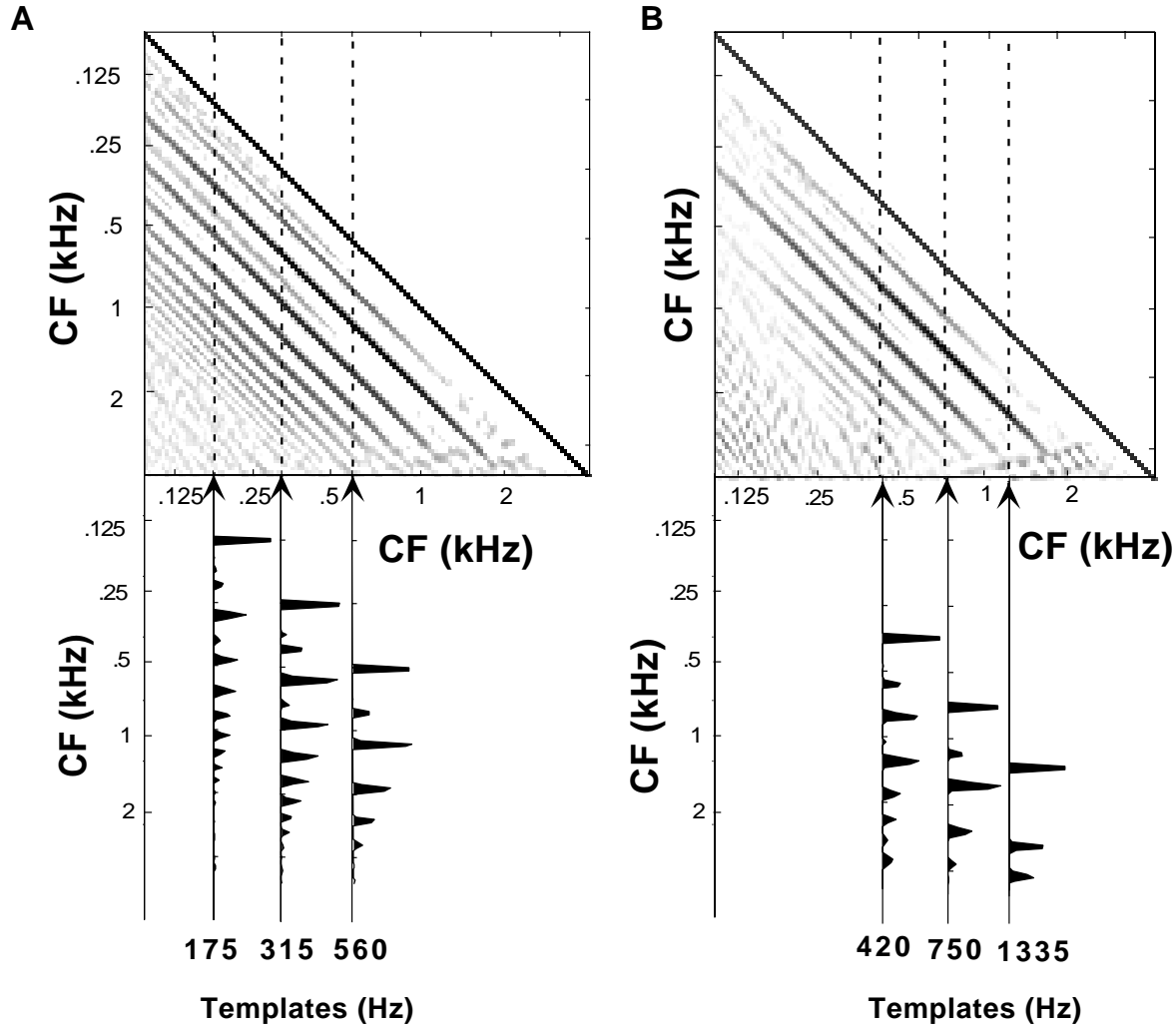
Figure 3: *The harmonic templates in the integrated output of the coincidence matrix.* Templates emerge as regions of high coincidence that run parallel to the main diagonal, and are exactly spaced at harmonically related CF distances. **A:** The templates generated by a broadband noise stimulus. Three templates are shown individually by the cross sections (fundamentals at 175, 315, 560 Hz). For each, the pattern shows prominent peaks at harmonically related CFs, that gradually decrease in amplitude for higher order harmonics. **B:** The templates generated by a random click train with random widths. Cross sections for three templates are shown below the figure (fundamentals at 420, 750, 1335 Hz).

10 - 4000 Hz) with random phases. Usually many examples of $s(t)$ are generated with different random phases. The final output of the network (T_{ij}) is the average over all these stimulus iterations (N=300 in Fig.3A). Fig.3B shows the average output T_{ij} for a random click train stimulus with random widths (N=300).

The simulations show strongly correlated outputs from channels that are separated exactly by harmonic distances from each other. These strong coincidences form a pattern of multiple diagonals that are spaced at exactly harmonic distances apart. For instance, consider the pattern of strong coincidences for the channel at CF=175 Hz displayed below the coincidence matrix outputs in Fig.3A. The pattern shows prominent peaks at CFs that are successive

multiples of 175 Hz. This pattern is interpreted as the "harmonic template" of the 175 Hz series. Similarly, the templates for all other harmonic series can be found across the diagonals of the network output (e.g., see harmonic series templates for several other fundamentals in Fig.3). Note also that the number of harmonics represented in each template decreases with increasing fundamentals as phase-locking diminishes gradually beyond 1 kHz.

2.4 Final Comments

The mathematical structure of the network and simulations described above are but one example of many variants that can be used. The two key operations are a cochlear-like filtering stage followed by coincidence detection. Relaxing the degree of spectral and temporal sharpening in the model only gradually reduces the clarity of the templates by either diminishing the height of the harmonic peaks or reducing their number. Similarly, replacing the "product" in the coincidence operation (Eq.6) with a squared sum or other "matching" operations does not alter the locations of the harmonic peaks. The reasons behind this robustness are discussed in the next section.

3 Why do the harmonic templates emerge?

In this section, we examine the reasons why the coincidences occur at harmonic distances between the cochlear channels, and specifically discuss the critical role played by three subtle but important factors in the model: the nonlinear transformations following the filtering stage; the rapid phase-shifts of the traveling wave near its resonance; and the spectral resolution of the cochlea.

3.1 Nonlinear transformations of the filter outputs

In the model outputs, the harmonic template lines emerge as a consequence of the strong coincidences between responses of harmonically related cochlear filters. To understand why this is so, consider a sharply-tuned filter bank driven by a broadband noise stimulus. Each filter in this bank produces a phase-locked response waveform that is semi-periodic and reflects predominantly its CF. This is exemplified in Figure 4 by the semi-sinusoidal responses at CF≈250 (Fig.4A). If the filter outputs are not half-wave rectified or otherwise nonlinearly distorted, then the outputs from any such pair of filters will be orthogonal or uncorrelated since each contains Fourier Coefficients only near its CF.

However, the situation is drastically different if the filter responses are half-wave rectified, because this creates "distortion" components and the waveform can be thought of as composed of a fundamental frequency (the CF of the filter) and its harmonics (Fig.4B). Consequently, the rectified waveform from any filter can now partially coincide with outputs of other filters that are at harmonically related CFs. For instance, the rectified waveform from the filter at CF=250 Hz contains harmonics of 250 Hz with gradually decreasing intensity (Fig.4B), and hence may coincide strongly with filter outputs at CF=500, 750, ... Hz. Note that the important role played by the half-wave rectification is not unique to this operation, but is a
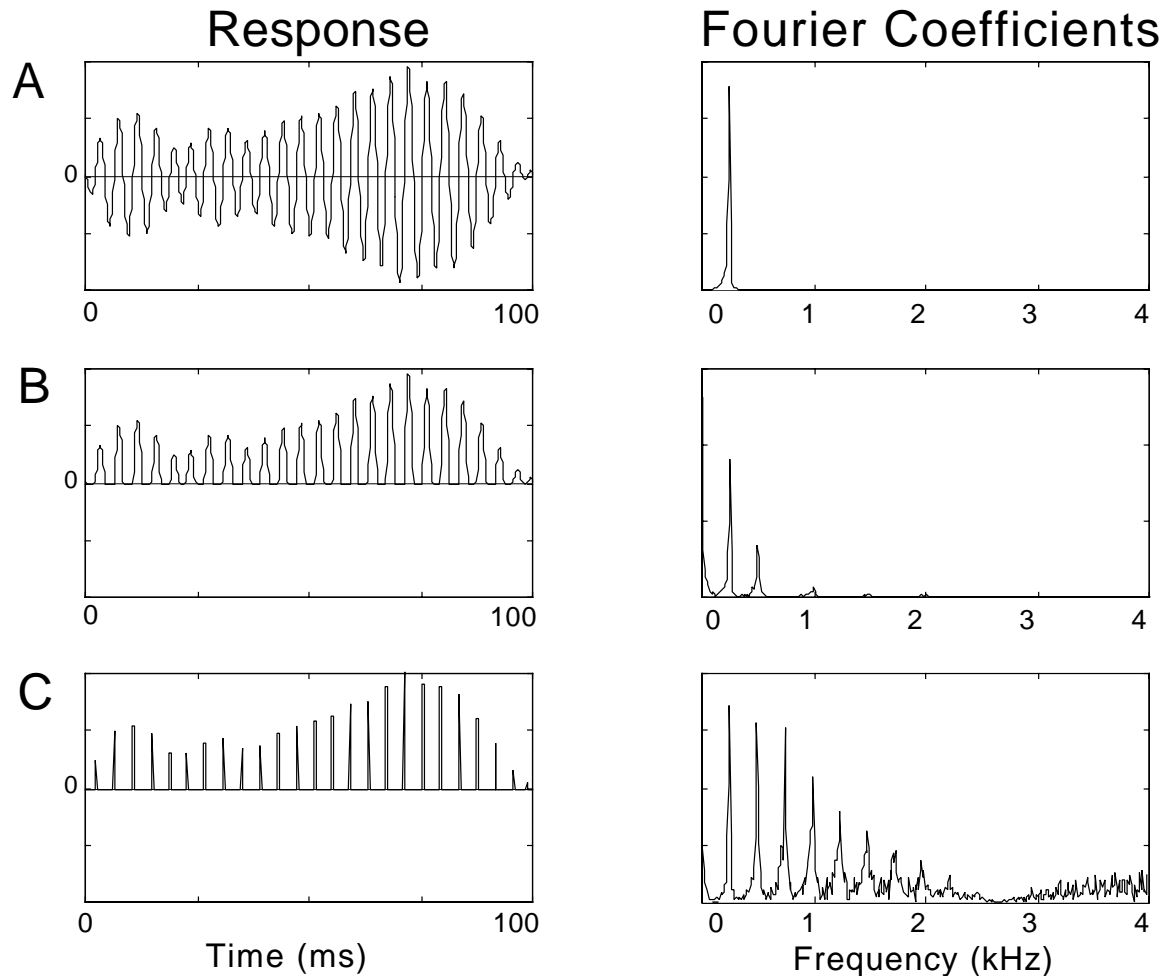
Figure 4: *The effects of nonlinear deformations and temporal sharpening.* The response waveforms within each channel implicitly convey harmonic distortion components with varying strength. *The left column* illustrates response waveforms with increasing nonlinear distortion and synchrony; *the right column* illustrates the Fourier coefficients corresponding to each waveform. The number and amplitude of the harmonic distortion components increase with increasing synchrony and nonlinear deformation of the response waveform. **A:** linear filter response at CF ≈ 250 Hz; **B:** Half-wave rectified response; **C:** The synchronized impulse train corresponding to the 250 Hz response.

common consequence of many instantaneous nonlinear distortions of the filter outputs. For example, similar harmonic coincidence patterns emerge if the filter waveforms are distorted by a saturating nonlinearity, a limited dynamic range, or are converted to a series of synchronized impulses as is done in the model, Eq.4 (Fig.4C).

It is in this context that one can appreciate the role of enhanced temporal synchrony in the model. The synchronization of the filter response waveforms is a highly nonlinear operation that ensures that the impulse train from each filter contains within it the fundamental frequency (at the CF) and, prominently, many of its harmonics (Fig.4C). That is why the pulse train from a filter at CF=250Hz will correlate well with pulse trains produced by filters at harmonically related CFs up to a relatively high order.

3.2   The phase of the cochlear traveling wave

How is it possible that the highly synchronized waveform at a given CF (e.g. 250 Hz) be in just the right phase to coincide with outputs from other CFs (e.g., the response at 500 Hz) ? The answer highlights the role of the cochlear traveling wave, specifically its phase delays, in the formation of the templates.
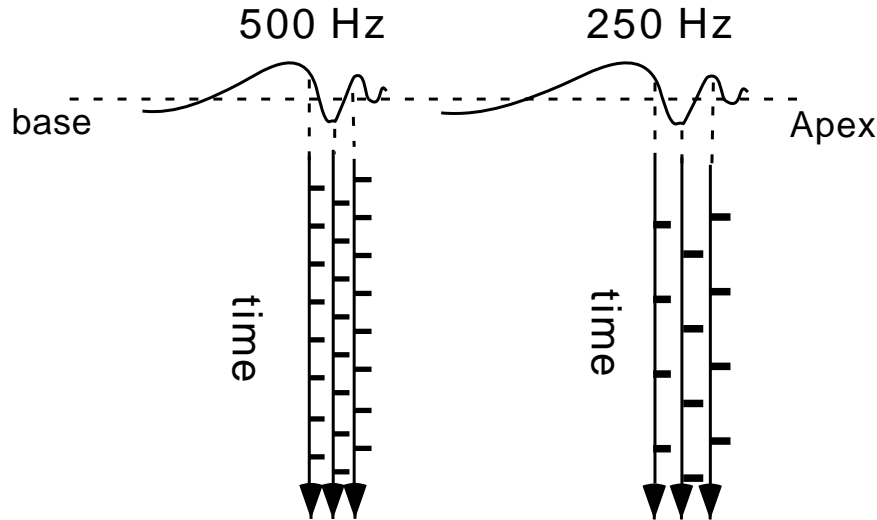

Figure 5: *Traveling wave phase-shifts near the resonance of a traveling wave.* The schematic illustrates that the response patterns near the resonance of the traveling waves can be significantly phase-shifted relative to each other over very short distances.

Figure 5 illustrates the typical features of two traveling waves evoked by two tones, say at 250 and 450 Hz. Near the resonance of each wave (CF=250 and 500 Hz), the travel velocity decreases rapidly, and the wave as a result accumulates phase-delays at an accelerated pace (Lyon and Shamma, 1996; Shamma, 1985a). Consequently, near the CF, one may find responses of widely different (even opposite) phases in closely spaced locations (or channels). That is, each of the CF regions of 250 and 500 Hz contains synchronized responses to these frequencies at various phases, and hence it is likely that at least a pair of channels will coincide and positively correlate. This argument still applies when the stimulus contains many tones (as with the broadband noise) because these phase-delays are characteristic of the cochlear filters and not of the stimulus. Thus, as long as the responses at a given CF are determined by a relatively sharply-tuned cochlear filter, they will necessarily exhibit these rapid phase-shifts near the CF. This was seen earlier in Fig.1B where the synchronized responses to the noise stimulus are similar in adjacent channels except for a rapid phase-delay towards the lower CFs.

3.3   The sharpness of frequency analysis

Cochlear frequency analysis and subsequent spectral sharpening of the filter outputs (by lateral inhibition (Shamma, 1985b); Eq.3) enhance the features of the harmonic templates. This is because sharp filters produce more regular synchronized responses, and hence "purer" harmonic templates compared to broadly tuned filters. This point is illustrated in Figure 6 where

we examine the effect of broadening of the cochlear filters on the synchronized responses to a broadband noise stimulus. Fig.6A shows the synchronized responses (*left plot*) and their corresponding Fourier series coefficients (*right plot*) using our regular filters. Here, the response at each CF contains well defined components at CF and its harmonic distortions as is evident by the well separated Fourier peaks. If the filters are made significantly broader (for instance, by removing the LIN stage), the synchronized responses from each filter become considerably more jittered due to the increased interference within each channel (Fig.6B, *left plot*). This in turn smears considerably the Fourier representation of the higher order distortion harmonics (Fig.6B, *right plot*). Therefore, cochlear frequency selectivity is critical for the formation of the harmonic templates: sharper filters result in clearer high order harmonic peaks in the templates.

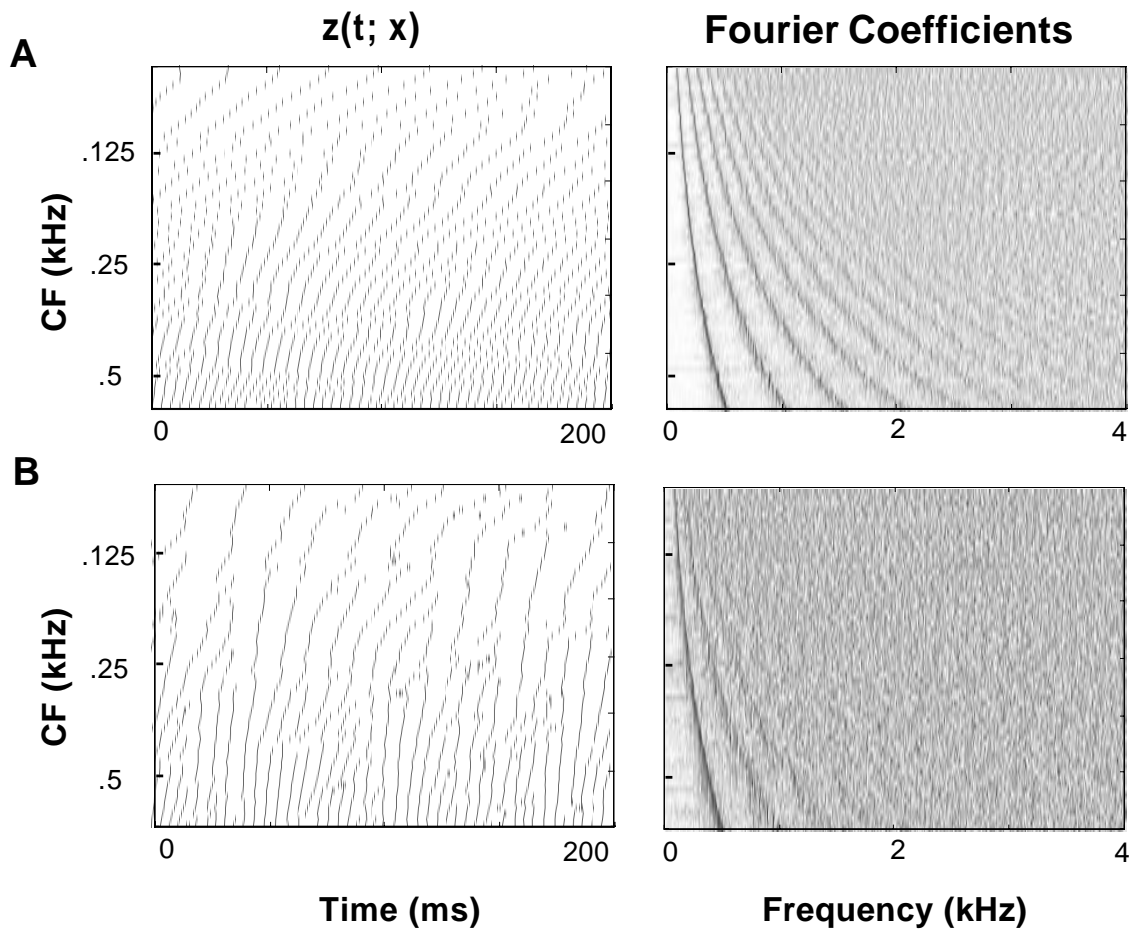Figure 6: *The effects of spectral resolution on the templates.* (*Left*) The synchronized responses to a broadband noise stimulus (as in Fig.1B). (*Right*) The corresponding Fourier series coefficients for all channels (each labeled by its CF along the ordinate). **A:** The responses due to the regular model filters (as in Fig.2). **B:** The responses using broader filters (by removing the LIN stage in Fig.1).

3.4 Summary

The harmonic templates arise from two basic processing stages: cochlear filtering, followed by a matrix of coincidence detectors. The exact shape of these templates, the clarity of their peaks, and the order of their highest harmonics is influenced by the details of these two operations. The following list summarizes these factors:

1. Phase-locking of the filter responses is a critical factor in the template formation. All templates are ultimately derived from the fine-time structure of the filter responses. Thus the gradual loss of phase-locking (or synchrony) to higher frequencies (approximately $> 2 - 3$ kHz) is the reason why they are not represented in the templates, and hence play little or no role in the perception of periodicity pitch. In the model, the degree of phase-locking can be simulated by changing the width of the pulse function p(t): the sharper the pulse, the better is the phase-locking to higher frequencies.

2. Nonlinear transformation of the filter responses is essential in generating the (distortion) harmonics that ultimately form the templates. Half-wave rectification and increasing temporal synchrony are two such transformations. Thus, increasing temporal synchrony, improves the representation of the higher harmonics.

3. High spectral resolution improves the representation of the harmonic peaks in the templates. In the model, the lateral inhibitory stage increases the effective tuning of the filters; removing this stage therefore reduces the number and sharpness of the harmonic peaks in the templates.

4. Phase-delays of the traveling wave provide locally phase-shifted copies of the responses at each CF. While such phase-shifts are typical near the resonance of any bandpass filter, they are especially large in the cochlear filters because of their steep high-frequency roll-off just above the CF. Note that it is important in the model to provide sufficiently dense sampling of the CF axis (number of channels/octave) in order to capture these phase-shifts; the sparser the sampling, the weaker are the coincidence peaks in the templates.

5. The formation of the harmonic templates and their parameters are solely determined by the intrinsic properties of the cochlear filters and coincidences and not of the stimulus. That is, given enough time, the same templates will emerge for any broadband stimulus whether it is noise, harmonic sequences, or impulses.

It is interesting to note that the combined effect of all these factors give rise to templates (Fig.3) with features that resemble closely those suggested by some of the algorithmic implementations of the spectral pitch theories (e.g., as in (Duifhuis et al., 1982; Cohen et al., 1995)). For example, in these implementations, the ideal harmonic templates with their equal amplitude spectral lines (as in (Goldstein, 1973b)) are modified in two ways: Harmonic peaks are gradually decreased in amplitude and/or increased in width with increasing order. These features arise in our templates due to the different factors discussed above.

4 Physiological Correlates of the Model

We discuss here the biological plausibility of the model and the correspondence between its stages and known physiological responses in the early auditory pathways. Some elements of the model have clear biological underpinnings, while others are speculative. For instance, the

frequency analysis, phase-shifts around the CF, half-wave rectification, and the phase-locking of the responses are all well known analogs of basilar membrane and hair cell function.

More speculative, however, is the anatomy and location of the coincidence matrix, and the identity of its immediate input pathway. Since phase-locking up to relatively high frequencies (at least 2 kHz) is necessary at the input of the matrix, this places it at, or prior to, the inferior colliculus. Furthermore, the synchronized responses at the input of the coincidence matrix are highly reminiscent of the responses of the variety of onset cells in the cochlear nucleus. While these observations suggest certain scenarios as depicted in Figure 7, the early auditory system is clearly complex and mysterious enough to support many other variant, or even drastically different substrates.

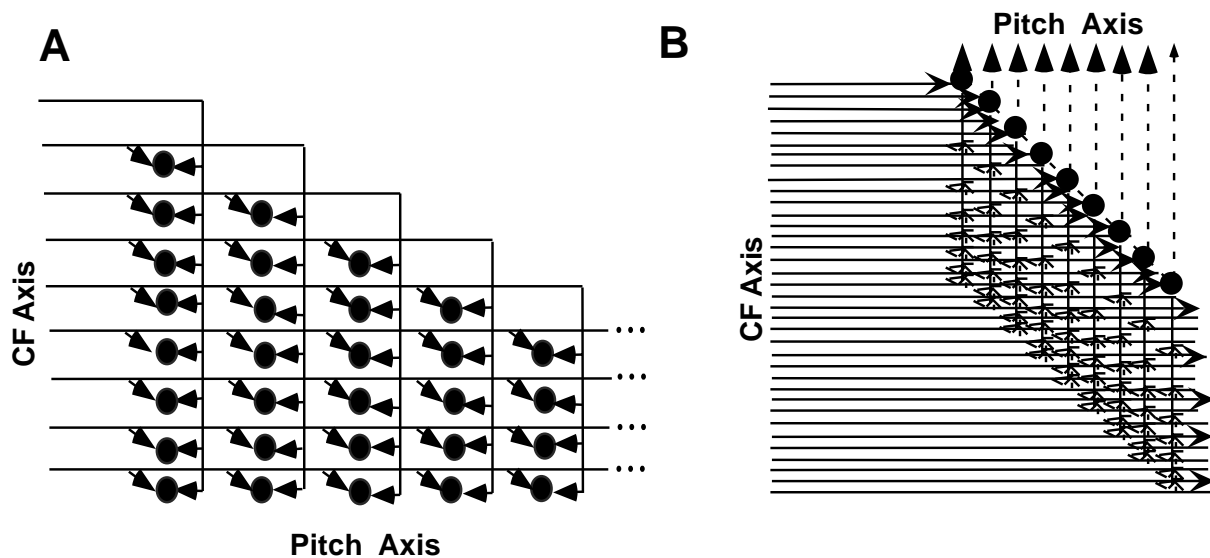## Physiological Realizations of the Coincidence Matrix

Figure 7: *Biological realizations of the coincidence detectors matrix.* **A:** The inputs from the auditory channel array are compared pair-wise by the network of coincidence detectors. Cells in each column have a common CF input from one side, and a progressively increasing CF input from the other side. The templates emerge along the columns (as illustrated earlier in Fig.3) when coincidence detectors at harmonic CF distances are strengthened, while others drop out. **B:** A different realization where pair-wise coincidences are measured and reinforced in the dendrites rather than in separate cells.

Figure 7 shows two examples of possible "neural" realizations of the coincidence matrix. Fig.7A is a more literal interpretation of the mathematical model. The matrix consists of tonotopically organized coincidence detectors, where all cells in a column have the same CF, and are also driven by inputs from systematically higher CFs. Thus, in the fully formed matrix, each cell ends up driven by a pair of CF inputs: one at its primary CF, and another from a higher harmonically related CF. In the alternative realization of Fig.7B, each cell is driven by its primary CF, but it also has an extensive dendritic tree which spans higher CFs. Initially the

dendrites are devoid of synapses. They begin to form during the learning phase at CF locations where the responses correlate well with the primary CF input. In the end, each coincidence cell will be driven by many CF inputs, and hence will appear very broadly tuned. Clearly, a mix of these two scenarios is also possible.

But where is the input pathway to the coincidence matrix? The candidate pathway must be spectrally well resolved and phase-locked as in the auditory-nerve. In the cochlear nucleus, many cell types exhibit the appropriate spectrally and temporally sharp responses, especially the onset and primary-like cells in the low CF regions (Rhode, 1995; Smith and Rhode, 1989; Evans and Zhao, 1998). These cells may project to the coincidence matrix in the limniscus nuclei or the IC. Alternatively, Fig.7B resembles closely the anatomical features of the Octopus cells (the presumed onset-I cells) (Oertel et al., 1990; Palmer et al., 1995), suggesting that they may serve themselves as the coincidence matrix. Unfortunately, most data available at present from various onset cells and other appropriate cell types in the cochlear nucleus are from units with relatively high CFs (> 3 kHz), and hence one cannot be certain of their role in periodicity pitch (Palmer et al., 1995; Rhode, 1995; Evans and Zhao, 1998). For instance, the strong dependence of onset cells (especially onset-I) on the phases of the components a complex tone stimulus observed in high CF cells may not occur in low CF cells (Evans and Zhao, 1998).

Finally, there are numerous pitch phenomena that are closely related to periodicity pitch, and derived exclusively from binaural stimuli (such as the Huggins pitch). These results suggest that the coincidence detectors maybe located at or post binaural convergence nuclei. For instance, it is conceivable that the MSO can serve both its traditional binaural coincidence role (Jeffress, 1948; Shamma et al., 1989), and a monaural coincidence role for the encoding periodicity pitch. Clearly, there is little solid support at present to indicate the existence of such structures in the IC or other central nuclei, and the only definite conclusion that can be made at this time is that much more physiological data are needed to disprove any of these hypotheses.

5 Discussion

We have described a model for how harmonic templates might arise during early development of the auditory system. The model demonstrates that the templates are a natural consequence of basic properties of processing in the early stages of the auditory system. Most important among these properties are cochlear filtering, phase-locked representation of its outputs, enhanced temporal synchrony, and, finally, coincidences across the channel array. We have discussed the contributions of each of these properties to the clarity of the template peaks and the highest harmonic order represented.

An important conclusion from this model is that the harmonic templates are robust and reflect fundamental features of peripheral auditory function. Thus, for the model to work at all, we must have cochlear frequency analysis; we must have rapid traveling wave delays near the wave's resonance; and we must have phase-locking and half-wave rectification on the auditory nerve. Beyond these fundamental features, all other details, such as enhanced temporal synchrony and spectral sharpness, are helpful in improving the templates in a graded fashion.

Another important conclusion is that template formation is largely independent of the

stimulus as long as there is energy available at all frequencies (< 3 kHz) over a period of time. That is, harmonic templates will appear if we had used harmonic sounds, impulses, or any other broadband stimulus provided that all frequencies are represented over the ensemble. However, even if the stimulus energy is not well balanced due, for instance, to partial threshold elevation or a notch in the audiogram, the templates will still arise, but with reduced contributions from these frequencies. For example, if the channel at CF = 400 Hz is removed at the outset (e.g., due to a localized hair cell death at that location), then the model predicts that the 400 Hz template will not be learned, and that this pitch will not be heard from a complex of higher order harmonics (e.g., 800, 1200, 1600 Hz). All other templates will form, but with contribution from the 400 Hz missing. For instance, the 200 Hz template will have all its peaks intact except for the 400 Hz. Note that this prediction is contrary to that obtained from a "temporal" model such as the correlogram (Slaney and Lyon, 1993), where the perceptual contribution to the 400 Hz pitch comes from all CF channels regardless of what is happening at the CF=400 Hz channel.

Finally, the model suggests a simple answer to the question of why harmonic templates postulated in psycho-acoustical studies have a prominent fundamental when natural harmonic sounds (e.g., speech) often have little or no energy at the fundamental. Equivalently, why does a partial set of upper harmonics evoke a pitch at the fundamental and not at any other arbitrary frequency, thus implying that the learned templates must be linked to the fundamental ? The answer is that harmonic templates are formed from exposure to broadband noise, clicks, and other stimuli where all frequencies are available, and not simply from examples of harmonic sounds such as voices which may not have the fundamental.

5.1   Repetition pitch

In the model, several factors conspire to diminish the representation of the high order harmonics ($> 10-15^{th}$), primary among them is the inadequate resolution of the filters. The consequences are easiest to see in the "neural" realization of Fig.7B, where each neuron is driven primarily at its CF, but also on its dendrites from a wide range of other CFs that are harmonically related to its own CF. During the learning phase with the white noise stimulus, broad tuning in the filters near the higher harmonics of the primary CF destroys the temporal regularity of their outputs (Fig. 6). Consequently, few coincidences will occur between these channels and the primary CF channel, and hence synapses will not form.

Humans, however, perceive a clear pitch from tone complexes of unresolved components that is equal to the period of the waveform envelope (repetition pitch). Unlike periodicity pitch, this percept is sensitive to the phase of the components and is weakest when they are in random phase. Clearly, this pitch is not related to any harmonicity in the stimulus, and its properties are dissimilar from those of periodicity pitch. There is, therefore, little reason to assume that this repetition pitch is derived from the harmonic templates, a conclusion also supported psycho-acoustically (Carlyon, 1998). Nevertheless, the need to unify these two pitch percepts in a single mechanism has been the primary motivation for the development of the "temporal" models of pitch alluded to earlier.

Ironically, it is possible to show that a simple scheme based on the coincidence matrix is also capable of measuring repetition pitch. The idea is illustrated in Figure 8, where we assume

the input to the coincidence matrix to be very broadly tuned and synchronized to the envelope of in-phase high frequency tone complexes. Such a pathway describes well the responses of the onset units (especially onset-I) at *high CFs* (Oertel et al., 1990; Rhode, 1995). Figure 8A illustrates such responses to a harmonic series stimulus consisting of the high harmonics (10th -110th harmonics of a 70 Hz fundamental). The last necessary ingredient is a simple monotonic increase in response latency for lower CFs (Fig.8B), similar to that found in the IC and cortex (Langner and Schreiner, 1988; Greenberg, 1997). This relative latency-shift effectively "tilts" the input waves, allowing the coincidence matrix to indicate the repetition period in terms of the distance separating coincident channels in the input array. Thus different repetition periods evoke different coincidence patterns, with faster rates (e.g. 140 Hz) causing coincidences closer to the center diagonal as shown in Fig.8C.
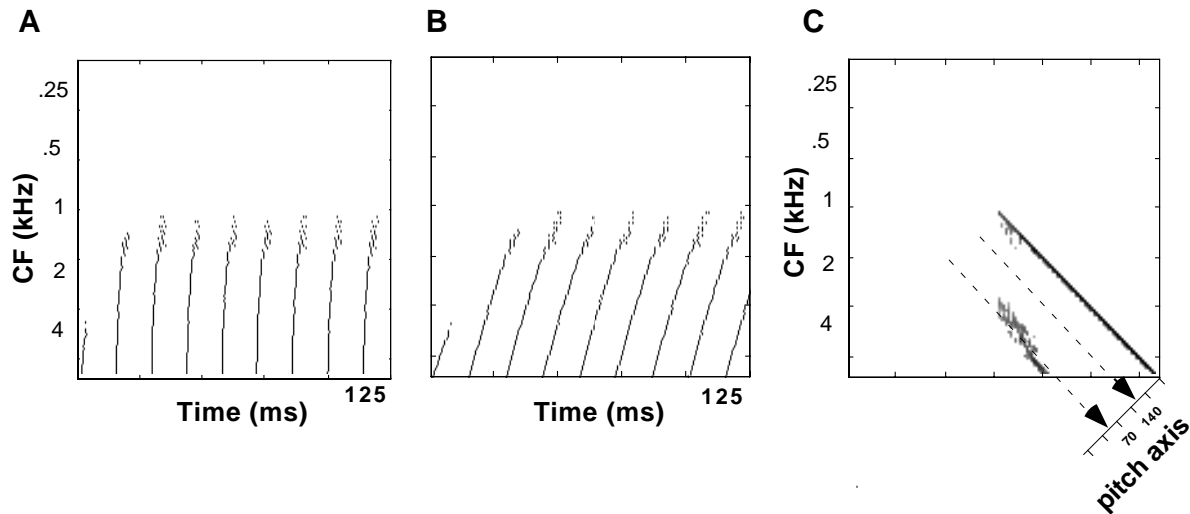

Figure 8: *Measuring repetition pitch with the coincidence matrix.* **A:** The synchronized responses to a stimulus consisting of $10^{th} - 110^{th}$ harmonics of 70 Hz. **B:** The synchronized waves are "tilted" by a gradual increase in latency from high to low CF channels. **C:** The average output of the coincidence matrix shows a strong line of coincidences parallel to the diagonal, and at a distance that reflects the repetition period of the stimulus. Faster rates cause this line to move gradually closer to the main diagonal.

Note that the input can also phase-lock to the periods of *single* low frequency tones (50-500 Hz), hence creating coincidence patterns identical to those evoked by high frequency complexes with the same repetition periods (Fig.8C). This explains the perceptual equivalence between a 10 kHz tone that is amplitude modulated at a 100 Hz, and a single 100 Hz tone. The question remains as to how and where periodicity and repetition pitch are combined and registered relative to each other . The simplest answer is that they are measured independently, and later linked (by coincidence?) to each other (Carlyon, 1998).

5.2   Where to search for physiological evidence

What physiological or anatomical evidence should we look for to confirm the presence of the harmonic templates ? Two sets of data are needed to shed light on the model. The first

concerns the inputs to the coincidence matrix, and the second deals with the coincidence cells themselves. The input pathway must be sharply tuned in its *synchronous responses*. This is an important consideration which is often ignored when reporting on the tuning or iso-intensity response curves of these cochlear units. To establish the relevance of any cells for the encoding of periodicity pitch, it is essential that the units studied receive phase-locked auditory-nerve inputs, and hence must have low CFs (< 3 kHz). It is also best if their tuning properties are measured with reference to their phase-locked, and not their average rate, inputs. For example, an onset cell may appear very broadly tuned due to its relatively high threshold and very limited dynamic range (Rhode, 1995; Evans and Zhao, 1998). However, the unit may be sharply tuned if one considers how well it is synchronized to one of several closely spaced stimulus components. Alternatively, onset cells may constitute the coincidence matrix themselves, and hence receive input from several CFs (e.g. a pair), and appear broadly tuned.

Coincidence detectors, wherever they may reside, should exhibit distinctive response patterns to harmonic series stimuli such as click trains. For instance, coincidence cells as in Fig.7B should be selective to the rate of a click train (tuned at the fundamental of the cell's template); they must also be insensitive to the phase of the harmonics in the stimulus; and finally, they should be broadly tuned, or at least broadly facilitated (Jiang et al., 1996). The response patterns are different if the coincidence cells have pair-wise inputs as in Fig.7A. Here cells should be doubly tuned or facilitated, and must exhibit predictable tunings to multiple click rates in a phase-insensitive manner. None of these response properties have been reliably demonstrated in the IC or lower auditory nuclei, and it remains to be seen if more controlled recordings in the low CF regions can shed light on these questions.

5.3 The Principle of Coincidence Detection

In biologically inspired models of various auditory tasks, it has been common to postulate neural delays so as to affect various correlation operations. These delays are explicit in some algorithms, e.g., in the binaural processing of inter-aural time delays (Jeffress, 1948; Colburn and Durlach, 1978), or in computing the correlograms for pitch (Slaney and Lyon, 1993). In other algorithms, these delays are implicit only within purely temporal operations that must use them, e.g., in the ALSR and dominant frequency algorithms for spectral shape extraction from auditory-nerve responses(Young and Sachs, 1979; Lyon and Shamma, 1996), or in the use of intrinsic oscillations of variable rates for pitch estimation (Langner, 1992). As mentioned earlier, the need for neural delays stems almost entirely from the need to make interval measurements on single channels independent of other channels.

In previous reports, we have demonstrated that simple coincidence measurements of responses across the auditory channels can extract the same kinds of information robustly, without need for functional neural delays. Thus, lateral inhibition across the outputs of the auditory-nerve fiber array (which is essentially a form of coincidence detection) can extract a highly resolved spectrum of a broadband complex stimulus over wide stimulus levels (Shamma, 1985a; Shamma, 1985b). Similarly, a coincidence matrix identical to the one discussed here (Fig.1) for auditory channels from the two ears (the stereausis network (Shamma et al., 1989)) potentially can explain binaural phenomena accounted for by traditional cross-correlation models. The algorithm described in this paper repeats the same theme discussed above. That is, coinci-

dences across the fiber array carry sufficient information to generate the harmonic templates, and hence obviate the need for the neural delay-lines invoked in many of the current pitch models.

To summarize, the need to invoke neural delay-lines stems from a common view of auditory processing as primarily temporal in the sense defined earlier. This limited view is a direct consequence of the experimental difficulty of measuring the distribution of auditory responses across the tonotopic axis, and hence of appreciating the richness and subtlety of the *spatio*temporal cues created by the cochlea. Decoding such cues often requires simple coincidence detection to measure time-differences between channels rather than absolute time intervals within a channel. Temporal models, instead, essentially re-invent the cochlea centrally by demanding additional ordered delay-lines, correlators, and narrowly-tuned filters with microsecond accuracy to detect and measure stimulus parameters.

6   Summary and Conclusions

We presented a biologically plausible model for forming harmonic templates in the early stages of the auditory system *with broadband noise stimulation, and without need for neural delay-lines and other temporal structures.* The model consists of two key operations: a cochlear filtering stage followed by coincidence detection. The cochlear stage provides responses analogous to those seen on the auditory-nerve and cochlear nucleus. The second stage is a matrix of coincidence detectors that compute the long-term average of pair-wise instantaneous correlation (or products) between responses from all CFs across the channels. Model simulations show that for any broadband stimulus, high coincidences occur between cochlear channels that are exactly harmonic distances apart. Accumulating coincidences over time results in the formation of harmonic templates for all fundamental frequencies in the phase-locking frequency range. The model explains the critical role played by such important factors in cochlear function as the nonlinear transformations following the filtering stage, the rapid phase-shifts of the traveling wave near its resonance, and the spectral resolution of the cochlear filters. More specifically, the following items summarize the major findings of the model:

1. Phase-locking of the filter responses is a critical factor in the template formation. All templates are ultimately derived from the fine-time structure of the filter responses. Thus, the gradual loss of phase-locking (or synchrony) to higher frequencies (approximately $> 2 - 3$ kHz) is partially the reason why they are not represented in the templates, and hence play little or no role in the perception of periodicity pitch.

2. Nonlinear transformation of the filter responses is essential in generating the "distortion" harmonics that ultimately form the templates. Half-wave rectification and increasing temporal synchrony are two such transformations.

3. High spectral resolution improves the representation of the harmonic peaks in the templates. Broadening the analysis filters smears the representation of the higher order harmonics considerably.

4. Phase-delays of the traveling wave provide locally phase-shifted copies of the responses at each CF, which in turn insures there is always a pair of channels at harmonic CFs that can be highly correlated regardless of the phase of the stimulus harmonics.

5. The exact nature of the sound stimulus is immaterial for the formation of the harmonic templates and their parameters. Instead they are solely determined by the intrinsic properties of the cochlear filters and coincidences. That is, given enough time, the same templates will emerge for any broadband stimulus whether it is noise, harmonic sequences, or impulses.

Acknowledgments

References

Cariani, P. and Delgutte, B. (1996). Neural correlates of the pitch of complex tones. i: Pitch and pitch salience. *Journal of Neurophysiology*, 76:1698–1716.

Carlyon, R. (1998). Comments on a unitary model of pitch perception. *JASA*, 104:1118–1121.

Clarkson, M. and Rogers, E. (1995). Infants require low-frequency energy to hear the pitch of the missing fundamental. *J Acoust Soc Am*, 98(1):148–54.

Cohen, M., Grossberg, S., and Wyse, L. (1995). A spectral network model of pitch perception. *JASA*, 98:862–879.

Colburn, S. and Durlach, N. (1978). Models of binaural interactions. In Carterette, E. and Friedman, M., editors, *Handbook of Perception*, volume IV. Academic, New York.

de Cheveingne, A. (1998). Cancellation model of pitch perception. *JASA*, 103:1261–1271.

de Cheveingne, A., McAdams, S., and Marin, C. (1995). Concurrent vowel identification. ii: Effect of phase, harmonicity, and task. *JASA*, 101:2848–2856.

Duifhuis, H., Willems, L., and Sluyter, R. (1982). Measurement of pitch in speech: An implementation of goldstein's theory of pitch perception. *JASA*, 71:1568–1580.

Evans, E. and Zhao, W. (1998). Periodicity coding of the fundamental frequency of harmonic complexes. physiological and pharmalogical study of onset units in the ventral cochlear nucleus. In Palmer, A. R., Ress, A., Summerfield, A. Q., and Meddis, R., editors, *Psychophysical and Physiological advances in hearing. Proceedings of the 11th international symposium on hearing*, pages 186–194, London. Whurr Publishers.

Goldstein, J. (1973a). An optimum processor theory for the central formation of pitch of complex tones. *J. Acoust. Soc. Am.*, 54:1496–1516.

Goldstein, J. (1973b). An optimum processor theory for the central formation of the pitch of complex tones. *JASA*, 54:1496–1516.

Greenberg, S. (1997). The significance of cochlear travelling wave for theories of frequency analysis and pitch. In Lewis, E., Long, G., Lyon, R., Narins, P., Steele, C., and Hecht-Poinar, E., editors, *Diversity in Auditory Mechanics*, pages 183–189. World Scientific, Singapore.

Jeffress, A. (1948). A place theory of sound localization. *J. Comp. Physiol. Psych.*, 61:468–486.

Jiang, D., Palmer, A., and Winter, I. (1996). Frequency extent of two-tone facilitation in onset units in the ventral cochlear nucleus. *J. Neurophysiol.*, 75(1):380–395.

Langner, G. (1992). Periodicity coding in the auditory system. *Hearing Res.*, 6:115–142.

Langner, G. and Schreiner, C. (1988). Periodicity coding in the inferior colliculus of the cat. *Journal of Neurophysiology*, 60:1805–1822.

Lickleider, J. (1951). A duplex theory of pitch perception. *Experientia*, 7:128–133.

Lyon, R. and Shamma, S. (1996). Auditory representation of timbre and pitch. In Hawkins, H., McMullen, E. T., Popper, A., and Fay, R., editors, *Auditory Computations*, pages 221–270. Springer Verlag.

Meddis, R. and Hewitt, J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification. *JASA*, 89:2866–2882.

Montgomery, C. and Clarkson, M. (1997). Infants' pitch perception: masking by low- and high-frequency noises. *J Acoust Soc Am*, 102(6):3665–72.

Moore, B. (1989). *An Introduction of the Psychology of Hearing.* Academic Press, London, third edition.

Oertel, D., Wu, S., and Dizack, C. (1990). Morphology and physiology of cells in slice preparation of posteroventral cochlear nucleus of mice. *Journal of Comp. Neurol,*, 295:136–154.

Palmer, A., Winter, I., Jiang, D., and James, N. (1995). Across frequency integration by neurons in the ventral cochlear nucleus. In Manley, J., Klump, G., Kopple, C., Fastl, H., and Oeckinghaus, H., editors, *Advances in Hearing Research*. World Scientific Publishers, Singapore.

Patterson, R. and Holdsworth, J. (1991). A functional model of neural activity patterns and auditory images. In Ainsworth, W. A., editor, *Advances in Speech, Hearing and Language Processing*, volume 3. JAI Press, London.

Plomp, R. (1976). *Aspects of tone sensation.* Academic Press, New York.

Rhode, W. (1995). Interspike intervals as a correlate of periodicity pitch in cat cochlear nucleus. *J. Acoust. Soc. Am.*, 97:2414–2429.

Schreiner, C. and Langner, G. (1988). Periodicity coding in the inferior colliculus of the cat. ii. topographical organization. *J. Neurophysiol.*, 60:1823–1840.

Schreiner, C. and Urbas, J. (1988). Representation of amplitude modulation in the auditory cortex of the cat. i: The anterior field (aaf). *Hear. Res.*, 21:227–241.

Schwartz, D. and Tomlinson, R. (1990). Spectral response patterns of auditory cortex neurons to harmonic complex tones in alert monkey (*macaca mulatta*). *Journal of Neurophysiology*, 64:282–299.

Shamma, S. (1985a). Speech processing in the auditory system: I. representation of speech sounds in the responses of the auditory nerve. *JASA*, 78:1612–1621.

Shamma, S. (1985b). Speech processing in the auditory system: Ii. later inhibition and the central processing of speech evoked activity in the auditory nerve. *JASA*, 78:1622–1632.

Shamma, S., Chadwick, R., Wilbur, J., Morrish, K., and Rinzel, J. (1986). *A biophysical model of cochlear processing: Intensity dependence of pure tone responses* . J. Acoust. Soc. Am., 80(1):133–145.

Shamma, S. and Morrish, K. (1986). Synchrony suppression in complex stimulus responses of a biophysical model of the cochlea. *JASA*, 81:1486–1498.

Shamma, S., Shen, N., and Gopalaswamy, P. (1989). Stereausis: Binaural processing without neural delays. *J. Acoust. Soc. Am.*, 86:989–1006.

Slaney, M. and Lyon, R. (1993). On the importance of time - a temporal representation of sound. In Cooke, M., Beet, S., and Crawford, M., editors, *Visual Representations of Speech Signals*. J. Wiley and Sons, Sussex England.

Smith, P. and Rhode, W. (1989). Structural and functional properties distinguish two types of multipolar cells in the cat ventral cochlear nucleus. *JASA*, 282:595–616.

Summerfield, A. and Assmann, P. (1990). Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of the Acoustical Society of America (JASA)*, 88:680–697.

Terhardt, E. (1979). Calculating virtual pitch. *Hearing Res*, 1:155–182.

Wang, K. and Shamma, S. A. (1994). Self-normalization and noise-robustness in early auditory representations. *IEEE Transactions on Speech and Audio Processing*, pages 421–435.

Wightman, F. (1973). A pattern transformation model of pitch. *JASA*, 54:397–406.

Yang, X., Wang, K., and Shamma, S. A. (1992). Auditory representations of acoustic signals. *IEEE Transactions on Information Theory, Special Issue on Wavelet Transforms and Multiresolution Signal Analysis*, 38(2):824–839.

Young, E. and Sachs, M. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *J Acoust. Soc. Am.*, 66:1381–1403.