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.

Analyses 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. 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. 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.

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 anatomic 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. 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.

‘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 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...
The 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. 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

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

**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\(^1\), 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\(^13\), 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
derived from the comparisons with image analysis in the visual cortex, where significant organizational features occur, such as ocular dominance and orientation columns. 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. 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. 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 (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, $\Omega$) 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 (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 cell's spectral, as well as dynamic, response selectivity. 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.

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. Furthermore, the shape of a
Opinion

Relations 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 ($\Omega_x, \Omega_y$). Therefore, VI cell is (simple or complex) tuned to specific grating orientations are implicitly tuned to specific combinations of spatial frequencies (e.g. a $45^\circ$ 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.

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. 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.

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. 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. 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.

http://tics.trends.com

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 ($\Omega$) of 0.8 cycles octave$^{-1}$ and (b) as a function of ripple density, $\Omega$ at a fixed drift speed of 12 Hz. For a complete measurement of all ripples, this test is repeated at all $w, \Omega$ combinations to construct the transfer function [TF or $T(w, \Omega)$] 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.
other mammals and birds. 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.

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 or other related variants of temporal cues and operations. The results are then combined from across all fibers to get the final estimate without reference to an ordered spatial tonotopic axis. 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. Both types of algorithms are successful in explaining and predicting the pitches of complex tones. Their main shortcoming is the lack of convincing 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.

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. 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.
in the phase-locking frequency range (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. 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.

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. 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
Contralateral cochlea

(b) The stereausis network computes the instantaneous spatial cross-peak of the correlation function) is interpreted as the interaural time difference and, hence, the azimuthal 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 and encode the spatial disparity between the two patterns (contralaterally delayed in this case) that can be exploited to estimate the location of the source.

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.

Relation to vision

The above coincidence algorithm is fundamentally identical to those proposed to solve the stereopsis problem in vision, 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 channel 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.

References