eNeuro

Research Article: Theory/New Concepts | Sensory and Motor Systems

A structural theory of pitch

A structural theory of pitch

Jonathan Laudanski[1,5,†], Yi Zheng[1,2,3,4] and Romain Brette[1,2,3,4]

[1] *Institut d'Etudes de la Cognition, Ecole Normale Supérieure, Paris, France*
[2] *Sorbonne Universités, UPMC Univ. Paris 06, UMR_S 968, Institut de la Vision, Paris, F-75012, France*
[3] *INSERM, U968, Paris, F-75012, France*
[4] *CNRS, UMR_7210, Paris, F-75012, France*
[5] *Scientific and Clinical Research Department, Neurelec, Vallauris, France*

DOI: 10.1523/ENEURO.0033-14.2014

Received: 29 September 2014

Revised: 7 November 2014

Accepted: 7 November 2014

Published: 12 November 2014

Author contributions: J.L. and R.B. designed research; J.L. performed research; J.L. and Y.Z. analyzed data; J.L., Y.Z., and R.B. wrote the paper.

Funding: European Research Council
StG 240132

Conflict of Interest: Authors report no conflict of interest.

[†] deceased, 11 May 2014

Corresponding Author: Romain Brette, INSERM, 17 rue Moreau, Paris, France 75012, voice: +33153462536, Email: romain.brette@inserm.fr

Cite as: eNeuro 2014; 10.1523/ENEURO.0033-14.2014 eNeuro

http://eneuro.msubmit.net

eN-TNC-0033-14R1

A structural theory of pitch

Abstract

Musical notes can be ordered from low to high along a perceptual dimension called « pitch ». A characteristic property of these sounds is their periodic waveform, and periodicity generally correlates with pitch. Thus pitch is often described as the perceptual correlate of the periodicity of the sound's waveform. However, the existence and salience of pitch also depends in a complex way on other factors, in particular harmonic content: for example, periodic sounds made of high-order harmonics tend to have a weaker pitch than those made of low-order harmonics. Here we examine the theoretical proposition that pitch is the perceptual correlate of the regularity structure of the vibration pattern of the basilar membrane, across place and time - a generalization of the traditional view on pitch. While this proposition also attributes pitch to periodic sounds, we show that it predicts differences between resolved and unresolved harmonic complexes and a complex domain of existence of pitch, in agreement with psychophysical experiments. We also present a possible neural mechanism for pitch estimation based on coincidence detection, which does not require long delays, in contrast with standard temporal models of pitch.

Significance statement

Melodies are composed of sounds that can be ordered on a musical scale. "Pitch" is the perceptual dimension on that scale. To a large extent, the periodicity of the sound's waveform can be mapped to pitch. However, the existence and strength of pitch also depends on the harmonic content sounds, i.e., their timbre, which does not fit with this simple view. We propose to explain these observations by the fact that the input to the auditory system is the spatio-temporal vibration of the basilar membrane in the cochlea, rather than the acoustic waveform. We show that defining pitch as the regularity structure of that vibration can explain some aspects of the complexity of pitch perception.

Introduction

A musical note played by a piano or a trumpet has a perceptual attribute called "pitch", which can be low or high. The same key played on different instruments produces sounds with different spectral content but identical pitch. To a large extent, pitch can be mapped to the periodicity, or repetition rate (f0), of the acoustic waveform (Oxenham, 2012). For this reason, theories of pitch perception have focused on how the auditory system extracts periodicity. In the cochlea, the mechanical response of the basilar membrane (BM) to sounds has both a spatial and a temporal dimension. The BM vibrates in response to tones, following the frequency of the tone. The place of maximal vibration along the BM also changes gradually with tone frequency, from the base (high frequency) to the apex (low frequency). Accordingly, there are two broad types of theories of pitch, emphasizing either time or place (de Cheveigné, 2010).

Place theories (or pattern recognition theories) propose that the spatial pattern of BM vibration is compared to internal templates, consisting of harmonic series of fundamental frequencies (Terhardt, 1974). Pitch is then estimated from the fundamental frequency of the best matching template. This mechanism requires that harmonics of the sound produce clear peaks in the spatial pattern of BM vibration, i.e., that harmonics are "resolved" by the cochlea, but this is typically not the case for high-order harmonics because the bandwidth of cochlear filters increases with center frequency. In contrast, tone complexes with only unresolved harmonics can elicit a pitch (Ritsma, 1962; Oxenham et al., 2011). In addition, the firing rate of auditory nerve fibers as well as most neurons in the cochlear nucleus saturates at high levels, but pitch perception does not degrade at high levels (Cedolin and Delgutte, 2005).

Temporal theories propose that periodicity is estimated from the temporal waveform in each auditory channel (cochlear place), and estimates are then combined across channels (Licklider, 1951; Meddis and O'Mard, 1997; de Cheveigné, 2010). Sound periodicity is indeed accurately reflected in the patterns of spikes produced by auditory nerve fibers (Cariani and Delgutte, 1996a, 1996b; Cedolin and Delgutte, 2005). Resolvability plays little role in these theories, but pitch based on resolved harmonics is more salient and easier to discriminate than pitch based on unresolved harmonics (Houtsma and Smurzynski, 1990; Carlyon and Shackleton, 1994; Carlyon, 1998; Bernstein and Oxenham, 2003). Finally, detecting the periodicity of a waveform with repetition rate f0 = 30 Hz (the lower limit of pitch (Pressnitzer et al., 2001)) would require delays of about 30 ms, of which there is no clear physiological evidence.

In addition, the domain of existence of pitch is complex, which neither type of theory explains: the existence of pitch depends not only on f0 but also on resolvability of harmonics and spectral content (Pressnitzer et al., 2001; Oxenham et al., 2004b, 2011). For example, high frequency complex tones (>4 kHz) with f0 = 120 Hz do not have a clear pitch while a pure tone with the same f0 does (Oxenham et al., 2004b); but high frequency complex tones with f0>400 Hz do have a clear pitch (Oxenham et al., 2011). Finally, while pitch is generally independent of sound intensity (contradicting place theories (Micheyl and Oxenham, 2007)), a few studies suggest a small but significant intensity dependence of pitch for low frequency pure tones (Morgan et al., 1951; Verschuure and Van Meeteren, 1975; Burns, 1982) (contradicting temporal theories).

Here we propose to address these issues by reexamining the postulate that pitch is the perceptual correlate of the periodicity of the acoustic waveform. Starting from the observation that the input to the auditory system is not the acoustic waveform but the vibration pattern of the BM, we propose instead that pitch is the perceptual correlate of the regularity structure of the BM vibration pattern, across place and time. While this proposition also attributes pitch to periodic sounds, we show that it predicts differences between resolved and unresolved harmonic complexes and a complex domain of existence of pitch. We also present a possible neural mechanism for pitch estimation based on coincidence detection, which does not require long delays.

Materials and Methods

Auditory filters

Auditory filters were modeled as gammatone filters (Slaney, 1993; Fontaine et al., 2011), which approximate reverse correlation filters of cat auditory nerve fibers (Boer and Jongh, 1978; Carney and Yin, 1988) and have been matched to psychophysical measurements in humans (Glasberg and Moore, 1990). Their impulse response defined by: $H(t) = t^{n-1}e^{-t/\tau}\cos(2\pi.CF.t)$, where CF is the characteristic frequency, n is the order and the bandwidth is set by $\tau = \left(2\pi \cdot 1.019 \cdot (24.7 + 0.108 \cdot CF)\right)^{-1}$. Filters were spaced uniformly in ERB scale (Glasberg and Moore, 1990) with CF between 100 and 8000 Hz.

Neural model of pitch estimation

The neural model of pitch estimation includes two layers: 1) the input layer (putatively cochlear nucleus) and 2) coincidence detector neurons.

Input layer

Each neuron receives the output x(t) of a gammatone filter, after half-wave rectification and compression with a power law with exponent $\gamma = 0.3$ (Stevens, 1971; Zwislocki, 1973): $y(t) = \kappa([x(t)]^{+})^{\gamma}$ (varying the exponent between 0.2 and 0.5 did not affect the results).

We tested different spiking neuron models (Fig. 4), defined by a membrane equation of the following form:

$$C\frac{dV}{dt} = g_L(E_L - V) + y(t) + \sigma\xi(t) + I(V), \tag{1}$$

where V is the membrane potential, $g_L(E_L - V)$ represent the non-specific leak current, σ is the noise level, C is the membrane capacitance and I(V) represents currents from voltage-gated channels.

The chopper cell model (T-multipolar) is based on the model of Rothman and Manis (Rothman and Manis, 2003a), with maximal conductances $g_{Na} = 1000$ nS, $g_{KHT} = 150$ nS, and $g_h = 0.5$ nS. Octopus cells are also based on the same model but include a low threshold potassium channel (KLT) and model of I_h taken from (Khurana et al., 2011), with $g_{Na} = 1000$ nS, $g_{KHT} = 150$ nS, $g_{KLT} = 600$ nS, and $g_h = 40$ nS. These two models were used only in Fig. 4.

We also used a leaky integrate-and-fire model (LIF), a phenomenological model with good predictive value for a broad class of neurons (Jolivet et al., 2004; Gerstner and Naud, 2009). The membrane time constant was $\tau = g_L/C = 1.5$ ms. The model spikes when V(t) reaches the threshold $\theta = -40$ mV, and V(t) is then reset to $V_r = -60$mV and clamped at this value for a refractory time of $\tau_r = 1$ ms. This model was used in all simulations, unless otherwise specified.

Coincidence detectors

The second layer consists of coincidence detectors, which are modeled as integrate-and-fire models (as above) with an adaptive threshold governed by the following equation (Platkiewicz and Brette, 2010, 2011; Fontaine et al., 2014):

$$\tau_\theta \frac{d\theta}{dt} = \theta_0 - \theta + V - E_L, \tag{2}$$

where $\theta_0 = -40$mV is the value of threshold at rest and $\tau_\theta = 5$ ms. (note that half-wave rectification can be discarded here because V is always above E_L, as there are only excitatory synapses). This equation ensures that the neuron is always in a fluctuation-driven regime where it is sensitive to coincidences (Platkiewicz and Brette, 2011). The response of the coincidence detectors was only considered after 30 ms following note onset.

Synaptic connections

For each possible f0, we build a group of coincidence detectors whose inputs are synchronous when a sound of period 1/f0 is presented. For any sound, the *synchrony partition* is defined as the set of groups of input neurons that fire in synchrony for that particular sound (Brette, 2012) (synchrony is within group, not across groups). One coincidence detector neuron is assigned to each group (synaptic connections from each input neuron to the coincidence detector), so that the synchrony partition corresponds to a set of coincidence detector neurons.

To build a group of coincidence detector neurons tuned to periodic sounds with fundamental frequency f0, we consider the synchrony partition of the complex tone made of all harmonics of f0, i.e., tones of frequency k.f0. For each harmonic, we select all pairs of channels in our filter bank that satisfy the following properties (Fig. 2D): 1) the gain at k.f0 is greater than a threshold G_{min} = 0.25 (dashed line in Fig. 2D); 2) the two gains at k.f0 are within $\epsilon = 0.02$ of each other; 3) the gain at neighboring harmonics (order k-1 and k+1) is lower than the threshold G_{min} (resolvability criterion). For each selected pair of channels, we connect the corresponding input neurons to a single coincidence detector neuron. The connection from the neuron with higher CF has an axonal delay $\delta = \Delta\phi / k f_0$, where $\Delta\phi$ is the phase difference between the two filters at k.f0 (which is known analytically for a gammatone (Zhang et al., 2001)). In addition, for each channel, multiple neurons receiving inputs from the same filter project to a single coincidence detector neuron with axonal delays $\delta = \mathrm{k}/f_0$ (as in Licklider's model), where k in the integer varying between 1 and a value determined by the maximum delay δ_{max}.

Sounds

Musical instruments

To test the neural model in a pitch recognition task, we used recordings of musical instruments and vowels from the RWC Music Database (Musical Instrument Sound), including 762 notes between A2 and A4, 41 instruments (587 notes) and 5 sung vowels (175 notes). Notes were gated by a 10 ms cosine ramp and truncated after 500 ms.

Environmental noises

We also used a set of 63 environmental sounds containing stationary noises including: airplanes, thunderstorm, rain, water bubbles, sea waves, fire and street sounds (recordings obtained from www.freesound.org). We selected 500 ms segments from these sounds, gated by a 10 ms cosine ramp.

Analytical description of the auditory nerve phase response

To analyze the discriminability of cross-channel structure (Fig. 6E-F), we fitted an analytical formula to the phase $\phi(L, f, CF)$ of auditory nerve responses recorded at different levels L and tone frequencies f in fibers with different CF, using a data set from Palmer and Shackleton (2009) (Palmer and Shackleton, 2009), similarly to Carlyon et al. (Carlyon et al., 2012). For each level, we fitted a function corresponding to the phase response of a gammatone filter bank:

$$\phi(L, f, CF) = f\psi(L, CF) + n\arctan\big(2\pi\tau(CF, L)(f - CF)\big)$$

where $\psi(L, CF)$ is the initial delay of the travelling wave (a parameterized function of CF, equation (3) in (Zhang et al., 2001)), n in the order of the gammatone filter and $\tau(CF, L) = \alpha(L)CF^{\beta(L)}$ is inversely related to the bandwidth of the filter.

We also tested another function: $\phi(L, f, CF) = \alpha(L, f) + \beta(L, f)\arctan\big(\mathrm{CF}/\gamma(L, f)\big)$ as in Carlyon et al. (Carlyon et al., 2012), where α, β and γ were second-order polynomial functions of L and f. The fits gave similar results.

Discriminability of cross-channel and within-channel structure

We used signal detection theory (GREEN and SWETS, 1966) to estimate the discriminability of tone frequency based on regularity structure, using only phase information (to simplify). We consider two places on the cochlea tuned to frequencies f_A and f_B. A tone of frequency f is detected when the two waveforms at places A and B are in phase after a delay d is introduced in channel B: $\phi(f_B, f) + fd = \phi(f_A, f) + n$, where n is an integer (phases are expressed in cycles). Note that n is related to related to the maximum delay $\delta_{\max}$ (when f < 1/ $\delta_{\max}$, there is at most one possible value for n).

We note $\Delta\phi_{AB}(f) = \phi(f_B, f) - \phi(f_A, f)$ the phase difference between the two places (before the delay is introduced), so that the equation reads:

$$\Delta\phi_{AB}(f) + f\delta = n \qquad (3)$$

That is, the phase difference after the delay is introduced is 0 cycle. When a tone of frequency f+df is presented, the phase difference after the delay is introduced is $\Delta\phi_{AB}(f + df) + (f + df)\delta = \Delta\phi_{AB}(f) + f\delta + (\Delta\phi'_{AB}(f) + \delta).df = n + (\Delta\phi'_{AB}(f) + \delta).df$. Thus, a frequency shift of df induces a phase shift of $(\Delta\phi'_{AB}(f) + \delta).df$ between the two channels, after introduction of the delay.

We consider that neurons corresponding to channels A and B fire spikes in a phase-locked manner with precision σ (standard deviation of spike phase). Then the discriminability index d' is the mean phase shift divided by the precision:

$$d' = \frac{(\Delta\phi'_{AB}(f) + \delta).df}{\sigma}$$

The just-noticeable difference (JND) for 75% correct discrimination is then:

$$JND = 1.35\frac{\sigma}{\Delta\phi'_{AB}(f) + \delta}$$

The Weber fraction is JND/f. For two identical channels (within-channel structure), δ =1/f and the formula simplifies to:

$$JND_{75\%} = 1.35\sigma f$$

For distinct channels (cross-channel structure), d is determined by equation (3), and the formula reads:

$$JND_{75\%} = 1.35\frac{\sigma f}{(\mathrm{f}.\Delta\phi'(f) + n - \Delta\phi_{AB}(f))}$$

Finally, we relate phase precision with vector strength VS using the following formula, based on the assumption that phases are distributed following a wrapped-normal distribution:

$$\sigma = \sqrt{-\ln(VS^2)}\,/2\pi$$

Results

The proposition

In the cochlea, the BM vibrates in response to sounds. We denote by S(x,t) the displacement of the BM at time t and place x. This displacement is represented in Fig. 1A as the output of a gammatone filterbank with bandwidth based on psychophysical measurements (see Methods). Each auditory nerve fiber transduces the temporal vibration S(x,t) at a specific place into a spike train. In Licklider's delay line model (the classical temporal model (Licklider, 1951)), the periodicity of the mechanical vibration is detected by a coincidence detector neuron receiving synaptic inputs from a single cochlear place x. It fires when it receives coincidences between a spike train produced by a fiber originating from that place and the same spike train delayed by a fixed amount δ (Fig. 1B). Conceptually, this neuron detects the identity S(x,t+δ)=S(x,t) for all t, that is, the fact that S(x,.) is periodic with period T = δ. This mechanism must be slightly amended to account for the refractory period of fibers, which sets a lower limit to the period that can be detected. This issue can be addressed by postulating that the neuron receives inputs from two different fibers originating from the same place (Fig. 1C).

We now consider the possibility that these two fibers may originate from slightly different cochlear places x and y. In this case, the neuron detects the identity S(y,t+δ)=S(x,t), that is, similarity of sensory signals across both place and time (Fig. 1D). We note in this example (a harmonic sound) that the delay δ may now be different from the period T of the vibration. Compared to the detection of periodicity, this does not require any additional anatomical or physiological assumption. Thus we propose to examine the proposition that pitch is the perceptual correlate of the regularity structure of the BM vibration pattern, across both time and place, defined as the set of identities of the form S(x,t)=S(y,t+δ) for all t. A few previous models of pitch also use cross-channel comparisons (Loeb et al., 1983; Shamma, 1985; Carney et al., 2002), and we will relate them to our theory in the discussion.

To illustrate our proposition, Fig. 1E-F show the cochleograms obtained by filtering two sounds with a gammatone filterbank. A noise-like sea wave (Fig.1E) produces no regularity structure in the cochleogram, that is, there are no identities S(x,t)=S(y,t+δ) in the signals. A clarinet note, on the other hand, produces a rich regularity structure (Fig. 1F). Because this is a periodic sound, the BM vibrates at the sound's period T at all places (or more generally T/k, where k is an integer), as shown by horizontal arrows: S(x,t+T)=S(x,t) for all t and x. We call this set of identities the *within-channel structure*. More interestingly, we also observe identities across places, as shown by oblique arrows: S(x,t)=S(y,t+δ) for all t. These occur for specific pairs of places x and y, which tend to be in low frequency regions. We note that the time shift δ is different from the sound's period T. We call this set of identities the *cross-channel structure*.

Resolvability and regularity structure

We now examine the type of regularity structure produced by sounds. First, if the sound is periodic, then the BM vibrates at the sound's period T at all places, provided there is energy at the corresponding frequency. That is, S(x,t+T)=S(x,t) for all x and t. Conversely, the identity S(x,t+T)=S(x,t) means that the BM vibrates periodically, which can only occur if the sound itself is periodic, at least within the bandwidth of the cochlear filter at place x. Thus, within-channel structure is simply the periodicity structure at each cochlear place.

Cross-channel structure is less trivial. What kind of sound produces the same vibration (possibly delayed) at different places of the cochlea? To simplify the argument, we consider that cochlear filters are linear (we come back to this point in the discussion), and we examine the identity S(x,t)=S(y,t+δ) in the frequency domain. If the two signals at place x and y match, then all their frequency components must match, both in phase and amplitude. But these two signals originate from the same sound, filtered in two different ways. Fig. 2A shows the gain (left) and phase (right) of the two filters A and B as a function of frequency. The only way that a frequency component is filtered in the same way by the two filters is that the gains are identical at that frequency, which happens in this case at a single frequency f (illustrated on Fig. 2A, bottom). Additionally, the phases of the two filters must match at frequency f, taking into account the delay δ. That is, the phase difference $\Delta\phi$ must equal f.δ (modulo 1 cycle).

In summary, the only type of sound that produces cross-channel structure is a sound with a single frequency component within the bandwidth of the two considered cochlear filters. This is a notion of *resolvability*, and we will say that the frequency component is *resolved* with respect to the pair of filters. Fig. 2B illustrates what happens when a periodic sound with unresolved harmonics is passed through the two filters. Here the output of filter A is a combination of harmonics k and k-1, while that of filter B is a combination of harmonics k and k+1. Therefore, the two resulting signals are different (bottom): there is no cross-channel structure.

Thus, the amount of cross-channel structure produced by a harmonic sound depends on the resolvability on its frequency components. Fig. 2C shows the amplitude spectrum of a periodic sound with all harmonics k.f0 (bottom). Because harmonics are linearly spaced but cochlear filter bandwidth increases with frequency (filter amplitude in gray), the excitation pattern of the

BM as a function of center frequency (top) shows distinct peaks for low-order harmonics (which are thus considered "resolved") but not for high-order harmonics (unresolved). More precisely, low-order harmonics are resolved for many pairs of cochlear filters, meaning that they produce cross-channel structure for many filter pairs (Fig. 2D, left); high-order harmonics produce little or no cross-channel structure (Fig. 2D, right). The amount of cross-channel structure is directly determined by the spacing between frequency components (f0) relative to the cochlear filter bandwidth. With the approximation that filter bandwidth is proportional to center frequency (k.f0 if centered at the k^{th} harmonic), this means that the amount of cross-channel structure is determined by the harmonic number k. Therefore, there is a direct relationship between resolvability defined in a conventional sense and the amount of cross-channel structure produced by the sound.

Figure 2E illustrates this point with a resolved harmonic complex consisting of resolved components (left) and with an unresolved harmonic complex (right). Both sounds produce within-channel structure (horizontal arrows), but the resolved complex additionally produces cross-channel structure. Thus, the structural theory attributes a pitch to all periodic sounds, but the amount of regularity structure, and therefore of information about f0, depends on resolvability. It follows in particular that discrimination of f0 based on regularity structure should be more precise for resolved than unresolved sounds (Houtsma and Smurzynski, 1990; Carlyon and Shackleton, 1994; Carlyon, 1998; Bernstein and Oxenham, 2003), since there is more information (the exact quantitative assessment would depend on the specific estimator chosen).

The domain of existence of pitch

From the definitions above, the set of sounds that produce regularity structure is exactly the set of periodic sounds. However, perceptually, not all periodic sounds have a melodic pitch. In particular, pitch only exists for f0 between 30 Hz (Pressnitzer et al., 2001) and 5kHz (Semal and Demany, 1990). Within this range, periodic sounds may or may not have a clear pitch, depending on their harmonic content. In the structural theory, the domain of existence of pitch is restricted when we impose constraints on the comparisons between signals (cross- or within-channel) that the auditory system can do. Two physiological constraints seem unavoidable: 1) there is a maximum time shift δ_{max} (possibly corresponding to a maximum neural conduction delay), 2) temporal precision is limited (possibly corresponding to phase locking precision). We may also consider that there is a maximum distance along the BM across which signals can be compared, but it will not play a role in the discussion below. The temporal precision sets an upper limit to pitch, exactly in the same way as in standard temporal theories. Thus we shall restrict our analysis to the constraint of a maximum delay δ_{max}. We consider the simplest possible assumption, which is a constant maximal delay δ_{max}, independent of frequency.

We start by analyzing the domain of existence of within-channel structure (Fig. 3A). Since this is just the periodicity structure, its domain of existence is the same as in standard temporal theories of pitch. When the sound's period exceeds the maximum delay δ_{max}, periodicity cannot be detected anymore. Therefore, the lower limit (minimum f0) is the inverse of the maximum delay: f0 = $1/\delta_{max}$.

A different limit is found for cross-channel structure, because the delay δ between signals across channels is not the same as the sound's period (see e.g. Fig. 1F). In fact, this delay can be arbitrary small, if the two places are close enough on the BM. Figure 3B shows an example of a 100 Hz pure tone passed through two filters A and B. The gains of the two filters are the same at 100 Hz and there is a phase difference of 8/10 cycle, which is equivalent to -2/10 cycle. As a result, the output of the two filters is a pair of tones with identical amplitude and delay $\delta = 2$ ms (2/10 of 10 ms), much smaller than the sound's period. This delay would be even smaller if the center frequencies of the two filters were closer. Thus the lower limit of cross-channel structure is not set by the maximum delay δ_{max}. Instead, it is set by the center frequencies of the filters. Indeed the frequency of the tone (or resolved harmonic) must lie between the two center frequencies of the filters, and therefore the lowest such frequency corresponds to the lowest center frequency of cochlear filters. This minimum frequency is not known in humans, but the lower limit of the hearing range is about 20 Hz, which suggests a lower limit of cross-channel structure slightly above 20 Hz. This is consistent with psychophysical measurements of the lower limit of pitch, around 30 Hz for tones (Pressnitzer et al., 2001).

Therefore, the structural theory of pitch predicts different lower limits of pitch depending on whether the sound contains resolved harmonics or not. When it does, the lower limit is determined by cross-channel structure, and thus by the lowest center frequency of cochlear filters, on the order of a few tens of Hz. When it does not, the lower limit of pitch is determined by within-channel structure, and is thus $1/\delta_{max}$. We now compare these theoretical predictions with two recent psychophysical studies. In Oxenham et al. (2004) (Oxenham et al., 2004a), transposed stimuli were created by modulating a high frequency carrier (>4 kHz) with the temporal envelope of a half-wave rectified low frequency tone (<320 Hz) (Fig. 3C, top). Human subjects displayed poor pitch perception for these stimuli, even though the repetition rate f_0 was in the range of pitch perception for pure tones. This finding poses a challenge for temporal theories, but is consistent with the structural theory, as is illustrated in Fig. 3C. Indeed, these transposed tones do not contain resolved harmonics, and therefore only produce within-channel structure (horizontal arrows in Fig. 3C). As described above, the lower limit of pitch is $1/\delta_{max}$ in this case. If this maximal delay is $\delta_{max} < 3$ ms, then transposed tones do not produce a pitch when the frequency of the tone is lower than 330 Hz. On the other hand, for pure tones, the lower limit of pitch is much lower than 330 Hz because of the presence of cross-channel structure (oblique arrows in Fig. 3D). In Oxenham et al. (2011) (Oxenham et al., 2011), it was shown that complex tones with f0 between 400 Hz and 2 kHz and all harmonics above 5 kHz elicit a pitch. In the structural theory, all periodic sounds with f0 > $1/\delta_{max}$ produce a pitch, irrespective of their harmonic content. This is shown in Fig. 3E, which shows the cochlear filter responses to a complex tone with $f_0 = 1.2$ kHz and all harmonics above 5 kHz. Therefore, this psychophysical study is consistent with the structural theory if $\delta_{max} > 2.5$ ms. In summary, both psychophysical studies are consistent with the structural theory if δ_{max} is on the order of 3 ms.

A possible neural mechanism

We now propose a possible neural mechanism to estimate f0 based on the vibration structure of the BM. Since the theory is based on similarity between signals, the same mechanism as for temporal models can be suggested. A straightforward generalization of Licklider's model (Licklider, 1951) is illustrated in Fig. 1D: a neuron receives inputs from two presynaptic neurons

(X and Y), which encode the BM vibration at two cochlear locations x and y in precisely timed spike trains, and there is a mismatch δ in their conduction delays. We assume that the postsynaptic neuron responds preferentially when it receives coincident input spikes. Indeed, neurons are highly sensitive to coincidences in their inputs, under broad conditions (Rossant et al., 2011). By acting as a coincidence detector, the postsynaptic neuron signals a particular identity $S(y,t+\delta)=S(x,t)$.

Anatomically, neurons X and Y could be auditory nerve fibers and the postsynaptic neuron could be in the cochlear nucleus. Alternatively, neurons X and Y could be primary-like neurons in the cochlear nucleus, for example spherical bushy cells, and the postsynaptic neuron could be in the inferior colliculus or in the medial superior olive. Indeed, as demonstrated in Fig. 4A-B, the synchrony between two neurons depends on the similarity between the signals they encode, rather than on their specific cellular properties. Fig. 4A shows the cochleogram of a trumpet note with f0 = 277 Hz (top). The red and blue boxes highlight the BM vibration at characteristic frequencies 247 Hz and 307 Hz, around the first harmonic. This harmonic produces cross-channel similarity with delay δ, as seen on the red and blue signals shown below (grey shading is the mismatch). As a result, neurons that encode these two signals into spike trains fire in synchrony, as is shown below for three different models: a biophysical model of a type Ic chopper neuron (Rothman and Manis, 2003b), a type II model of an octopus cell, and a leaky integrate-and-fire model. In contrast, when an inharmonic sound is presented, such as a rolling sea wave (Fig. 4B), the inputs do not match and neural responses are not synchronous, for any of the three models.

The same mechanism applies for within-channel structure. In Fig. 4C, we consider two high-frequency neurons with the same characteristic frequency CF = 2700 Hz but a delay mismatch δ = 4.5ms. When a periodic sound with repetition rate 220 Hz is presented (here a harpsichord note), their input signals match, which results in synchronous discharges. We note that not all output spikes are coincident. This occurs because the neurons discharge in more complex spiking patterns (Laudanski et al., 2010) and do not fire one spike per cycle: they may miss a cycle or fire several times in one cycle. Nevertheless, coincidences of output spikes occur much less often with an inharmonic sound (Fig. 4D). This mechanism is analog to Licklider’s model (Licklider, 1951), in which each neuron signals a particular identity $S(x,t+\delta)=S(x,t)$. Thus the neural mechanism we describe is simply an extension of Licklider’s model to cross-channel similarity.

As a proof of concept, we now build a simple neural model that estimates f0 by detecting regularity structure. For each f0 between notes A2 and A4 (110 Hz to 440 Hz), we build a group of coincidence detector neurons, one for each similarity identity $S(y,t+\delta)=S(x,t)$ that is present for sounds with that particular f0. To this aim, we examine the BM response (modeled as gammatone filters) to a complex tone with all harmonics n.f0 (Fig. 4E, red comb on the left). On Fig. 4E-F, we represent the BM response using color disks arranged as a function of cochlear location (vertical axis) and delay (horizontal axis): color saturation represents the amplitude of the filter output while hue represents its phase. For low-order harmonics (resolved, bottom), the BM vibrates as a sine wave and therefore disks with the same color correspond to identical signals, and thus to encoding neurons firing in synchrony. For high-order harmonics (unresolved, top), the BM vibrates in a more complex way and there only identically colored disks within the same channel correspond to identical signals. We then set synaptic connections from neurons encoding the same BM signal to a specific coincidence detector neuron (all modeled as integrate-and-fire neurons). Thus we obtain a group of neurons that fire preferentially when the identities $S(y, t+\delta) = S(x, t)$ corresponding to a particular f0 occur (note that we have omitted a number of possible identities for simplicity, e.g. cross-channel identities occurring with high frequency pure tones). In this way, the mean firing rate of the group of neurons is tuned to f0.

We iterate this construction for every f0 between A2 and A4 (by semitone steps). As illustrated in Fig. 4F, a different f0 produces a different regularity structure (colored disks), from which we build a different set of synaptic connections to the pitch-tuned group of coincidence neurons (one group per f0). To estimate f0, we then simply look for the pitch-tuned group with the highest mean firing rate.

We presented two types of natural sounds to this model (spectrograms shown in Fig. 5A, top): inharmonic sounds (e.g. an airplane, a sea wave and street noise), and harmonic sounds (e.g. clarinet, accordion and viola) with f0 between A2 and G#4. For each sound, we measure the average firing rate of all pitch-tuned neuron groups (Fig. 5A, bottom). Inharmonic sounds generally produce little activation of these neurons, whereas harmonic sounds activate specific groups of neurons (with some octave confusions, see below). In Fig. 5A, musical notes were played in chromatic sequence, which appears in the response of pitch-tuned groups. Fig. 5B shows the distribution of group firing rates, measured in the entire neuron model, for inharmonic (grey) and harmonic sounds (blue), at three different sound levels. Although an increase in sound level produces an overall increase in population firing rate, there is little overlap between the rate distributions for harmonic and inharmonic sounds.

From the activity of these neurons, we estimate the pitch of a presented harmonic sound as the pitch associated to the maximally activated group of neurons. This estimation was correct in 77% of cases, and was within one semitone of the actual pitch in 88% of cases (Fig. 5C, top). Most errors greater than one semitone correspond to octaves or fifths (octaves: 5.5%, fifth: <2%), which also shows in the distribution of firing rate of pitch-tuned groups (Fig. 5C, bottom). This performance was obtained with 400 frequency channels spanning 50 Hz to 8 kHz, and it degrades if the number of channels is reduced (e.g. 35% score for N = 100, Fig. 5D, top), because the model relies on comparisons between neighboring channels. We then tested how performance was affected by constraints on the maximum delay (Fig. 5D, bottom). We found no difference in performance when maximum delay δ_{max} was varied between 2 and 15 ms. The highest f0 in our sound database was 440 Hz (A4), which corresponds to a period greater than 2 ms. Therefore with δ_{max} = 2 ms, the model reached the same level of performance with only cross-channel comparisons.

Pitch discriminability

Finally, we examine the discriminability of pure tones based on regularity structure. To simplify, we ignore amplitude differences and focus on phase differences between channels. We start with within-channel structure and consider two neurons (e.g. auditory nerve fibers) encoding BM vibration from the same place x (i.e., same characteristic frequency) into phase-locked spike trains, with a delay mismatch $\delta = 1/f$ (Fig. 6A). These two neurons fire in synchrony when a pure tone of frequency f is presented. More precisely, given that there is some stochasticity in neural firing, the two neurons produce spikes with the same mean phase relative to the tone, so the difference of phases of spikes $\Delta\Phi(f)$ is distributed around 0 (Fig. 6A, left). When a tone of frequency f+df is presented, $\Delta\Phi(f)$ shifts by an amount of $\delta.df = df/f$ (Fig. 6A, right).

The same analysis applies for cross-channel structure, where the two neurons encode BM vibration at two different places A and B (different CFs, Fig. 6B). Here the delay δ is related to the mismatch in phase response at the places at tone frequency f. When a tone of frequency f+df is presented, $\Delta\Phi(f)$ shifts because of both the delay and the relative change in response phase at the two places on the BM (see Methods).

Thus, discriminating between tones of nearby frequencies corresponds to discriminating between two circular random variables $\Delta\Phi(f)$ and $\Delta\Phi(f+df)$ with different means, which can be analyzed with signal detection theory (GREEN and SWETS, 1966). Specifically, the discriminability index d' is the mean phase shift μ divided by the precision σ (standard deviation of phase) (Fig. 6C). The precision of phase locking is often measured by the vector strength (VS), which is relatively independent of frequency below a critical frequency above which it decays rapidly to 0 (Fig. 6D, guinea pig auditory nerve). We estimate the standard deviation σ from VS assuming a wrapped normal distribution (see Methods). To estimate μ, we used spike trains recorded in guinea pig auditory nerve fibers with different CFs in response to tones with various frequencies (Palmer and Shackleton, 2009) and estimated the average spike phase as function of both CF and tone frequency (see Methods) (Fig. 6E).

We used these estimates to calculate the just-noticeable difference (JND) for 75% correct discrimination, which is the frequency change df producing a discriminability index d' = 1.35. Figure 6F shows the JND relative to tone frequency (JND(f)/f), called the Weber fraction, as a function of tone frequency, for within-channel structure (black) and for cross-channel structure (colors), for pairs of channels varying by their spacing in CF (1 to 6 semitones). For both types of structure, the Weber fraction increases in high frequency because of the loss of phase locking (VS goes to 0). The two types differ in the low-frequency end: while the Weber fraction is independent of frequency for within-channel structure, it tends to increase with lower frequency for cross-channel structure. We also note that discriminability is better for widely spaced channels (orange) than for neighboring channels (blue), but the former require larger delays.

Discussion

We have proposed that pitch is the perceptual correlate of the regularity structure of the BM vibration pattern, defined as the set of identities of the form $S(x,t)=S(y,t+\delta)$ for all t, where $S(x,t)$ is the displacement of the BM at time t and place x. The regularity structure generalizes the notion of periodicity. This proposition assigns a pitch to periodic sounds and therefore has many similarities with the standard view that pitch is the perceptual correlate of the periodicity of the acoustic waveform. However, it also predicts that resolved harmonic complexes elicit a stronger pitch than unresolved harmonic complexes (richer structure), and it predicts a complex region of existence of pitch that depends on harmonic content. In particular, it predicts that high frequency complex tones only elicit a clear pitch if f0 is high, in agreement with experiments (Oxenham et al., 2004b, 2011). Finally, it does not rely on the existence of long conduction delays in the auditory system.

Previous studies have proposed mechanisms to extract the fundamental frequency of either resolved or unresolved harmonic complexes (see detailed discussion in section "Related theories of pitch" below). Some share common ideas with our proposition: for example, classical temporal models address the extraction of within-channel periodicity (S(x,t) = S(x,t+T)) (de Cheveigné, 2010), which does not distinguish between resolved and unresolved components; other authors have proposed that the frequency of resolved components can be estimated with cross-channel comparisons or operations (Loeb et al., 1983; Shamma, 1985; Carney et al., 2002). These ideas are also present in our proposition. However, instead of proposing a particular mechanism to extract f0, we propose that pitch is not the correlate of the periodicity of the sound waveform but of the regularity structure of the BM vibration pattern (with a limited temporal window). The main implications for pitch perception (as shown in Fig. 3) are to a large extent independent of the particular mechanism that extracts that structure. In particular, this single proposition implies that resolved and unresolved harmonic complexes have different perceptual properties.

Neural mechanism

A separate issue is the physiological implementation of this theory, that is, how pitch defined according to the regularity structure of the BM vibration pattern might be estimated by the auditory system. There are different ways in which the auditory system might extract that information. It may also be the case that pitch is not conveyed by the increased firing of pitch-tuned neurons but by temporal relationships in their firing (Cariani, 2001). Here we have simply made a suggestion of a possible mechanism that makes minimal physiological assumptions. But we stress that our core proposition does not rely on a particular mechanism, but on the regularity structure of the BM vibration. The most straightforward implementation is a generalization of Licklider's delay line model (Licklider, 1951), in which a pitch-selective neuron detects coincidences between two inputs with different axonal conduction delays. In the original model, the two inputs originate from the same place in the cochlea. An implementation of the structural theory is obtained simply by allowing the two inputs to originate from slightly different places. If a neural circuit resembling Licklider's model indeed exists in the auditory brainstem, then it is plausible that inputs to these coincidence detector neurons are not exactly identical. Because our proposition relies on the temporal fine structure of sounds, the matching mechanism between the outputs of two channels (whether it is based on coincidence detection or not) should occur early in the auditory periphery. Input neurons could be auditory nerve fibers and the coincidence detector neuron could be in the cochlear nucleus. Alternatively, input neurons could be primary-like neurons in the cochlear nucleus, for example spherical bushy cells, and the coincidence detector neuron could be in the inferior colliculus or in the medial superior olive (MSO). The latter possibility has some appeal because neurons in the MSO are thought to receive few synaptic inputs (Couchman et al., 2010) and are known to act as coincidence detectors (Yin and Chan, 1990), although possibly not monaurally (Agmon-Snir et al., 1998), and there are cases of binaural pitch for sounds that have no monaural structure. In the inferior colliculus, there is some physiological evidence of tuning to pitch (Langner, 1992). Specifically, in a number of mammalian species, IC neurons are tuned in their firing rate to the modulation frequency of amplitude-modulated tones, up to about 1000 Hz, independently of their characteristic frequency, although the best modulating frequency may depend on carrier frequency. There is also some evidence of a topographic organization of periodicity tuning, orthogonal to the tonotopical organization.

As a proof of principle, we have shown with a spiking neural model that such a mechanism can indeed estimate the pitch of harmonic sounds, even with short conduction delays. Standard temporal models of pitch have been criticized because they require long delays for low f0, up to 30 ms for the lowest pitch (Pressnitzer et al., 2001). There is no experimental evidence of such long axonal delays in the auditory brainstem. In a recent anatomical study of axons of spherical bushy cells in cats (cochlear nucleus projections to the MSO) (Karino et al., 2011), the range of axonal delays was estimated to be just a few hundred µs, far from the required 30 ms (although these were anatomical estimates, not functional measurements). This range could be larger in humans as axons are presumably longer, but it could also be similar if axonal diameter scales in the same way (since conduction speed is approximately proportional to diameter in myelinated axons (Rushton, 1951)). In either case, the range of axonal delays is unlikely to be much greater than a few ms. Another possibility is to consider dendritic propagation delays or intrinsic delays induced by ionic channels. These could contribute additional delays, but the duration of postsynaptic potentials measured at the soma of auditory brainstem neurons tends to be short (Trussell, 1997, 1999), which makes this scenario rather implausible for large delays. We have shown that the structural theory is compatible with psychophysical results when the delays are limited to a few ms, and the neural mechanism based on coincidence detection remains functional even for low f0.

Related theories of pitch

Two previous propositions are directly related to the structural theory. Loeb et al. (Loeb et al., 1983) proposed that the frequency of a pure tone can be estimated by comparing signals across the BM: the distance that separates places that vibrate in phase is indeed related to the tone's frequency. This is a special case of the structural theory, when the maximal delay is 0 ms (i.e., identities of the form $S(x,t) = S(y,t)$ for all t). However, this proposition restricts pitch to resolved harmonic complexes only, and in fact to complexes made of widely separated tones.

The phase opponency model (Carney et al., 2002) is a similar proposition, in which a tone of a particular frequency is detected when signals at two different places on the BM are out of phase. This corresponds to detecting identities of the form $S(x,t) = -S(y,t)$ for all t. This model suffers from the same problem as Loeb's model, that is, it applies to a limited subset of pitch-evoking sounds.

We may also consider a variation of the structural theory, in which amplitude is discarded (as we did when analyzing frequency discrimination). This variation corresponds to considering identities of the form $S(x,y) = a.S(y,t+\delta)$ for all t. This variation has the same qualitative properties as the original formulation, and is physiologically motivated by the observation that low threshold AN fibers saturate quickly when intensity is increased (Sachs and Abbas, 1974).

Place theories of pitch are based on the comparison of internal templates with the spatial pattern of BM vibration encoded in the firing of auditory nerve fibers. A weakness of these theories is that the firing rate of auditory nerve fibers as well as of most neurons in the cochlear nucleus saturate at high levels (Sachs and Young, 1979; Cedolin and Delgutte, 2005). To address this problem, it has been proposed that the spatial profile is first sharpened by lateral inhibition, prior to template matching (Shamma, 1985). This preprocessing step enhances the responses at places where the phase changes rapidly, which occurs where the BM is tuned to the sound's frequency. A recent analysis of cat auditory nerve responses has shown that such preprocessing produces spatial profiles from which f0 can indeed be extracted even at high levels (Cedolin and Delgutte, 2010), although a more recent analysis (in guinea pigs and with different methods) suggested that the estimated f0 is very sensitive to level (Carlyon et al., 2012). Because this preprocessing step relies on temporal cues, template-based models of pitch using this stage as input are often described as spatiotemporal models (Cedolin and Delgutte, 2010). However, these are very different from the structural theory we have presented, as they are in fact models based on matching spatial templates where temporal information is discarded, only with an input that is obtained from a spatiotemporal transformation of the auditory nerve response. In contrast, matching in the structural theory as well as in the two related models mentioned above and in standard temporal models is performed on the entire temporal signals.

Unlike the structural theory, none of these three models addresses the pitch of unresolved harmonic complexes.

The nature of pitch in theories of pitch

In standard temporal theories of pitch, pitch is the perceptual correlate of the periodicity of the acoustical waveform. Independently of how the periodicity is physiologically extracted, this proposition implies for example that: periodic sounds have a pitch, non-periodic sounds do not have pitch, and pitch saliency is related to how close to periodic a sound is. It also implies that two sounds with the same periodicity are similar, and that two sounds with fundamental frequencies differing by an octave are similar, in the sense that they have a periodicity in common. Thus, this characterization of pitch entails a particular region of existence of pitch (what sounds produce pitch) and a particular topology of pitch (how pitch-evoking sounds relate to each other). These two aspects do not rely on learning, in the sense that they do not depend on the specific sounds the auditory system is exposed to. Instead, they derive from the existence of a general mechanism that identifies periodicity.

In a similar way, the structural theory of pitch defines pitch as the perceptual correlate of the regularity structure of the BM vibration pattern. It also entails an existence region of pitch, which is more complex than in temporal theories, and a particular topology of pitch, which is similar to that implied by temporal theories (but see below for the effect of level on pitch). In the same way, these two aspects do not rely on learning.

In standard place theories of pitch based on templates, what characterizes pitch-evoking sounds is that they are similar to some internal template (Terhardt, 1974). Thus pitch is the perceptual correlate of a particular category of sounds, which is formed by previous exposure to pitch-evoking sounds. There is an obvious problem of circularity in this characterization, which means that in addition to exposure to the sounds, these sounds must be labeled as having or not having a pitch. That is, pitch is characterized independently of the sounds themselves. An example would be that vocalizations are those special sounds that are considered as producing pitch. Accordingly, a more rigorous characterization of pitch in place theories is the following: pitch is the perceptual correlate of spectral similarity to vocalizations (or any other externally defined category of sounds).

This characterization is problematic for several reasons. First, it defines an existence region of pitch but not a topology of pitch, unless the spatial activation profiles produced by sounds with the same pitch are similar. This issue might be addressed to some extent by spatial sharpening as previously mentioned (Shamma, 1985), although there is no indication that such an operation occurs in the auditory system. A second problem is that not all pitch-evoking sounds are spectrally similar to vocalizations, for example low-frequency pure tones. Finally, infants have a sense of musical pitch (Montgomery and Clarkson, 1997). The latter two issues have been addressed in a model in which harmonic templates are learned from inharmonic sounds (Shamma and Klein, 2000). Indeed auditory nerve fibers with harmonically related CFs are expected to fire with some degree of correlation in response to noise, because of nonlinearities in their response. Thus a Hebbian mechanism could form harmonic templates by selecting temporally correlated fibers. In this scheme, pitch is then the perceptual correlate of the similarity between the places of activation on the BM and places that are generally expected to be correlated.

In addition to the fact that this only addresses the pitch of unresolved harmonic complexes, this proposition is somehow paradoxical. On one hand, the formation of internal templates critically relies on the temporal fine structure of the sounds, and fine correlations between channels. Indeed in Hebbian models, the learning signal is the correlation between input and output (pre- and postsynaptic neurons), and therefore it requires that the output firing is sensitive to input correlations. On the other hand, pitch estimation by template matching assumes that this temporal fine structure is then entirely discarded: only average spectrum is considered, and correlations between channels (relative phases of harmonics in a complex tone) are assumed to have no effect on pitch. To reconcile the two aspects of the model requires either that the neurons are initially sensitive to input correlations and become insensitive to them after a critical period (after learning), or that learning is based on input correlations but not through a Hebbian mechanism (i.e., not involving input-ouput correlations).

Experimental predictions

We can formulate two types of predictions, for psychophysical experiments and for physiological experiments. The strongest psychophysical prediction concerns the effect of level on pitch. The phase of the BM response to tones depends on level (Robles and Ruggero, 2001), because of nonlinear effects. Consequently, cross-channel structure should depend on level. However, within-channel structure should not depend on level because such nonlinearities have no effect on periodicity. If we assume that sounds are matched in pitch when they produce some common regularity structure on the BM, then a pitch-matching experiment between sounds with different levels should reveal an effect of level on the pitch of sounds that produce cross-channel structure but not within-channel structure. According to our analysis, these are pure tones of low frequency, i.e., with period larger than the maximum delay. The few studies on such effects support this prediction (Morgan et al., 1951; Verschuure and Van Meeteren, 1975; Burns, 1982), but a more exhaustive and controlled study would be required.

Predictions for physiological experiments can be made for specific hypotheses about the neural mechanism. For example, low-frequency spherical bushy cells are primary-like neurons of the cochlear nucleus with strong phase locking properties (Joris et al., 1994; Fontaine et al., 2013) (possibly stronger than the auditory nerve), and their pattern of synchrony in response to sounds could then reflect the regularity structure of the BM vibration. The prediction is then that the synchrony receptive field of two such cells, defined as the set of sounds that produce synchronous responses in the two cells (Brette, 2012), should consist of pitch-evoking sounds - in fact of a pure tone of specific frequency. Ideally, such recordings should be done simultaneously, because shared variability (e.g. due to local synaptic connections or shared modulatory input) affects phase locking and reproducibility but not synchrony (Brette, 2012).

References

Agmon-Snir H, Carr CE, Rinzel J (1998) The role of dendrites in auditory coincidence detection. Nature 393:268–272.

Bernstein JG, Oxenham AJ (2003) Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number? The Journal of the Acoustical Society of America 113:3323.

Boer E de, Jongh HR de (1978) On cochlear encoding: Potentialities and limitations of the reverse-correlation technique. The Journal of the Acoustical Society of America 63:115–135.

Brette R (2012) Computing with neural synchrony. PLoS computational biology 8:e1002561.

Burns EM (1982) Pure-tone pitch anomalies. I. Pitch-intensity effects and diplacusis in normal ears. The Journal of the Acoustical Society of America 72:1394.

Cariani PA (2001) Neural timing nets. Neural Netw 14:737.

Cariani PA, Delgutte B (1996a) Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. J Neurophysiol 76:1698.

Cariani PA, Delgutte B (1996b) Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch. J neurophysiol 76:1717.

Carlyon RP (1998) Comments on “A unitary model of pitch perception” [J. Acoust. Soc. Am. 102, 1811–1820 (1997)]. The Journal of the Acoustical Society of America 104:1118–1121.

Carlyon RP, Long CJ, Micheyl C (2012) Across-Channel Timing Differences as a Potential Code for the Frequency of Pure Tones. JARO 13:159–171.

Carlyon RP, Shackleton TM (1994) Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms? The Journal of the Acoustical Society of America 95:3541–3554.

Carney LH, Heinzy MG, Evilsizer ME, Gilkeyz RH, Colburn HS (2002) Auditory phase opponency: A temporal model for masked detection at low frequencies. Acta Acustica United with Acustica 88:334–347.

Carney LH, Yin TC (1988) Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model. J Neurophysiol 60:1653–1677.

Cedolin L, Delgutte B (2005) Pitch of Complex Tones: Rate-Place and Interspike Interval Representations in the Auditory Nerve. J Neurophysiol 94:347–362.

Cedolin L, Delgutte B (2010) Spatiotemporal Representation of the Pitch of Harmonic Complex Tones in the Auditory Nerve. J Neurosci 30:12712–12724.

Couchman K, Grothe B, Felmy F (2010) Medial Superior Olivary Neurons Receive Surprisingly Few Excitatory and Inhibitory Inputs with Balanced Strength and Short-Term Dynamics. The Journal of Neuroscience 30:17111 –17121.

De Cheveigné A (2010) Pitch perception. The oxford handbook of auditory science: Hearing:71–104.

Fontaine B, Benichoux V, Joris PX, Brette R (2013) Predicting spike timing in highly synchronous auditory neurons at different sound levels. J Neurophysiol:jn.00051.2013.

Fontaine B, Goodman DFM, Benichoux V, Brette R (2011) Brian Hears: Online Auditory Processing Using Vectorization Over Channels. Front Neuroinform 5 Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3143729/ [Accessed November 2, 2013].

Fontaine B, Pena JL, Brette R (2014) Spike-threshold adaptation predicted by membrane potential dynamics in vivo. PLoS Comput Biol 10:e1003560.

Gerstner W, Naud R (2009) How Good Are Neuron Models? Science 326:379–380.

Glasberg BR, Moore BC. (1990) Derivation of auditory filter shapes from notched-noise data. Hearing Research 47:103–138.

GREEN DM, SWETS JA (1966) SIGNAL DETECTION THEORY AND PSYCHOPHYSICS. Oxford, England: John Wiley.

Houtsma AJM, Smurzynski J (1990) Pitch identification and discrimination for complex tones with many harmonics. The Journal of the Acoustical Society of America 87:304–310.

Jolivet R, Lewis TJ, Gerstner W (2004) Generalized integrate-and-fire models of neuronal activity approximate spike trains of a detailed model to a high degree of accuracy. J Neurophysiol 92:959–976.

Joris PX, Carney LH, Smith PH, Yin TC (1994) Enhancement of neural synchronization in the anteroventral cochlear nucleus. I. Responses to tones at the characteristic frequency. J neurophysiol 71:1022.

Karino S, Smith PH, Yin TCT, Joris PX (2011) Axonal Branching Patterns as Sources of Delay in the Mammalian Auditory Brainstem: A Re-Examination. J Neurosci 31:3016–3031.

Khurana S, Remme MWH, Rinzel J, Golding NL (2011) Dynamic Interaction of Ih and IK-LVA During Trains of Synaptic Potentials in Principal Neurons of the Medial Superior Olive. J Neurosci 31:8936–8947.

Langner G (1992) Periodicity coding in the auditory system. Hearing Research 60:115–142.

Laudanski J, Coombes S, Palmer AR, Sumner CJ (2010) Mode-Locked Spike Trains in Responses of Ventral Cochlear Nucleus Chopper and Onset Neurons to Periodic Stimuli. Journal of Neurophysiology 103:1226–1237.

Licklider JCR (1951) A duplex theory of pitch perception. Experientia 7:128.

Loeb GE, White MW, Merzenich MM (1983) Spatial cross-correlation. Biol Cybern 47:149–163.

Meddis R, O'Mard L (1997) A unitary model of pitch perception. The Journal of the Acoustical Society of America 102:1811.

Micheyl C, Oxenham AJ (2007) Across-frequency pitch discrimination interference between complex tones containing resolved harmonics. The Journal of the Acoustical Society of America 121:1621.

Montgomery CR, Clarkson MG (1997) Infants' pitch perception: Masking by low- and high-frequency noises. The Journal of the Acoustical Society of America 102:3665–3672.

Morgan CT, Garner WR, Galambos R (1951) Pitch and Intensity. The Journal of the Acoustical Society of America 23:658–663.

Oxenham AJ (2012) Pitch Perception. J Neurosci 32:13335–13338.

Oxenham AJ, Bernstein JG, Penagos H (2004a) Correct tonotopic representation is necessary for complex pitch perception. Proceedings of the National Academy of Sciences of the United States of America 101:1421–1425.

Oxenham AJ, Bernstein JGW, Penagos H (2004b) Correct Tonotopic Representation Is Necessary for Complex Pitch Perception. PNAS 101:1421–1425.

Oxenham AJ, Micheyl C, Keebler MV, Loper A, Santurette S (2011) Pitch perception beyond the traditional existence region of pitch. Proceedings of the National Academy of Sciences 108:7629–7634.

Palmer AR, Shackleton TM (2009) Variation in the Phase of Response to Low-Frequency Pure Tones in the Guinea Pig Auditory Nerve as Functions of Stimulus Level and Frequency. J Assoc Res Otolaryngol 10:233–250.

Platkiewicz J, Brette R (2010) A threshold equation for action potential initiation. PLoS Comput Biol 6:e1000850.

Platkiewicz J, Brette R (2011) Impact of fast sodium channel inactivation on spike threshold dynamics and synaptic integration. PLoS Comput Biol 7:e1001129.

Pressnitzer D, Patterson RD, Krumbholz K (2001) The lower limit of melodic pitch. The Journal of the Acoustical Society of America 109:2074–2084.

Ritsma RJ (1962) Existence Region of the Tonal Residue. I. The Journal of the Acoustical Society of America 34:1224–1229.

Robles L, Ruggero MA (2001) Mechanics of the mammalian cochlea. Physiol Rev 81:1305–1352.

Rossant C, Leijon S, Magnusson AK, Brette R (2011) Sensitivity of noisy neurons to coincident inputs. The Journal of Neuroscience 31:17193–17206.

Rothman JS, Manis PB (2003a) The Roles Potassium Currents Play in Regulating the Electrical Activity of Ventral Cochlear Nucleus Neurons. J Neurophysiol 89:3097–3113.

Rothman JS, Manis PB (2003b) The roles potassium currents play in regulating the electrical activity of ventral cochlear nucleus neurons. Journal of neurophysiology 89:3097–3113.

Rushton WAH (1951) A theory of the effects of fibre size in medullated nerve. J Physiol 115:101–122.

Sachs MB, Abbas PJ (1974) Rate versus level functions for auditory-nerve fibers in cats: tone-burst stimuli. The Journal of the Acoustical Society of America 56:1835–1847.

Sachs MB, Young ED (1979) Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. The Journal of the Acoustical Society of America 66:470–479.

Semal C, Demany L (1990) The Upper Limit of “Musical” Pitch. Music Perception: An Interdisciplinary Journal 8:165–175.

Shamma S, Klein D (2000) The case of the missing pitch templates: How harmonic templates emerge in the early auditory system. The Journal of the Acoustical Society of

America 107:2631–2644.

Shamma SA (1985) Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. The Journal of the Acoustical Society of America 78:1622.

Slaney M (1993) An efficient implementation of the Patterson-Holdsworth auditory filter bank. Apple Computer, Perception Group, Tech Rep Available at: http://rvl4.ecn.purdue.edu/~malcolm/apple/tr35/PattersonsEar.pdf [Accessed November 2, 2013].

Stevens SS (1971) Sensory Power Functions and Neural Events. In: Principles of Receptor Physiology (Loewenstein WR, ed), pp 226–242 Handbook of Sensory Physiology. Springer Berlin Heidelberg. Available at: http://link.springer.com/chapter/10.1007/978-3-642-65063-5_7 [Accessed November 2, 2013].

Terhardt E (1974) Pitch, consonance, and harmony. J Acoust Soc Am 55:1061–1069.

Trussell LO (1997) Cellular mechanisms for preservation of timing in central auditory pathways. Curr Opin Neurobiol 7:487–492.

Trussell LO (1999) SYNAPTIC MECHANISMS FOR CODING TIMING IN AUDITORY NEURONS. Annu Rev Physiol 61:477–496.

Verschuure J, Van Meeteren AA (1975) The effect of intensity on pitch. Acta Acustica united with Acustica 32:33–44.

Yin TC, Chan JC (1990) Interaural time sensitivity in medial superior olive of cat. J Neurophysiol 64:465–488.

Zhang X, Heinz MG, Bruce IC, Carney LH (2001) A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. The Journal of the Acoustical Society of America 109:648–670.

Zwislocki JJ (1973) On intensity characteristics of sensory receptors: A generalized function. Kybernetik 12:169–183.

Figures

Figure 1. Regularity structure of the basilar membrane (BM) vibration pattern. (A) Vibration of the basilar membrane produced by a periodic sound S(x,t) (clarinet musical note), at places x tuned to different frequencies (modeled by band-pass filters). (B) The vibration at one place is transformed into spikes produced by an auditory nerve fiber (bottom : post-stimulus time histogram of spikes). In Licklider's model, the fiber projects to a coincidence detector neuron through two axons with conduction delays differing by δ. The neuron fires maximally when the signal's periodicity T equals δ. (C) If the signal's period T is smaller than the neuron's refractory time, then the neuron must detect coincidences between spikes coming from different fibers. (D) If the fibers originate from slightly different places x and y on the cochlea, then the neuron responds to similarities between BM vibrations at different places. (E) Vibration pattern of the BM produced by a non-periodic sound (noise): there is no regularity structure across place and time. (F) Vibration pattern produced by a musical note: there are signal similarities across time (horizontal arrows) and place (oblique arrow).

Figure 2. Harmonic resolvability and cross-channel structure. (A) Amplitude and phase spectrum of two gammatone filters. Only a pure tone of frequency f ("Input" waveform) is attenuated in the same way by the two filters (red and blue waveforms: filter outputs). At that frequency, the delay between the outputs of the two filters is $\delta = \Delta\phi/f$. (B) If several harmonic components fall within the bandwidths of the two filters, then the outputs of the two filters differ (no cross-channel similarity). (C) Excitation pattern produced on the cochlea by a harmonic complex. Top: amplitude vs. center frequency of gammatone filters; bottom: spectrum of harmonic complex and of gammatone filters. Harmonic components are "resolved" when they can be separated on the cochlear activation pattern. Higher frequency components are unresolved because cochlear filters are broader. (D) Resolved components produce cross-channel similarity between many pairs of filters (as in A). Unresolved components produce little cross-channel structure (as in B). (E) Thus the vibration pattern produced by resolved components displays both within-channel and cross-channel structure (left), while unresolved components only produce within-channel structure (right).

Figure 3. Domain of existence of pitch. (A) Within-channel structure produced by a periodic sound can be decoded if the sound's period is smaller than the maximal neural delay δ_{max}. With δ_{max} = 4 ms, it occurs for sounds of fundamental frequency greater than 250 Hz. (B) A pure tone or resolved harmonic produces cross-channel structure with arbitrarily small delays between channels, corresponding to the phase difference between the two filters at the sound's frequency: here a 100 Hz tone produces two identical waveforms delayed by δ = 2 ms, while the sound's period is 10 ms. (C) A transposed tone with a high-frequency carrier (>4 kHz) modulated by a low-frequency envelope (<320 Hz) elicits a very weak pitch (Oxenham et al., 2004a) (top: f0 = 120 Hz). Such sounds produce only within-channel structure because they only have high-frequency content (middle). The structural theory of pitch predicts an absence of pitch when the envelope's periodicity is larger than δ_{max}, which is consistent with psychophysics if δ_{max}< 3 ms. (D) A pure tone with the same fundamental frequency (f0 = 120 Hz) produces cross-channel structure with short delays. The structural theory of pitch predicts the existence of pitch in this case, consistently with psychophysical results (Oxenham et al., 2004a). (E) Complex tones with f0 between 400 Hz and 2 kHz and all harmonics above 5 kHz elicit a pitch (Oxenham et al., 2011) (top, spectrum of a complex tone; middle, temporal waveform). Such tones produce only within-channel structure in high-frequency (bottom), and the structural

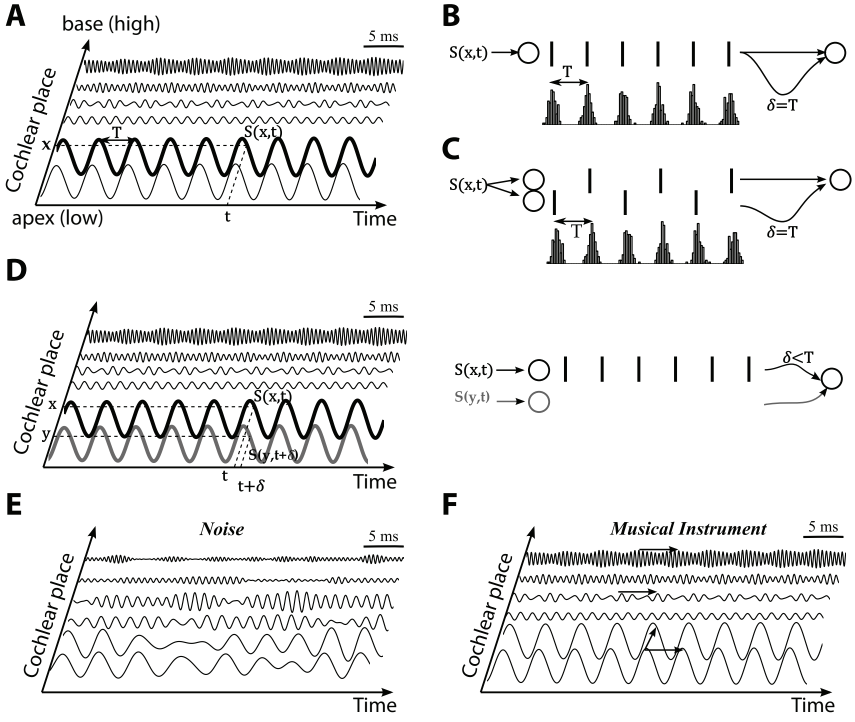
theory of pitch predicts the existence of pitch if the sound's period is smaller than δ_{max}, which is consistent with psychophysics if $\delta_{max} > 2.5$ ms.

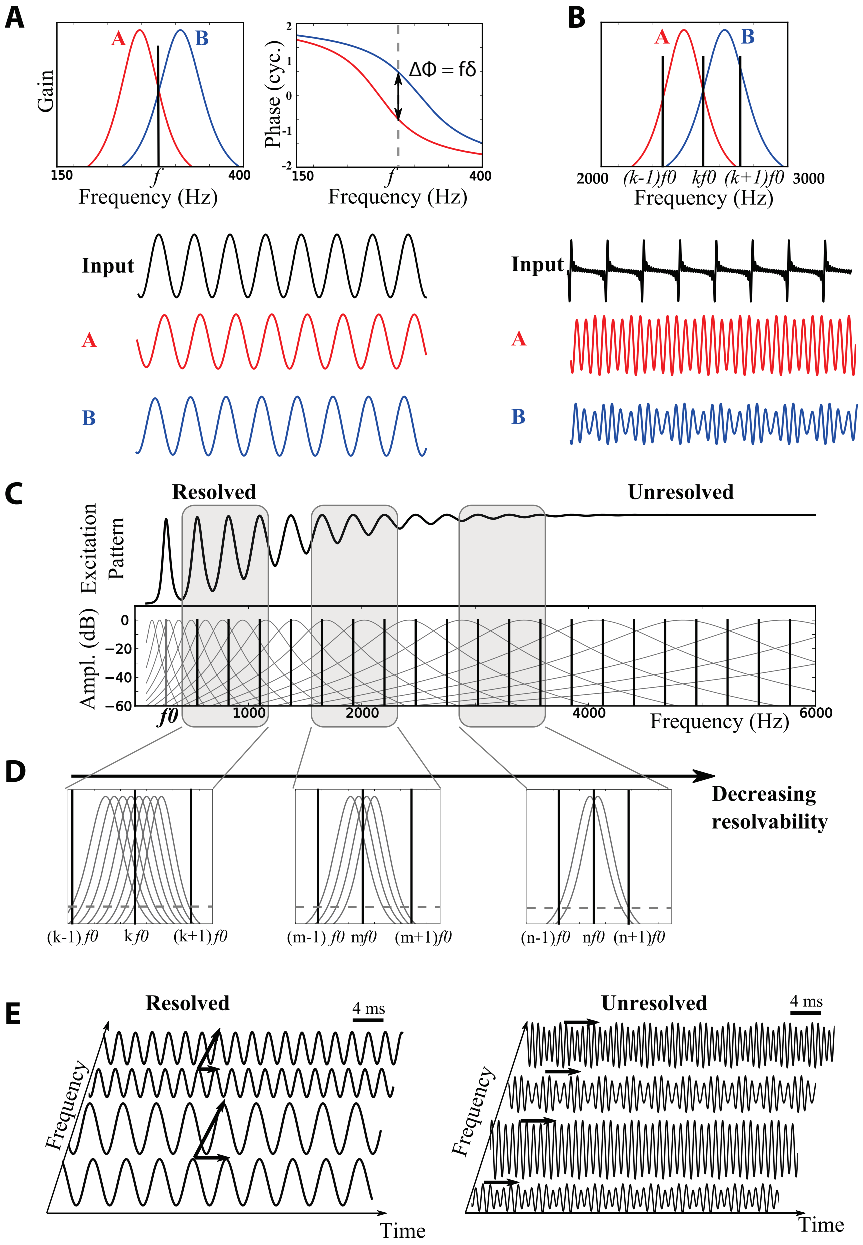
Figure 4. Neural network model of pitch estimation using within- and cross-channel structure. (A) Spectrogram of a trumpet sound showing the first two harmonics. Two neurons with CF around the first harmonic and input delay δ receive the same signal (red and blue rectangles and input signals below). As a result, the two neurons fire synchronously, for all 3 neuron models used: biophysical model of chopper and octopus cells, leaky integrate-and-fire model (voltage traces). (B) Spectrogram of a rolling sea wave sound, which shows no regularity structure. In particular, the two neurons do not receive the same signals (input, shaded area: difference between the two signals) and thus do not fire synchronously. (C) Spectrogram of a harpsichord sound with unresolved harmonics in high frequency. The inset shows the periodicity of the envelope. Two neurons fire synchronously if they receive inputs from the same place delayed by $\delta = 1/f0$. (D) In the same high frequency region, the inharmonic sound of a sea wave does not produce within-channel structure and therefore the two neurons do not fire synchronously. (E) Synaptic connections for a pitch-selective group tuned to f0 = 220 Hz. Harmonics are shown on the left (red comb) superimposed on auditory filters. Resolved harmonics (bottom) produce regularity structure both across and within channels: color saturation represents the amplitude of the filter output while hue represents its phase, for different delays (horizontal axis) and characteristic frequencies (vertical axis). Neurons with the same color fire synchronously and project to a common neuron. Unresolved harmonics (top) produce regularity structure within channels only. Here two identical colors correspond to two identical input signals only when the neurons have identical CF (same row). (F) Same as (E) for a f0 = 261 Hz, producing a different regularity structure, corresponding to a different synchrony pattern in input neurons. Synchronous neurons project to another group of neurons, selective for this pitch.

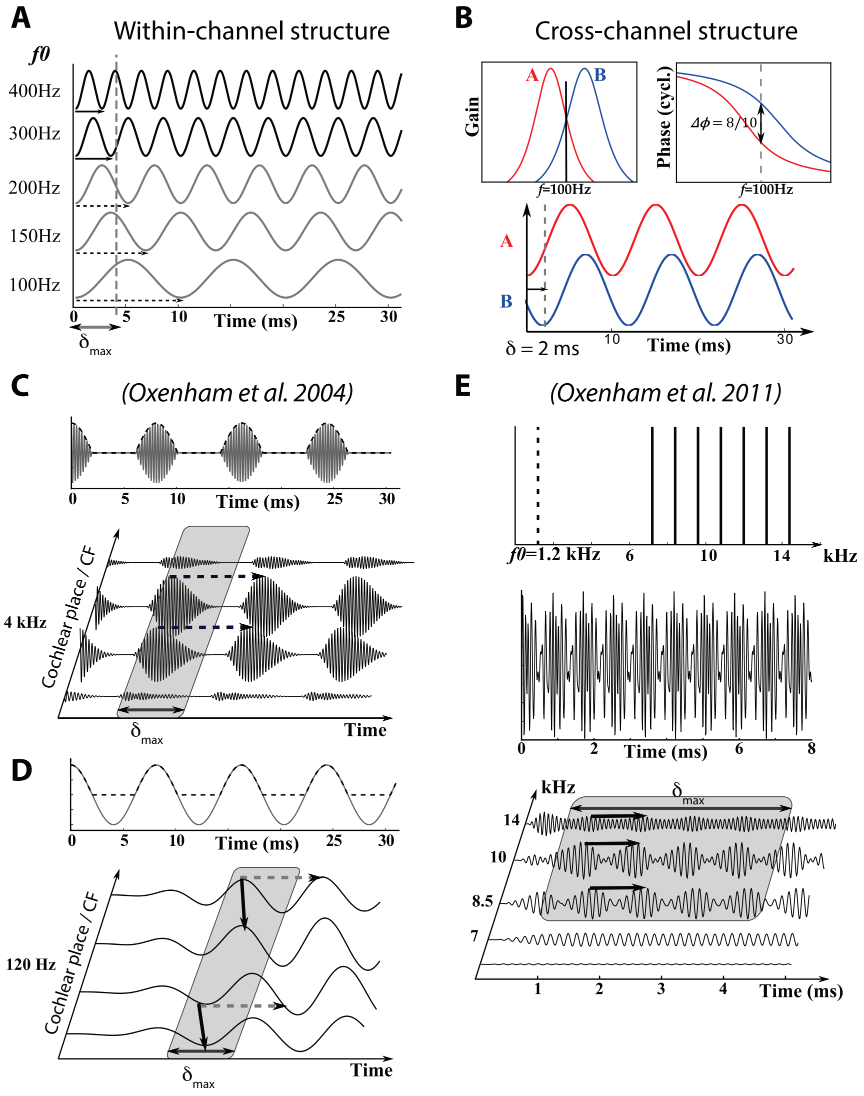
Figure 5. Pitch recognition by a neural network model based on the structural theory. (A) Top, Spectrogram of a sequence of sounds, which are either either environmental noises (inharmonic) or musical notes of the chromatic scale (A3-A4) played by different instruments. Bottom, Firing rate of all pitch-specific neural groups responding to these sounds (vertical axis: preferred pitch, A3-A4). (B) Distribution of firing rates of pitch-specific groups for instruments played at the preferred pitch (blue) and for noises (grey), for 3 different sound levels. (C) Top, Pitch recognition scores of the model (horizontal axis: error in semitones) on a set of 762 notes between A2 and A4, including 41 instruments (587 notes) and 5 sung vowels (175 notes). Bottom, Firing rate of all pitch-specific groups as a function of the difference between presented f0 and preferred f0, for all sounds (solid black: average). Peaks appear at octaves (12 semitones) and perfect fifths (7 semitones). (D) Impact of the number of frequency channels (top) and maximal delay δ_{max} (bottom) on recognition performance.

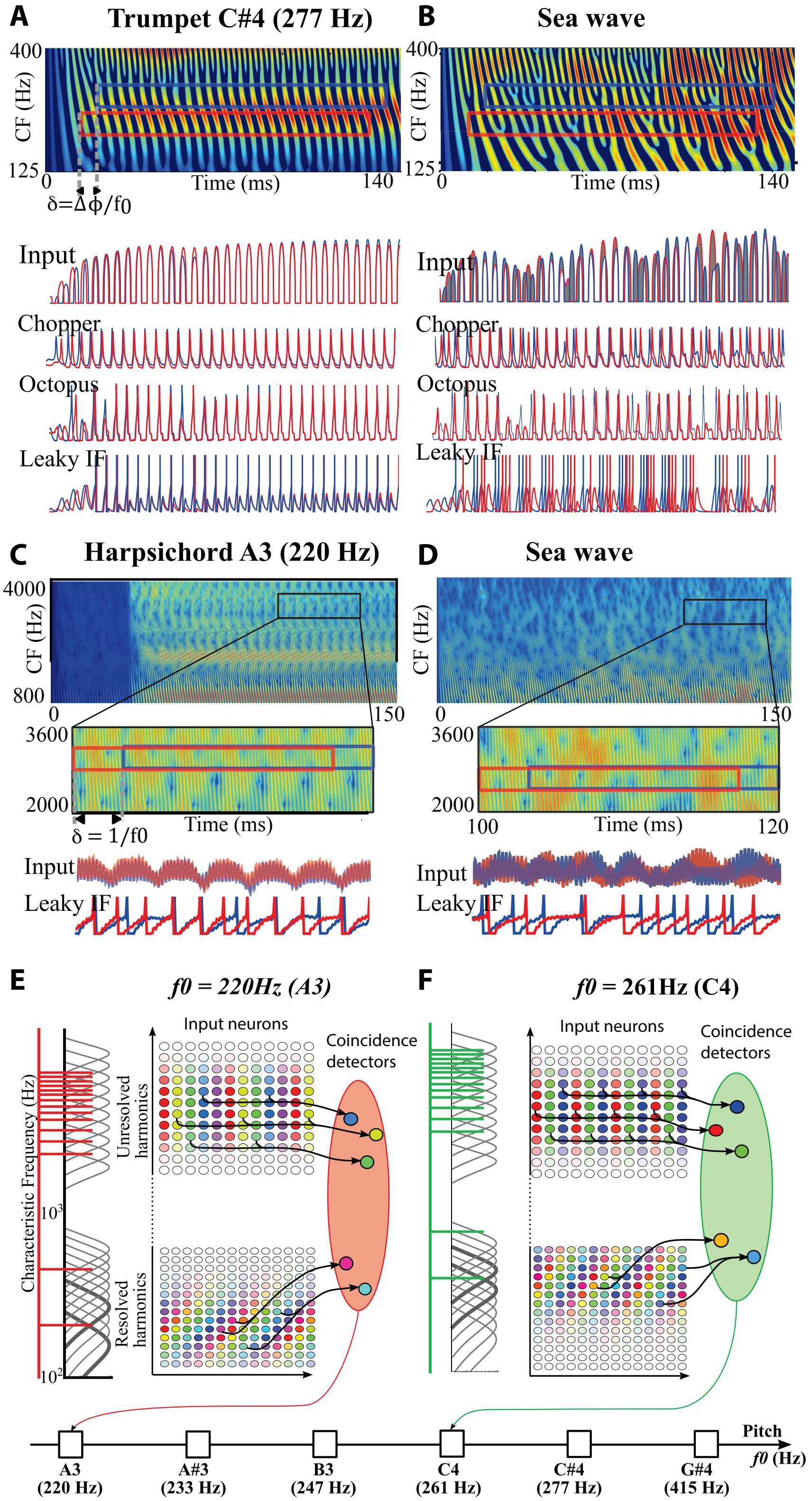
Figure 6. Pitch discriminability. (A) Two neurons tuned to the same frequency (within-channel) but with delay mismatch $\delta = 1/f$ produce phase-locked spikes (red and blue crosses) in response to a tone (sine waves). When the tone frequency is f (left), the two input signals match and the difference of phases of spikes $\Delta\Phi(f)$ between the two neurons is distributed around 0 (shaded curve). When the tone frequency is f+df (right), the two signals are slightly mismatched and the distribution of $\Delta\Phi(f)$ is not centered on 0. (B) Two neurons tuned to different frequencies (cross-channel) respond at different mean phases to tones (red and blue curves). (C) The discriminability index d' is defined as the distance μ between the centers of be two phase difference distributions ($\Delta\Phi(f)$ and $\Delta\Phi(f+df)$) relative to their standard deviation σ. (D) The

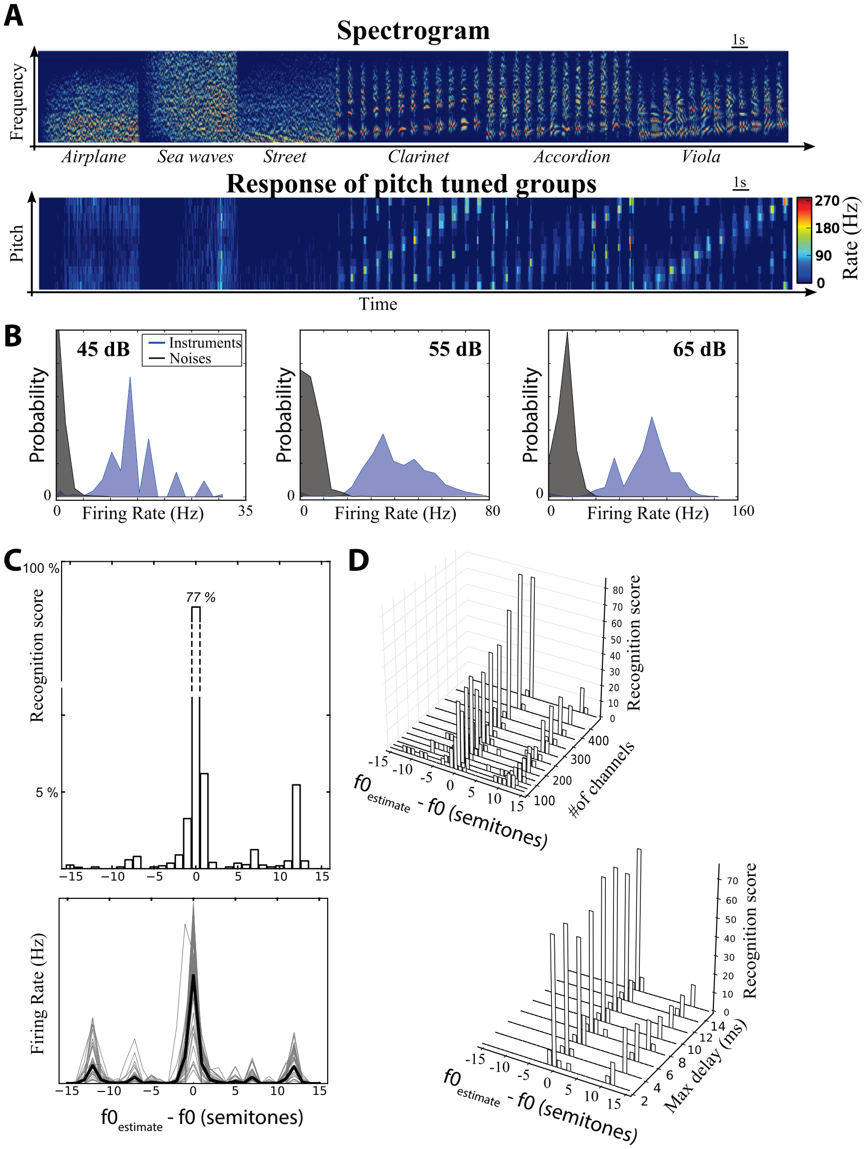
standard deviation of the phase distribution is related to the precision of phase locking, measured by the vector strength (dots: vector strength vs. characteristic frequency for guinea pig auditory fibers; solid curve: fit). (E) Mean phase of spikes produced by auditory nerve fibers of guinea pigs for different tone frequencies (data from Palmer and Shackleton (2009) (Palmer and Shackleton, 2009)), as a function of CF (crosses), with fits (solid lines). (F) Weber fraction ($\Delta f/f$, where Δf is the just noticeable difference in frequency) as a function of tone frequency for cross-channel structure (colored curves) and within-channel structure (black curve). Color represent different frequency spacings between the two channels (1-6 semitones). Dotted lines represent the limitations implied by a maximal delay δ_{max} = 5 ms.

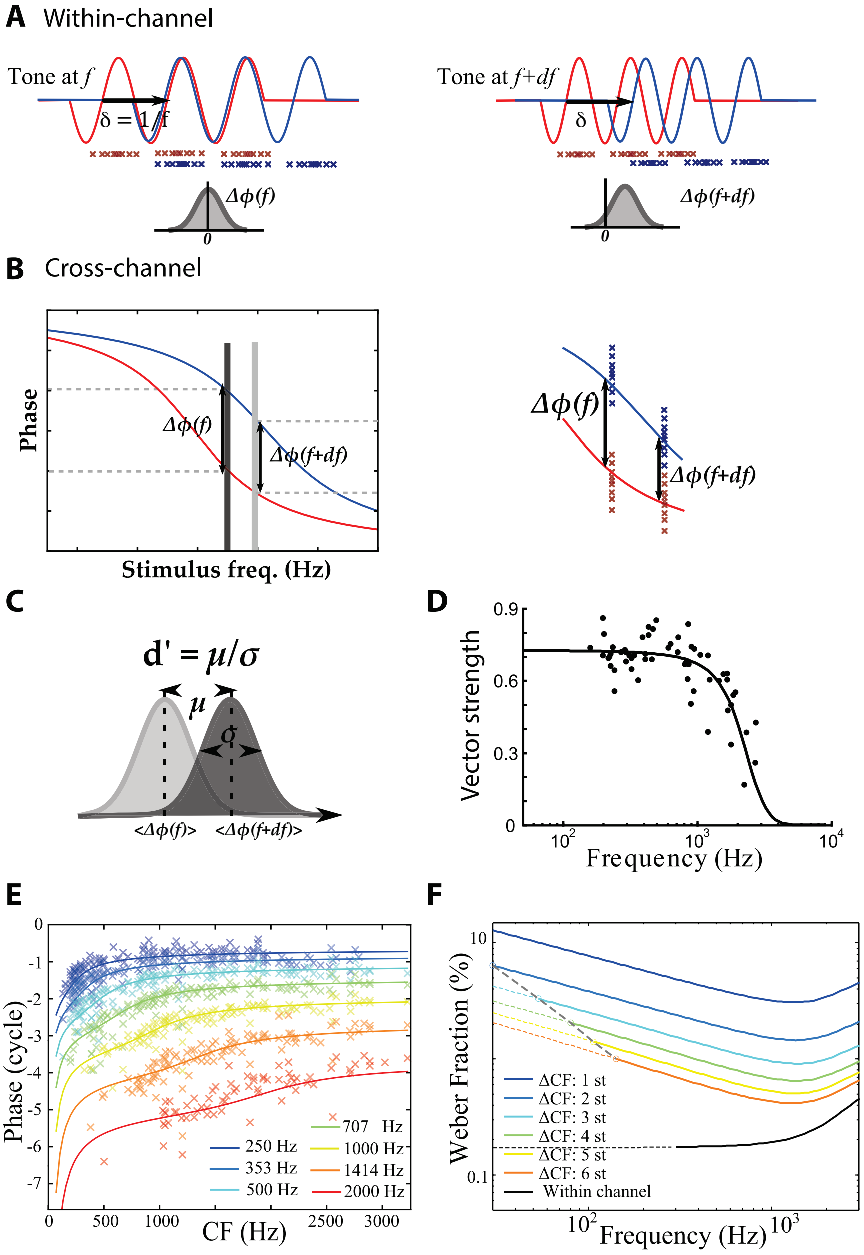

A
base (high)
5 ms
Cochlear place
x
T
S(x,t)
apex (low)
t
Time
B
S(x,t)
T
δ=T
C
S(x,t)
T
δ=T
D
5 ms
Cochlear place
x
y
S(x,t)
S(y,t+δ)
t
t+δ
Time
S(x,t)
S(y,t)
δ<T
E
Noise
5 ms
Cochlear place
Time
F
Musical Instrument
5 ms
Cochlear place
Time


A
Gain
A
B
150
f
400
Frequency (Hz)
Phase (cyc.)
2
1
0
-1
-2
150
f
400
ΔΦ = fδ
Frequency (Hz)
B
A
B
2000
(k-1)f0
kf0
(k+1)f0
3000
Frequency (Hz)
Input
A
B
Input
A
B
C
Resolved
Unresolved
Excitation Pattern
Ampl. (dB)
0
−20
−40
−60
f0
1000
2000
4000
Frequency (Hz)
6000
D
Decreasing resolvability
(k-1) f0
k f0
(k+1) f0
(m-1) f0
mf0
(m+1)f0
(n-1)f0
nf0
(n+1)f0
E
Resolved
4 ms
Frequency
Time
Unresolved
4 ms
Frequency
Time


A
Within-channel structure
f0
400Hz
300Hz
200Hz
150Hz
100Hz
0
5
10
Time (ms)
25
30
δ max
B
Cross-channel structure
Gain
A
B
f=100Hz
Phase (cycl.)
Δφ = 8/10
f=100Hz
A
B
δ = 2 ms
10
Time (ms)
30
C
(Oxenham et al. 2004)
0
5
10
Time (ms)
25
30
Cochlear place / CF
4 kHz
δ max
Time
D
0
5
10
Time (ms)
25
30
Cochlear place / CF
120 Hz
δ max
Time
E
(Oxenham et al. 2011)
f0=1.2 kHz
6
10
14
kHz
0
2
Time (ms)
6
8
kHz
δ max
14
10
8.5
7
1
2
3
4
Time (ms)


A
Trumpet C#4 (277 Hz)
400
CF (Hz)
125
0
Time (ms)
140
δ=Δϕ/f0
B
Sea wave
400
CF (Hz)
125
0
Time (ms)
140
Input
Chopper
Octopus
Leaky IF
Input
Chopper
Octopus
Leaky IF
C
Harpsichord A3 (220 Hz)
4000
CF (Hz)
800
0
150
3600
2000
δ = 1/f0
Time (ms)
Input
Leaky IF
D
Sea wave
CF (Hz)
0
150
3600
2000
100
Time (ms)
120
Input
Leaky IF
E
f0 = 220Hz (A3)
Characteristic Frequency (Hz)
10³
10²
Input neurons
Coincidence detectors
Unresolved harmonics
Resolved harmonics
F
f0 = 261Hz (C4)
Input neurons
Coincidence detectors
Pitch
f0 (Hz)
A3 (220 Hz)
A#3 (233 Hz)
B3 (247 Hz)
C4 (261 Hz)
C#4 (277 Hz)
G#4 (415 Hz)


A
Spectrogram
1s
Frequency
Airplane
Sea waves
Street
Clarinet
Accordion
Viola
Response of pitch tuned groups
1s
Pitch
Time
270
180
90
0
Rate (Hz)
B
45 dB
Instruments
Noises
Probability
0
0
Firing Rate (Hz)
35
55 dB
Probability
0
0
Firing Rate (Hz)
80
65 dB
Probability
0
0
Firing Rate (Hz)
160
C
100 %
77 %
Recognition score
5 %
−15 −10 −5 0 5 10 15
Firing Rate (Hz)
−15 −10 −5 0 5 10 15
f0 estimate - f0 (semitones)
D
80 70 60 50 40 30 20 10 0
Recognition score
-15 -10 -5 0 5 10 15
100 200 300 400
#of channels
f0 estimate - f0 (semitones)
70 60 50 40 30 20 10 0
Recognition score
-15 -10 -5 0 5 10 15
2 4 6 8 10 12 14
Max delay (ms)
f0 estimate - f0 (semitones)


A Within-channel
Tone at f
δ = 1/f
Δϕ(f)
0
Tone at f+df
δ
Δϕ(f+df)
0
B Cross-channel
Phase
Δϕ(f)
Δϕ(f+df)
Stimulus freq. (Hz)
Δϕ(f)
Δϕ(f+df)
C
d' = μ/σ
μ
σ
<Δϕ(f)>
<Δϕ(f+df)>
D
Vector strength
0.9
0.6
0.3
0
10^2
10^3
10^4
Frequency (Hz)
E
Phase (cycle)
0
-1
-2
-3
-4
-5
-6
-7
0
500
1000
2500
3000
CF (Hz)
250 Hz
353 Hz
500 Hz
707 Hz
1000 Hz
1414 Hz
2000 Hz
F
Weber Fraction (%)
10
0.1
10^2
10^3
Frequency (Hz)
ΔCF: 1 st
ΔCF: 2 st
ΔCF: 3 st
ΔCF: 4 st
ΔCF: 5 st
ΔCF: 6 st
Within channel