On the role of space and time in auditory processing

Shihab Shamma

Unlike visual and tactile stimuli, auditory signals that allow perception of timbre, pitch and localization are temporal. To process these, the auditory nervous system must either possess specialized neural machinery for analyzing temporal input, or transform the initial responses into patterns that are spatially distributed across its sensory epithelium. The former hypothesis, which postulates the existence of structures that facilitate temporal processing, is most popular. However, I argue that the cochlea transforms sound into spatiotemporal response patterns on the auditory nerve and central auditory stages; and that a unified computational framework exists for central auditory, visual and other sensory processing. Specifically, I explain how four fundamental concepts in visual processing play analogous roles in auditory processing.

Shihab Shamma
Center for Acoustic and Auditory Research, Electrical and Computer Engineering Dept, Institute for Systems Research, University of Maryland College Park, College Park, MD 20742, USA.
e-mail: sas@isr.umd.edu

Analogies between auditory and visual perceptions have been sought and discussed for centuries and are reinforced by neurological disorders where the experience of the two senses are closely intermingled[1]. Examination of the current scientific literature concerning sound and image representations, and the functional principles and neural networks underlying their transformations and perception, reveals strongly divergent theoretical views of auditory and visual processing. A primary cause of this seems to be the different nature of the two inputs. Sound is a pressure wave represented by one-dimensional temporal waveform at the eardrum, whereas an image is a two-dimensional spatially distributed pattern of activation on the retina. Consequently, a characteristic aspect of most proposed auditory processing strategies is their 'temporal' nature in which a neural network computes its output by a systematic analysis of the time history of its input signal. Examples of such algorithms include computations of auto- and cross-correlation functions of the time course of a neural response or measurement of its absolute periodicity, relative delays and intervals[2,3]. In contrast, processing in the visual and somatosensory systems is primarily 'spatial' and these networks derive their output primarily from the spatial distribution of the input pattern, for example, edge-enhancement of stationary images in the retina and spatial disparity measurements in stereopsis[4–6].

Such contrasting views of auditory and visual processing have profound implications for the architecture of the neural networks that implement these computations. For instance, in temporal processing, a network must possess an organized range of time delays, which may arise through systematic variations in the morphological features of its neurons, for example, axons or dendrites with regularly changing lengths, diameters or membrane time constants. The network topology is different for spatial processing, emphasizing axonal and dendritic arborizations and precise patterns of interneuronal connectivity. This temporal–spatial distinction leads to the conclusion that profound functional differences must exist between the neural networks that underlie auditory and visual processing, a conclusion that remains largely unsubstantiated. Thus, although anatomical differences are clearly significant early in these pathways (e.g. retina versus cochlear nucleus), they tend to reflect the different physical nature of sound and light, and not necessarily the function of the neural networks and the cues and features they extract. In fact, in more central neural structures, the theoretical views and anatomical studies support the notion of a unified proto-cortical plan for all primary sensory areas[7]. These views are backed up by recent developmental experiments in which optical projections to the primary auditory cortex (AI) of newborn ferrets resulted in the eventual development of classical visual sensitivity and receptive fields in the AI region[8].

'The theoretical views and anatomical studies support the notion of a unified proto-cortical plan for all primary sensory areas.'

In this article, I argue in favor of a strong spatial view of auditory processing and, hence, a unified computational framework for auditory and visual perception. Specifically, I explain how computational algorithms and neural architectures commonly proposed for early vision may also operate in the auditory system to give rise to the most important perceptual attributes of sound – timbre, pitch and location.

In the context of the auditory system, the term spatial refers to the sensory epithelial axis of the cochlea, depicted schematically by the row of transducer hair cells situated along the length of the cochlea (Fig. 1). This axis is also referred to as the 'tonotopic' or spectral axis of the cochlea because of its ordered frequency selectivity. The cochlea separates a complex sound into its constituent tonal components and distributes their responses spatially along its length[9,10] by the distinctive spatial and temporal vibration patterns of its basilar membrane that reflect the frequency of the sound. For instance, vibrations evoked by a single tone appear as travelling waves that propagate down the

http://tics.trends.com 1364-6613/01/$ – see front matter PII: S1364-6613(00)01704-6

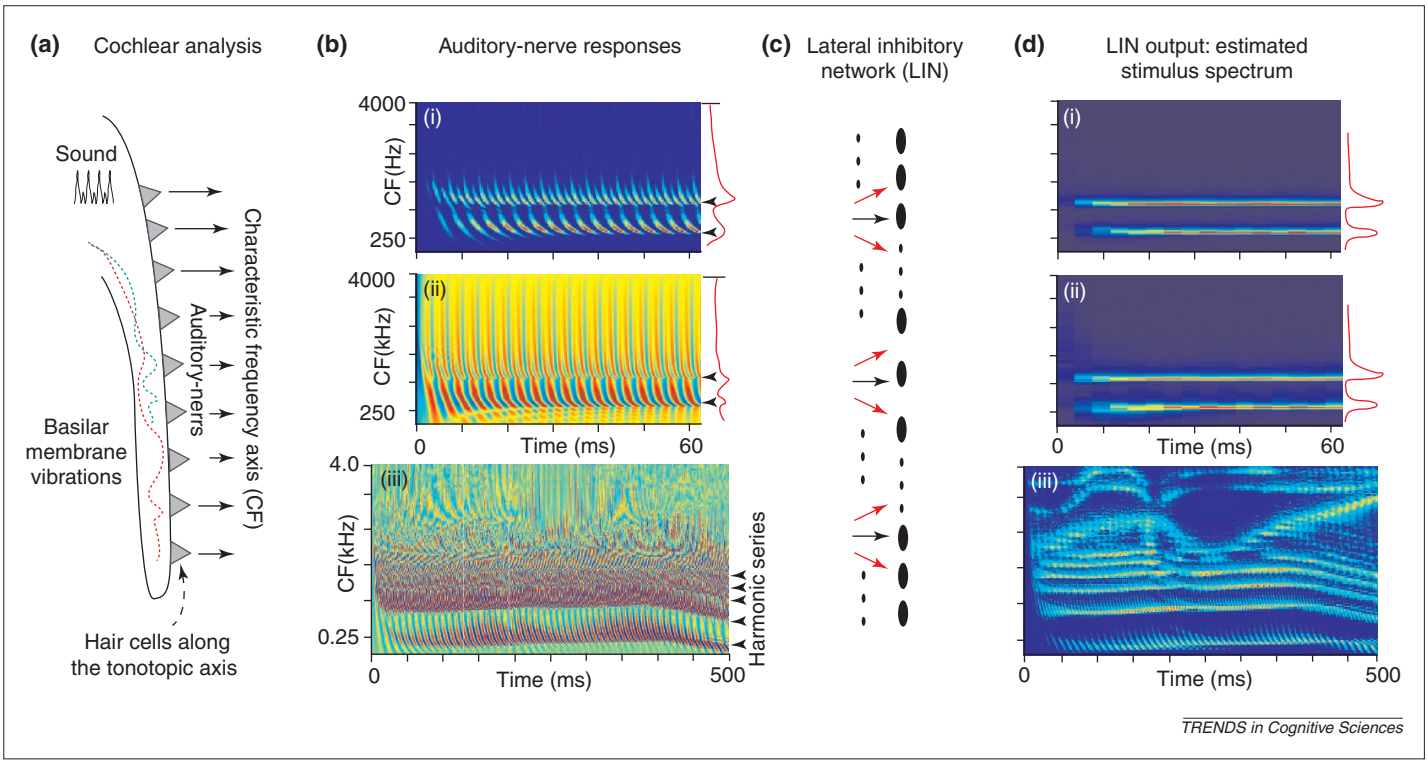

Fig. 1. Schematic model of the early stages in auditory processing. (a) Sound entering the cochlea initiates traveling-wave displacement patterns on the basilar membrane. Vibrations caused by low frequencies propagate and achieve their maximum amplitude further down the cochlea (red) compared with high frequencies (green), which creates the tonotopically ordered (spatial) axis of the auditory system. Vibrations of the basilar membrane are transduced into spatiotemporal responses on the auditory nerve by hair cells. (b) The spatiotemporal response patterns on the auditory nerve caused by a two-tone stimulus of 300 and 600 Hz (i, ii) or the phrase 'right away' (iii) . The ordinate in each panel represents the tonotopic axis, labeled by the characteristic frequency (CF) at each location. Each component in the stimulus initiates a localized traveling wave pattern that ends abruptly, creating a prominent discontinuity near the appropriate CF (arrowheads). In (i) a quiet stimulus does not saturate the nerve responses. The response amplitudes are therefore strongest near the CFs of the two tones, resulting in clear peaks in the average response curve (red). Louder tones (ii) saturate the average response rates and reduce the peaks. Speech (iii) contains many frequency components that are harmonics of a fundamental corresponding to the voice pitch. (c) The lateral inhibitory network (LIN) is a classical feedforward (or feedback) network that detects and enhances the presence of discontinuities in its input pattern. Black and red arrows depict excitatory and inhibitory projections, respectively, that detect phase-locked inputs from the auditory nerve. (d) The LIN detects the discontinuities along the tonotopic axis due to the stimulus frequency components, thus extracting its spectrum. (i, ii) The output is similar, regardless of stimulus level. (iii) The LIN estimates the spectrogram of the speech signal, showing the (pitch) harmonics and the formants of the speech phonemes.

cochlea and reach a maximum amplitude at a particular point before slowing down and decaying rapidly (Fig. 1). The lower the frequency of the tone, the further its waves propagate down the cochlea. Thus, each point along the cochlea has a characteristic frequency (CF) to which it is most responsive. This tonotopic order or axis is an important organizational principle of the entire primary auditory pathway and is preserved through several point-to-point topographic mappings all the way to the auditory cortex.

Hair cells transform displacements of the basilar membrane into intracellular potentials, which are subsequently encoded by spiking patterns of neurons from the auditory nerve that innervate them (Fig. 1). A key feature of auditory-nerve responses is their ability to encode the fine temporal structure of basilar membrane vibrations, by synchronizing (termed 'phase-locking') to them up to fairly high frequencies (4 kHz in mammals). This is a critical range of frequencies for speech and music perception in humans[10]. Within this range, the frequency of a tone is encoded in the auditory nerve both spatially, by its CF location, and temporally, by the periodicity of the responses in the fibers that innervate this CF (Fig. 1b). At much higher frequencies, auditory-nerve responses cease to be phase-locked. Thus, the average firing rate of a fiber reflects the (frequency tuned) amplitude of the traveling wave at its CF, just as retinal rod cells encode only the intensity (not the color) of the light stimulus at that location. The response pattern evoked by a complex sound consisting of several distinct tones is approximately the superposition of the responses initiated by each component. For example, in the responses to the two-tone stimulus (300 and 600 Hz in Fig. 1), each tone synchronizes the responses of a different band of auditory-nerve fibers along the tonotopic axis. Hence responses are not only organized tonotopically (or spatially), but are also phase-locked to the two-tone frequencies (or temporally organized).

In the remainder of this article, I demonstrate how a detailed consideration of the spatiotemporal distribution of auditory-nerve responses leads to the conclusion that the major auditory percepts of timbre, pitch and location can be derived using neural computational principles that are well known in vision processing. The correspondence between these auditory and visual percepts and principles is summarized in Table 1.

Table 1. Correspondences between auditory and visual tasks

Auditory task	Analogous visual task	Common principle
Extracting the profile of the sound spectrum	Extracting the form of an image	Lateral inhibition for edge/peak enhancement
Cortical spectro-temporal profile analysis	Cortical image form analysis; orientation and direction of motion selectivity	Multiscale analysis
Periodicity pitch perception (of the missing fundamental)	Perception of bilateral symmetry; figure–background segregation	Detecting temporal coincidences
Binaural azimuthal localization (stereausis)	Binocular depth perception (stereopsis)	Detecting spatial coincidence

Lateral inhibition: extracting the spectral profile

The spatiotemporal representation of sounds in the frequency range <4 kHz in the auditory nerve has given rise to a range of opinions on how the early stages of the auditory system extracts the acoustic spectrum of the stimulus. At one extreme is the purely spatial representation[11], which views the cochlea as a frequency analyzer that maps the stimulus spectral profile onto the tonotopic axis. A simple, central neural network would estimate this profile from the short-time average firing-rate of auditory-nerve responses as a function of CF (Fig. 1b, upper panel). Experimental support for this hypothesis is equivocal because the representation of important spectral features, such as peaks and valleys in these profiles, deteriorates at moderate-to-high sound levels owing to the limited dynamic range of auditory-nerve responses[10,12] (Fig. 1b, middle panel).

The alternative, extreme view of early auditory processing asserts that the sound spectrum is primarily encoded in the temporal aspect of the responses[13]. To derive the spectrum, neural networks must be able to perform periodicity or time-interval measurements rather than simply averaging the firing rates. For example, measuring response periodicity of all auditory-nerve fibers in Fig. 1b reveals two strongly represented periods, 3.33 and 1.66 ms, which reflect the frequencies of the stimulus components. Saturating the channel responses by increasing the sound level does not affect this outcome and, hence, the representation of this two-tone spectrum is robust. However, for such computations to occur, the underlying neural networks in the auditory system must operate on the time-history of the response and hence exhibit precise series of time-delays that are organized in a regular topology. No such networks have yet been found in the auditory system, or in any other system of the mammalian CNS.

A simple, alternative, robust strategy that circumvents these physiological and anatomical difficulties is based on the principle of lateral inhibition and exploits the detailed spatiotemporal structure of the responses of the auditory nerve. For example, considering two-tone responses (Fig. 1b), the phase-locked responses to each tone reflect two fundamental properties of the underlying traveling waves near their maximum (or point of resonance): an abrupt decay of the amplitude; and a rapid accumulation of phase-lag that appears as a sharp increase in the curvature of the response waves[9,14]. These two features create sharp boundaries (termed edges or discontinuities) between response regions that are phase-locked to different tones. The saliency and CF location of these edges depend on the amplitude and frequency of each tone and, hence, a spectral estimate of the input can be derived by detecting these edges with lateral inhibitory networks such as those found in the retina[4] (Fig. 1c). Such networks may exist in the anteroventral cochlear nucleus, especially involving T-Stellate cells, which exhibit fast inhibitory surrounds and a robust representation of the input spectrum regardless of level[15], that is they mimic the lateral inhibitory network (LIN) outputs in Fig. 1d (Ref. 11).

Relation to vision

Traditionally, lateral inhibition in vision is thought to detect and highlight edges and peaks in the spatial patterns derived from the average firing-rates of the ganglion cells. This is exactly the case in the auditory system for sound stimuli where phase-locking is absent (e.g. frequencies >4 kHz). For lower frequencies, however, the auditory system also expresses edges temporally as borders between response regions that are phase-locked to different frequencies. Note that as far as the LIN is concerned, the temporal structure of the responses is only a 'carrier', or a means of expressing the edges that it detects. This view is fundamentally different from that of the purely temporal algorithms, which seek to derive direct temporal measures (e.g. the absolute frequency of phase-locking) and, hence, require specialized neural delays for their neural implementation.

Multiscale cortical decomposition: spectral profile analysis

The spectral profile extracted at the cochlear nucleus is projected to the auditory cortex via a tonotopically organized pathway through the midbrain and thalamus. However, the details of the representation of the spectral profile in these structures are vague[16]. As with other cortical sensory areas, the auditory cortex is subdivided, with a primary auditory field (AI) in the center, surrounded by a belt of secondary areas that are distinguished both anatomically and physiologically. The responses in AI have been recorded and analyzed for a wide range of acoustic stimuli, natural and artificial, spectrally narrow and broad, species specific and otherwise[17–22]. However, beyond the obvious analogy between the retinotopic and tonotopic maps, few useful insights have been

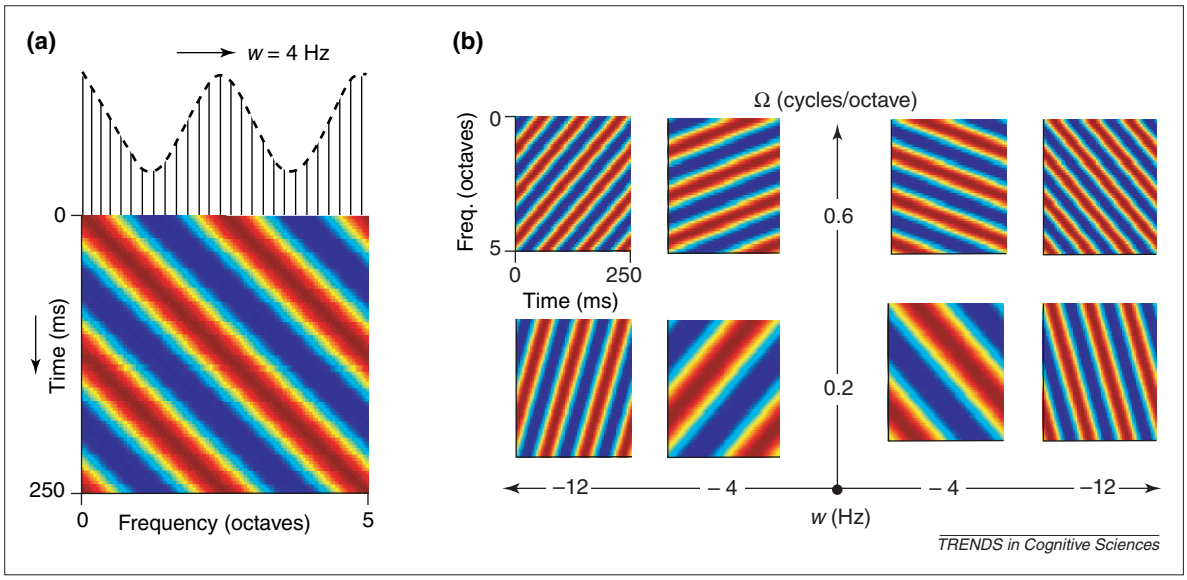

Fig. 2. The dynamic ripple stimulus. (a) The spectral profile of a moving ripple usually consists of many simultaneously presented tones, depicted schematically by vertical lines along the frequency axis. Usually, these are equally spaced along the logarithmic frequency axis and span 5 octaves (e.g. 0.25–8 kHz or 0.5–16 kHz). The sinusoidal spectral profile is depicted by the dashed curve. The spectrogram of one ripple profile is shown in the bottom panel [spectral peak density (Ω) = 0.4 cycles octave^{-1}; peak drift speed (w) = 8 Hz]. (b) Examples of spectrograms of ripple profiles with various spectral densities and velocities (in both directions), ranging from w = 4–12 Hz in two directions, and Ω = 0.2–0.6 cycle octave^{-1}. Ripples over a wider range of parameters are used to measure the responses of a unit.

derived from the comparisons with image analysis in the visual cortex, where significant organizational features occur, such as ocular dominance and orientation columns[23]. Part of the difficulty stems from the fact that orientation and the associated stimuli of oriented bars or edges are intuitively two-dimensional constructs with no obvious analogs in a one-dimensional pattern such as the auditory spectrum.

An alternative approach is to consider responses to the sinusoidal luminance grating, a stimulus that has provided a systematic and mathematically more accessible framework for investigations into organization of the visual cortical for many decades[24]. Gratings have been used to distinguish between simple (linear) and complex (nonlinear) cells and to measure the receptive field of simple cells from their transfer functions to the grating[24]. Furthermore, the properties of a unit's response to different parameters of the grating (e.g. its selectivity to a spatial frequency or phase) could be directly related to the bandwidth, directional selectivity and orientation of its two-dimensional receptive field. Consequently, response maps, such as the orientation columns generated using the oriented bar stimulus, can be characterized equivalently in terms of parameters of the grating.

To relate these findings to the auditory cortex, I first review a series of experiments by two groups, using 'ripples', the acoustic analog of visual gratings[25–29] (Fig. 2). These noise-like broadband stimuli typically consist of hundreds of densely packed tones that are equally spaced along the logarithmic frequency axis. A key feature of the ripple is its sinusoidal spectral envelope, which is created by adjusting the amplitudes of the tones (Fig. 2a). By analogy to a luminance grating, the profile of the ripple is fully characterized by four parameters: overall level; contrast ratio (or the amplitude of the sinusoidal envelope); spectral peak density (or spacing between the peaks, Ω) in units of cycles octave^{-1}; and the constant leftward or rightward peak drift speed in cycles s^{-1} (or Hz). Several ripples with different parameters are shown in Fig. 2b.

AI cells respond well to ripples and are usually selective to a narrow range of ripple parameters[27] (Fig. 3). This selectivity reflects details of the shape of the spectro-temporal response field (STRF) of the cell (the analog of a one-dimensional dynamic receptive field of a visual cell). For instance, cells preferring low ripple-peak-densities are usually more broadly tuned around their CF's than cells that prefer high densities. Similarly, the selectivity to a particular ripple phase is a strong indicator of the asymmetry of the inhibition around the CF. By compiling a complete description of the responses of a unit to all ripple densities and velocities it is possible to compute the STRF and, hence, characterize both the cells' spectral, as well as dynamic, response selectivity[27–29]. If the responses to the ripples are linear with respect to the ripple envelope, the STRF can be considered a complete descriptor of the response properties of the cell and can be used to predict the responses to novel complex-ripple stimuli[29].

STRFs recorded in AI (such as that shown in Fig. 3d) have a wide range of spectral bandwidths and temporal dynamics and, hence, indirectly support the existence of different response maps mentioned earlier[18–20]. Furthermore, the shape of a

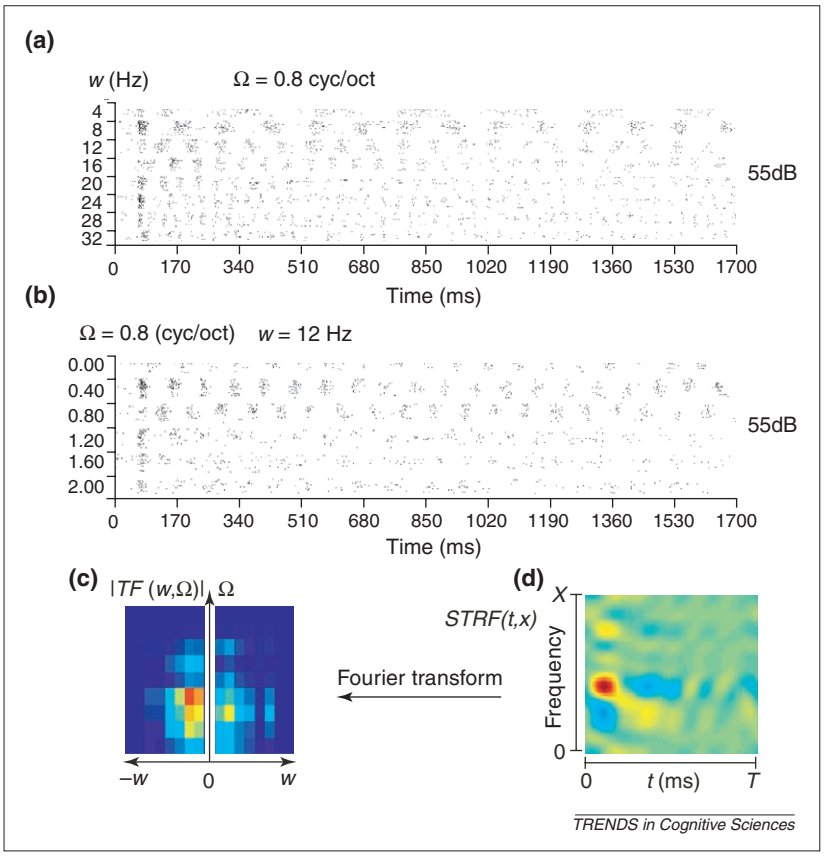

Fig. 3. Constructing the STRF from ripple responses. Raster responses to ripples of different spectral densities and velocities: (a) as a function of peak drift speed (w) at a fixed ripple density (Ω) of 0.8 cycles octave^{-1}, and (b) as a function of ripple density, Ω at a fixed drift speed of 12 Hz. For a complete measurement of all ripples, this test is repeated at all w,Ω combinations to construct the transfer function [TF, or T(w,Ω)] of a cell. The cell depicted responds best to a ripple of $\Omega = 0.4$ cycles octave^{-1} and $w = 8$ Hz. (c) The full two-dimensional TF. (d) The spectrotemporal response field (STRF). TF is computed by constructing period histograms of the raster responses in (a) and (b). The amplitude and phase of the best fit to the histogram at each TF are estimated and plotted as one complex point in the transfer function. Only the magnitude of the TF is shown in the left panel. The STRF is then computed by an inverse-Fourier transform of TF. The STRF is effectively a two-dimensional (spectrotemporal) impulse response that captures the linear response properties of the cell.

unit's STRF remains unchanged whether it is measured with one ripple at a time or multiple ripples simultaneously, for example, using the reverse-correlation method[30]. This affirms the linearity of the ripple responses and the partial separability of the temporal and spectral dimensions of the STRF. It remains to be seen, however, how the extensive variety of STRFs arises and the relationship between them and the morphology and connectivity of different cell types in AI (Ref. 31).

Relation to vision

The responses of cells in the primary visual cortex (VI) to luminance gratings are similar to those described here in AI (see Ref. 24 for a review). For instance, the transfer function of a VI cell is tuned around a specific grating frequency (usually called 'spatial frequency') and its inverse transform predicts the receptive field of the cell measured by impulse-like stimuli as light dots. Thus, just as in AI, responses of the visual cortex have a substantial linear component. The dynamics of the responses in AI and VI regions to moving ripples or gratings and perception of these stimuli are also comparable as they are tuned around velocities of the order of 10 Hz (Fig. 3 and Refs 24,27).

It is also possible to describe a simple, direct relationship between orientation selectivity in VI and ripple selectivity in AI. Visual luminance gratings are spatially two-dimensional and can be uniquely parametrized (within a quadrant) by a combination of vertical and horizontal spatial frequencies (Ω_x, Ω_y). Therefore, VI cells (simple or complex) tuned to specific grating orientations are implicitly tuned to specific combinations of spatial frequencies (e.g. a 45° orientation corresponds to $\Omega_x = \Omega_y$). If one of these two axes is ignored, then orientation selectivity reduces to selectivity to spatial gratings on one dimension. This is analogous to ripple selectivity in AI. Thus, apart from the dimensionality of the input signal, the mechanisms giving rise to orientation selectivity in VI might be identical to those in AI.

The similarity of auditory and visual principles of cortical processing is consistent with conclusions from studies into the generation of the neocortex and subsequent division into distinct areas[7]. Particularly relevant here are experiments by Sur and colleagues in newborn ferrets, in which visual inputs from the optic nerve are induced to project to the auditory cortex through the medial geniculate body[8]. In such adult animals, AI cells possess many of the response characteristics typical of the normal primary visual cortex, such as orientation selectivity. These and other manipulations, such as the transplanting of pieces of fetal neocortex to different positions, point to the homogeneity of the neocortex during early stages of development and the importance of subsequent influences, especially through afferent inputs, in differentiating the adult neocortical areas[7].

Temporal coincidence: estimating periodicity pitch

Pitch is a fundamental percept of sound that is critical to our appreciation of prosody of speech and the melody in music, and in organizing the acoustic environment into different sources[32,33]. Pitch refers to many distinct percepts that are illustrated in Fig. 4. These include: (1) spectrally pitch evoked by a single tone; (2) periodicity pitch (also known as virtual and missing fundamental pitch) evoked by harmonic tone complexes that are spectrally resolved by the cochlea[34]; and (3) residue pitch associated with unresolved harmonics of a common fundamental, very slow click trains, the envelope of amplitude-modulated noise and tones[35]. Periodicity pitch is the most important of these and is associated with musical intervals, melodies, 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

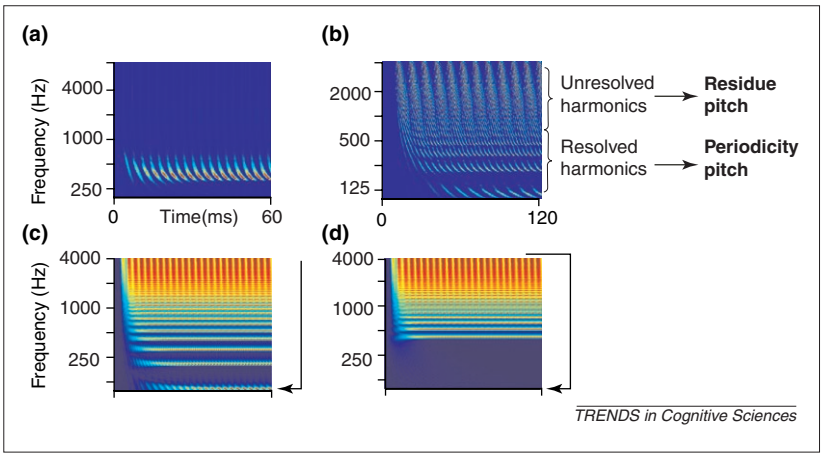

Fig. 4. The representation of stimulus periodicity of a single tone versus a harmonic complex. (a) The auditory nerve responses to a single 300 Hz tone. The phase-locked responses display the periodicity of the tone and the pitch percept heard is that of a single tone at 300 Hz. (b) A harmonic complex of 125 Hz fundamental. The low-order harmonics are well resolved along the tonotopic axis in that each excites an appropriately phase-locked pattern near its characteristic frequency (CF), as with a single tone. These harmonics evoke a pitch sensation at the fundamental frequency of the series (i.e. at 125 Hz). The high order harmonics (>8th harmonic or 1 kHz) are unresolved and they produce a pattern that 'beats' at the difference frequency of 125 Hz. These harmonics evoke a weaker pitch (called the residue pitch) at the beating frequency (125 Hz, in this case). (c,d) The lateral inhibitory network (LIN) outputs of two harmonic series of a 125 Hz fundamental. (c) This harmonic complex contains all 40 harmonics and evokes a strong pitch at 125 Hz. (d) This harmonic complex lacks the lowest three harmonics (125, 250 and 375 Hz). Nevertheless, it still evokes a strong pitch at the 'missing fundamental' frequency of 125 Hz.

other mammals and birds[3,34]. It is most salient when evoked by harmonically related tone complexes that are at least partially resolved spectrally; the pitch heard is normally that of the fundamental frequency of these harmonics regardless of the energy in that fundamental component; the pitch is roughly in the range 50–2000 Hz. The most effective, or dominant, are the low order 2nd–5th harmonics; the salience of the pitch increases in proportion to the number of resolved stimulus harmonics. Multiple pitches are often perceived if there are only a few harmonics in the complex or if the tones form an inharmonic sequence.

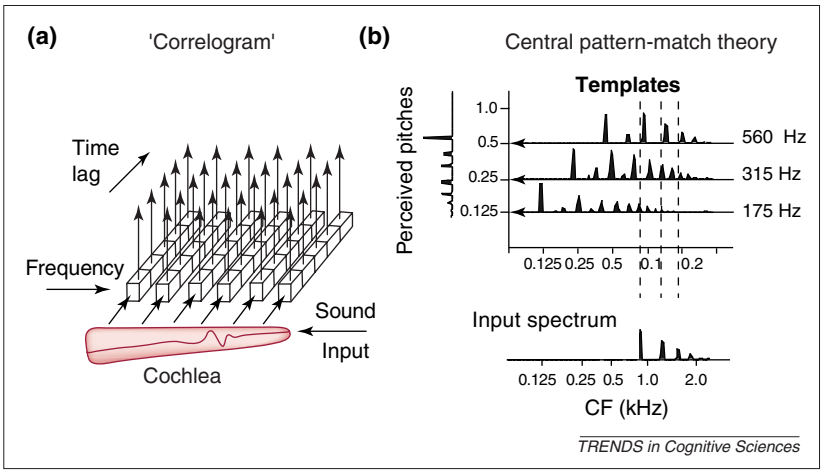

Fig. 5. Two algorithms for extracting periodicity pitch. (a) Schematic illustration of an autocorrelogram implementation (from Ref. 36). This presumes the existence of organized delay lines that autocorrelate the responses from each auditory nerve channel (or fiber) prior to computing the pitch. (b) Schematic illustration of a template-matching algorithm (as described in Ref. 40). This is spatial (spectral) in character and presumes the existence of harmonic templates in the brain that are matched to and measure the pitch of incoming spectra.

Hypotheses proposed for the encoding of periodicity pitch in the auditory system again fall into two major categories: temporal and spectral. A temporal hypothesis is illustrated in Fig. 5a, where it is assumed that the brain estimates the periodicity of the response waveform in each auditory nerve fiber by autocorrelation[36–39] or other related variants of temporal cues and operations[3,37]. The results are then combined from across all fibers to get the final estimate without reference to an ordered spatial tonotopic axis[9]. The spectral (or spatial) hypothesis is radically different and takes as its starting point the incoming spectral profile (defined along the tonotopic axis). Using the specific algorithm in Fig. 5b, the pattern is compared with internally stored spectral templates consisting of the harmonic series of all possible fundamentals to find the closest match and, hence, the perceived pitch[40,41]. Both types of algorithms are successful in explaining and predicting the pitches of complex tones. Their main shortcoming is the lack of convincing biological evidence for the existence of the necessary temporal structures, such as the delay lines or harmonic templates, and how they could be generated. One cannot assume that the templates are 'learned' from frequent exposure to speech and natural sounds early in life because recent evidence suggests that infants are born with an innate sense of periodicity pitch[42].

'The major auditory percepts (...) can be derived using neural computational principles that are well known in vision processing.'

The mystery of how harmonic templates are formed can be largely resolved by examining carefully the spatiotemporal distribution of responses on the auditory nerve to broadband sounds. Specifically, it can be demonstrated that any broadband stimulus, including noise and random click trains, can give rise to the harmonic templates, without the need for delay lines, oscillators or other neural temporal structures[43]. The proposed mechanism consists of two key stages (Fig. 6). The first stage is early auditory filtering (as in Fig. 1a) coupled with temporal enhancement (i.e. by spiking only at the peaks of the input waveform) to produce more highly synchronized response waveforms, such as those seen commonly in Onset 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 CF's across the input array (Fig. 6b). Simulations show that, for any broadband stimulus, a degree of high coincidence occurs among cochlear channels that are spaced precisely at harmonic intervals. Accumulating coincidences over time results in the formation of harmonic templates for all fundamental frequencies

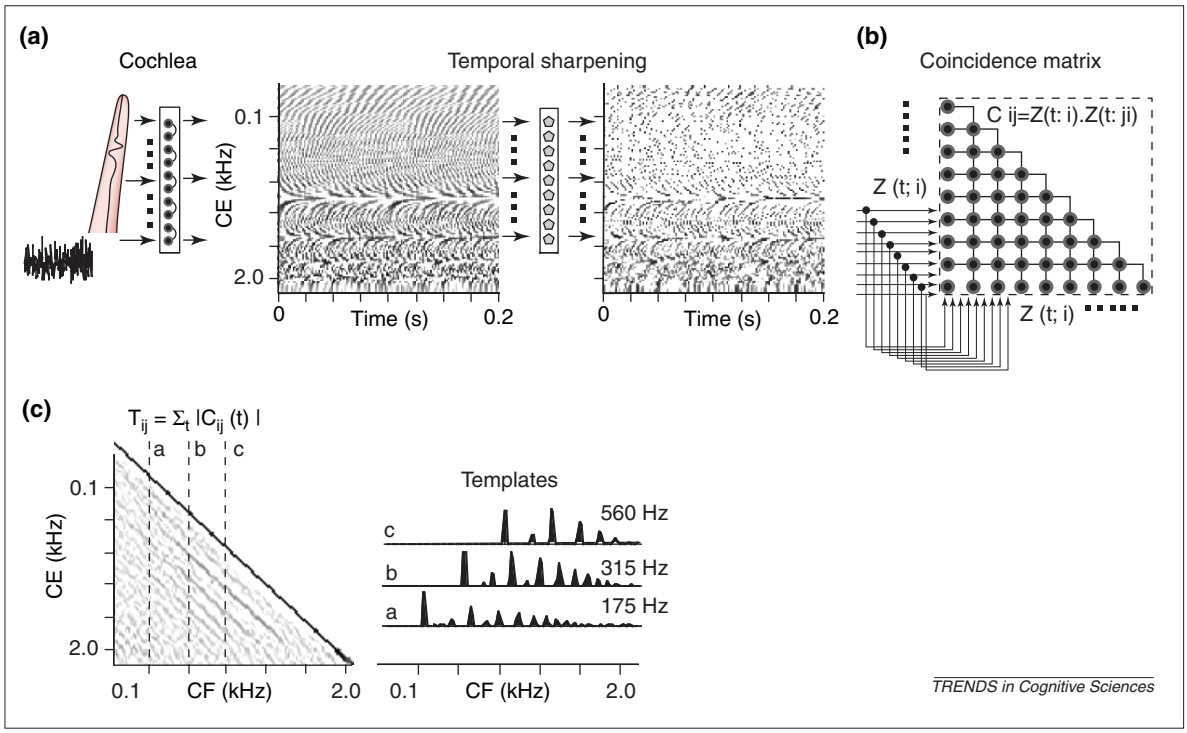

Fig. 6. Schematic model of early auditory stages involved in the formation of harmonic templates[43]. (a) Sound is analyzed by the cochlea and lateral inhibition (as in Fig.1) and is then temporally sharpened. (b) The final stage is a matrix of coincidence detectors that compute the instantaneous product of responses from all pairs of channels across the array. (c) The output of the coincidence matrix integrated over several iterations. Harmonic templates emerge as regions of high coincidence that run parallel to the main diagonal of the coincidence matrix and are exactly spaced at harmonically related characteristic frequency (CF) distances. Three templates are shown individually by the cross sections (fundamentals at 175, 315 and 560 Hz). For each, the pattern has prominent peaks at harmonically related CFs that gradually decrease in amplitude for higher-order harmonics.

in the phase-locking frequency range[43] (Fig. 6c). This model illustrates once again that the auditory system can make use of relatively common coincidence detection mechanisms across spatially distributed inputs to extract highly precise temporal intervals and correlations without need for tapped delay lines and oscillators.

Relation to vision

Relating pitch to visual perception is both intuitive and difficult. It is intuitive because musical melodies evoke a wide range of emotions that are commonly expressed in colors and more formally articulated as a correspondence between musical composition and painting[44]. However, a profound, objective relationship between pitch and vision must be based on similar underlying neurophysiological mechanisms. Delineating such an analogy has been difficult because of the strongly temporal flavor of most proposed pitch extraction mechanisms, which suggests that sensation of periodicity pitch is a percept and process unique to the auditory system.

So what kind of visual percepts could be involved if we imagine that the topographically ordered (one-dimensional) optic nerve projects to a matrix of coincidence detectors as described in Fig. 6? If the outputs of the coincidence detectors are summed diagonally across the matrix, then the network computes an instantaneous spatial cross-correlation between its input pattern and its reverse. Thus, large outputs result if the input image is bilaterally symmetric. They also occur where input regions are temporally synchronized or are comodulated in intensity, as would be the case for the edges of a moving figure against a static background. Detection of bilaterally symmetric and temporally synchronized cues serve a similar functional role in vision as periodicity pitch does in audition, namely to group common harmonics into unitary percept, hence segregating and organizing the perceptual scene into distinct objects that differ in pitch[32,33]. Similarly, in vision, bilateral symmetry is often a property of living (and many inanimate) objects in our world, and provides a powerful cue to segregate and identify different objects in a scene. Figure–surround opposition based on direction of motion cues also has a similar role.

Spatial coincidence: binaural localization

In binaural sound processing the central auditory system compares the signals impinging on the two ears, detecting and utilizing various imbalances (e.g. sound level, time of arrival and phase) to perform such perceptual tasks as sound localization in space and signal-to-noise enhancement[45]. In this sense, binaural hearing is analogous to binocular vision in endowing perception with an extra spatial dimension based primarily on disparity measures in the stimulus projection upon the sensory organs. Numerous computational models have been

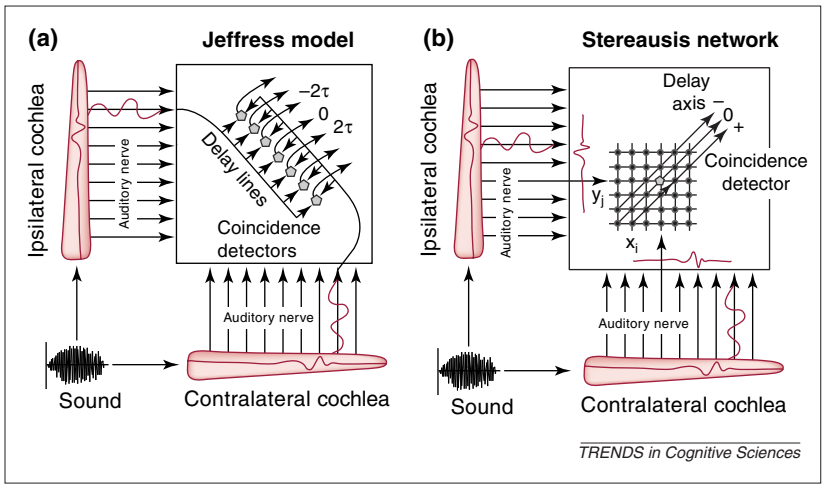

Fig. 7. Binaural processing of interaural time differences. (a) The Jeffress delay-line model for the detection and encoding of interaural delays. Matched-characteristic frequency inputs from the two ears project through delay lines to a series of coincidence detectors whose integrated outputs effectively cross correlate the two inputs. Therefore, the location of the coincidence detector with the maximal output (or peak of the correlation function) is interpreted as the interaural time difference and, hence, the azimuthal location of the sound source. (b) The stereausis network[49] computes the instantaneous spatial cross-correlation between the simultaneous input patterns from the two ears. It consists of a matrix of coincidence detectors that receives direct projections from the cochlear nuclei. All inputs are phase-locked, and hence snap shots of the traveling wave (see Fig. 8) are relayed faithfully to the coincidence detectors. At a given instant, the projected patterns are compared (e.g. multiplied) against each other with either zero shift (along the main diagonal) or with progressively larger shifts along the secondary diagonals.

proposed to account for these phenomena in vision[6] and audition[45,46]. These models have the same spatial–temporal dichotomy discussed earlier for the monaural percepts of pitch and timbre. For instance, in vision, stereopsis algorithms usually detect and process spatial disparities between coincident images from the two retinas. Binaural models, specifically those for the processing of inter-aural time disparities (or time delays between the two ears), derive azimuthal location based on measurement of phase-shifts between the responses from corresponding CFs (or cochlear locations) through explicit time-domain operations.

An important example is the Jeffress model[47] (Fig. 7a), which postulates the existence of an organized array of neural delays to facilitate the computation of cross-correlation between the ipsilateral and contralateral cochlear outputs. This

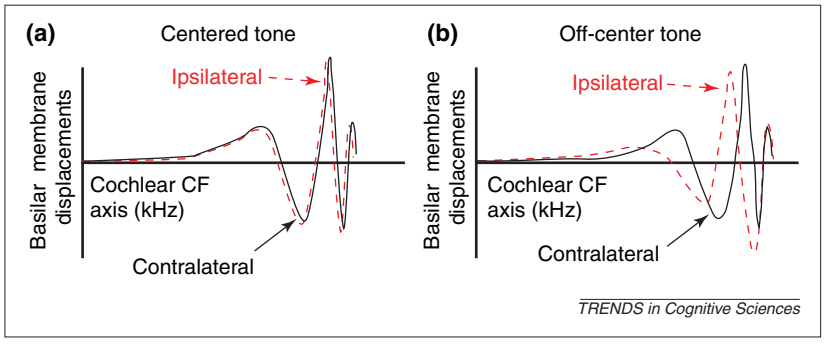

Fig. 8. Snapshots of the basilar membrane traveling waves on the two ears. Traveling waves are evoked by a single tone from a source located on the midline (a) or to one side of the midline (b). The interaural delay in the latter case causes a slight phase shift in the waves relative to each other, thus creating a spatial disparity between the two patterns (contralaterally delayed in this case) that can be exploited to estimate the location of the source.

and similar correlation-based models account well for many psychophysical observations within a convenient mathematical framework. Hence they indirectly support the notion of organized neural delay lines, despite the lack of unambiguous physiological and anatomical evidence of their existence, especially in mammals[48].

A fundamentally different way of estimating the interaural-time-difference is by the proportional spatial disparity that occurs between the simultaneously evoked traveling waves on the two ears. This disparity can be estimated by a spatial cross-correlation of the responses from the two ears using the network of coincidence detectors in Fig. 7b. Figure 8 illustrates the nature of the spatial disparity between traveling waves evoked by a centered tone (Fig. 8a) and an off-center tone (Fig. 8b). The phase-locked responses on the auditory nerve transmit these views of the basilar membrane to the coincidence network. For a centered tone (Fig. 8a), the identical inputs cause maximal activation along the center diagonal of the coincidence matrix. For a binaurally delayed tone, the input patterns appear spatially shifted (Fig. 8b) and maximal activation shifts off the diagonal. Thus, the binaural processing of interaural-time-differences can be reduced to purely spatial operations. Many other possible inequalities in binaural inputs, for instance in their envelopes, degree of correlation, amplitudes and bandwidths, can be similarly detected and consistently represented via the spatial disparities between the resulting traveling waves[49].

Relation to vision

The above coincidence algorithm is fundamentally identical to those proposed to solve the stereopsis problem in vision[6], where spatial disparities between the binocular images play an analogous role to interaural differences in hearing by endowing the percept with an additional perceptual (spatial) axis. Once again, however, the auditory and visual systems differ in the 'means' for expressing the spatial disparity cues, with temporal phase-locking in the auditory system fundamentally serving as the carrier of spatial cues to the CNS. Without phase-locking, the central coincidence processor is blind to the structure of the traveling waves and, hence, cannot detect the relative disparity cues. It is in this light that one may interpret the significance of the temporal specialization observed in the early auditory pathways and nuclei, such as the rapid, extraordinarily large synapses of the bushy cells of the anteroventral cochlear nucleus.

Conclusion

The perception of sound involves a complex array of attributes, ranging from the sensation of timbre and pitch to the localization and fusion of sound sources. Computational strategies proposed to describe these

phenomena have emphasized temporal cues and features in the representation of sound in the auditory system. They have also postulated temporal algorithms, such as correlations and absolute period measurements, and utilized delay-lines, intrinsic oscillators and other temporal structures to extract them. These arguments have led to the conclusion that auditory and visual processing must be quite different, as are the neural networks that underlie them.

I argue here that simple coincidence measurements of responses across the tonotopically ordered auditory channels could extract the same kinds of temporal information robustly, without need for neural delays and associated structures. The key idea is that the basilar membrane acts as the universal, effective, mechanical, delay-line of the auditory system. Through its traveling wave and related frequency analysis, the basilar membrane transforms acoustic temporal cues into spatial cues that can be subsequently analyzed by spatially distributed neural networks much like those found in the visual system.

Acknowledgement
This work has been supported in part by a grant from the Office of Naval Research under the ODDR\&E MURI97 Program to the Center for Auditory and Acoustic Research.

References

1 Cytowic, R. (1998) *The Man Who Tasted Shapes*, Bradford Books
2 Hawkins, H. *et al*., eds (1996) *Auditory Computations*, Springer-Verlag
3 Moore, B. (1989) *An Introduction of the Psychology of Hearing*, Academic Press
4 Hartline, H. (1974) *Studies on Excitation and Inhibition in the Retina* (Ratliff, E., ed.), Rockefeller University Press
5 Poggio, G. (1984) Processing of stereoscopic information in primate visual cortex. In *Dynamic Aspects of Neocortical Function* (Edelman, G. *et al*., eds), pp. 613–636, John Wiley & Sons
6 Marr, D. and Poggio, T. (1979) A computational theory of human stereo vision. *Proc. R. Soc. London Ser. B* 204, 301–328
7 Dennis, D. and O'Leary, M. (1989) Do cortical areas emerge from a protocortex? *Trends Neurosci.* 12, 400–406
8 Sur, M. *et al*. (1988) Experimentally induced visual projections into auditory thalamus and cortex. *Science* 242, 1437–1441
9 Lyon, R. and Shamma, S. (1996) Auditory representation of timbre and pitch. In *Auditory Computations* (Hawkins, H. *et al.,* eds), pp. 221–270, Springer-Verlag
10 Pickles, J.O. (1988) *An Introduction to the Physiology of Hearing*, Academic Press
11 Shamma, S. (1985) Speech processing in the auditory system: II. Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *J. Acoust. Soc. Am.* 78, 1622–1632
12 Sachs, M.B. and Young, E.D. (1979) Encoding of steady state vowels in the auditory-nerve: representation in terms of discharge rate. *J. Acoust. Soc. Am.* 66, 470–479
13 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
14 Shamma, S. (1985) Speech processing in the auditory system: I. Representation of speech sounds in the responses of the auditory nerve. *J. Acoust. Soc. Am*. 78, 1612–1621
15 Blackburn, C. and Sachs, M. (1990) The representation of the steady-state vowel /e/ in the discharge patterns of cat anteroventral cochlear nucleus neurons. *J. Neurophysiol*. 63, 1191–1212
16 Clarey, J. *et al.* (1992) Physiology of thalamus and cortex. In *The Mammalian Auditory Pathway: Neurophysiology* (Webster, D. *et al.*, eds), pp. 232–334, Springer-Verlag
17 Nelken, I. and Versnel, H. (2000) Responses to linear and logarithmic frequency-modulated sweeps in ferret primary auditory cortex. *Eur. J. Neurosci*. 12, 549–562
18 Shamma, S. *et al*. (1993) Response area organization in the ferret primary auditory cortex. *J. Neurophysiol*. 69, 367–383
19 Evans, E. and Whitfield, I. (1964) Classification of unit responses in auditory cortex of the unanesthetized and unrestrained cat. *J. Physiol.* 171, 476–493
20 Schreiner, C. and Urbas, J. (1988) Representation of amplitude modulation in the auditory cortex of the cat. i: the anterior field. *Hear. Res*. 21, 227–241
21 Middlebrooks, J.C. *et al*. (1980) Binaural response-specific bands in primary auditory cortex of the cat: topographical organization orthogonal to isofrequency contours. *Brain Res*. 181, 31–48
22 Wang, X. *et al*. (1995) Representation of species-specific vocalizations in the primary auditory cortex of common marmosets: temporal and spectral characteristics. *J. Neurophysiol*. 74, 2685–2706
23 Hubel, D. and Wiesel, T. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol*. 160, 106–154
24 De-Valois, R. and De-Valois, K. (1990) *Spatial Vision*, Oxford University Press
25 Calhoun, B. and Schreiner, C. (1998) Spectral envelope coding in cat primary auditory cortex. *Eur. J. Neurosci.* 10, 926–940
26 Shamma, S.A. *et al*. (1995) Ripple analysis in ferret primary auditory cortex: I. Response characteristics of single units to sinusoidally rippled spectra. *Aud. Neurosci.* 1, 233–254
27 Kowalski, N. *et al*. (1996) Analysis of dynamic spectra in ferret primary auditory cortex: I. Characteristics of single unit responses to moving ripple spectra. *J. Neurophysiol*. 76, 3503–3523
28 Shamma, S.A. and Versnel, H. (1995) Ripple analysis in ferret primary auditory cortex: II. Prediction of unit responses to arbitrary spectral profiles. *Aud. Neurosci*. 1, 255–270
29 Kowalski, N. *et al*. (1996) Analysis of dynamic spectra in ferret primary auditory cortex: II. Prediction of unit responses to arbitrary dynamic spectra. *J. Neurophysiol*. 76, 3524–3534
30 Klein, D.J. *et al*. (1999) Robust spectro-temporal reverse correlation for the auditory system: optimizing stimulus design. *J. Comput. Neurosci.* 9, 85–111
31 Depireux, D.A. *et al*. (2001) Spectrotemporal response field characterization with dynamic ripples in ferret primary auditory cortex. *J. Neurophysiol.* 85, 1220–1234
32 Assmann, P.F. and Paschall, D.D. (1998) Pitches of concurrent vowels. *J. Acoust. Soc. Am*. 103, 1150–1160
33 Culling, J.F. and Darwin, C.J. (1993) Role of timbre in the segregation of simultaneous voices with intersecting F0 contours. *Percept. Psychophys*. 54, 303–309
34 Plomp, R. (1976) *Aspects of Tone Sensation*, Academic Press
35 Schouten, J.F. (1940) The residue and the mechanism of hearing. *Proc. Kon. Ned. Akad. Wet*. 43, 991–999
36 Slaney, M. and Lyon, R. (1993) On the importance of time: a temporal representation of sound. In *Visual Representations of Speech Signals* (Cooke, M. *et al*., eds), pp. 95–116, John Wiley & Sons
37 Langner, G. (1992) Periodicity coding in the auditory system. *Hear. Res*. 6, 115–142
38 Licklieder, J. (1951) A duplex theory of pitch perception. *Experientia* 7, 128–133
39 Meddis, R. and Hewitt, J. (1991) Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *J. Acoust. Soc. Am*. 89, 2866–2882
40 Goldstein, J. (1973) An optimum processor theory for the central formation of pitch of complex tones. *J. Acoust. Soc. Am.* 54, 1496–1516
41 Terhardt, E. (1979) Calculating virtual pitch. *Hear. Res*. 1, 155–182
42 Montgomery, C. and Clarkson, M. (1997) Infants' pitch perception: masking by low- and high-frequency noises. *J. Acoust. Soc. Am*. 102, 3665–3672
43 Shamma, S. and Klein, D. (2000) The case of the missing pitch templates: how harmonic templates emerge in the early auditory system. *J. Acoust. Soc. Am*. 107, 2631–2644
44 Boulez, P. (1989) *Le Pays Fertile: Paul Klee*, Editions Gallimards, Paris
45 Durlach, N. and Colburn, S. (1978) Binaural phenomena. In *Handbook of Perception* (Carterette, E.C. and Friedman, M.P., eds), pp. 365–466, Academic Press
46 Colburn, S. and Durlach, N. (1978) Models of Binaural Interactions. In *Handbook of Perception* (Carterette, E.C. and Friedman, M.P., eds), pp. 467–518, Academic Press
47 Jeffress, A. (1948) A place theory of sound localization. *J. Comp. Physiol. Psychol*. 61, 468–486
48 McAlpine, D. *et al*. (2000) Convergent input from brainstem coincidence detectors onto delay-sensitive neurons in the inferior colliculus. *J. Neurosci.* 18, 6026–6039
49 Shamma, S. *et al*. (1989) Stereausis: binaural processing without neural delays. *J. Acoust. Soc. Am*. 86, 989–1006