Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception

Mark R. Saddler[1,2,3,*], Ray Gonzalez[1,2,3,*], Josh H. McDermott[1,2,3,4]

1 Department of Brain and Cognitive Sciences, MIT
2 McGovern Institute for Brain Research, MIT
3 Center for Brains, Minds and Machines, MIT
4 Program in Speech and Hearing Biosciences and Technology, Harvard University
* co-first authors

ABSTRACT

Computations on receptor responses enable behavior in the environment. Behavior is plausibly shaped by both the sensory receptors and the environments for which organisms are optimized, but their roles are often opaque. One classic example is pitch perception, whose properties are commonly linked to peripheral neural coding limits rather than environmental acoustic constraints. We trained artificial neural networks to estimate fundamental frequency from simulated cochlear representations of natural sounds. The best-performing networks replicated many characteristics of human pitch judgments. To probe how our ears and environment shape these characteristics, we optimized networks given altered cochleae or sound statistics. Human-like behavior emerged only when cochleae had high temporal fidelity and when models were optimized for natural sounds. The results suggest pitch perception is critically shaped by the constraints of natural environments in addition to those of the cochlea, illustrating the use of contemporary neural networks to reveal underpinnings of behavior.

INTRODUCTION

A key goal of perceptual science is to understand why sensory-driven behavior takes the form that it does. In some cases, it is natural to relate behavior to physiology, and in particular to the constraints imposed by sensory transduction. For instance, color discrimination is limited by the number of cone types in the retina (Wandell, 1995). Olfactory discrimination is similarly constrained by the receptor classes in the nose (Hildebrand and Shepherd, 1997). In other cases, behavior can be related to properties of environmental stimulation that are largely divorced from the constraints of peripheral transduction. For example, face recognition in humans is much better for upright faces, presumably because we predominantly encounter upright faces in our environment (Yin, 1969).

Understanding how physiological and environmental factors shape behavior is important both for fundamental scientific understanding and for practical applications such as sensory prostheses, the engineering of which might benefit from knowing how sensory encoding constrains behavior. Yet the constraints on behavior are often difficult to pin down. For instance, the auditory periphery encodes sound with exquisite temporal fidelity (Rose et al., 1967; Palmer and Russell, 1986), but the role of this information in hearing remains controversial (Attneave and Olson, 1971; Javel and Mott, 1988; Jacoby et al., 2019). Part of the challenge is that the requisite experiments – altering sensory receptors or environmental conditions during evolution or development, for instance – are practically difficult (and ethically unacceptable in humans).

The constraints on behavior can sometimes instead be revealed by computational models. Ideal observer models, which optimally perform perceptual tasks given particular sensory inputs and sensory receptor responses, have been the method of choice for investigating such constraints (Geisler, 2011). While biological perceptual systems likely never reach optimal performance, in some cases humans share behavioral characteristics of ideal observers, suggesting that those behaviors are consequences of having been optimized under particular biological or environmental constraints (Heinz et al., 2001a, 2001b; Weiss et al., 2002; Burge and Geisler, 2011; Girshick et al., 2011). Ideal observers provide a powerful framework for normative analysis, but for many real-world tasks, deriving provably optimal solutions is analytically intractable. The relevant sensory transduction properties are often prohibitively complicated, and the task-relevant parameters of natural stimuli and environments are difficult to specify mathematically. An attractive alternative might be to collect many real-world stimuli and optimize a model to perform the task on these stimuli. Even if not fully optimal, such models might reveal consequences of optimization under constraints that could provide insights into behavior.

In this paper, we explore whether contemporary “deep” artificial neural networks (DNNs) can be used in this way to gain normative insights about complex perceptual tasks. DNNs provide general-purpose architectures that can be optimized to perform challenging real-world tasks (LeCun et al., 2015). While DNNs are not guaranteed to achieve optimal performance, they might reveal the effects of optimizing a system under particular constraints (Kell and McDermott, 2019a). Previous work has documented similarities between human and network behavior for neural networks trained on vision or hearing tasks (Yamins and DiCarlo, 2016; Jozwik et al., 2017; Kell et al., 2018; Watanabe et al., 2018). However, we know little about the extent to which human-DNN similarities depend on either biological constraints that are built into the model architecture or the sensory signals for which the models are optimized. By manipulating the properties of simulated sensory transduction processes and the stimuli on which the DNN is trained, we hoped to get insight into the origins of behaviors of interest.

Here, we test this approach in the domain of pitch – traditionally conceived as the perceptual correlate of a sound's fundamental frequency (F0) (Plack and Oxenham, 2005). Pitch is believed to enable a wide range of auditory-driven behaviors, such as voice and melody recognition (McPherson and McDermott, 2018), and has been the subject of a long history of work in psychology (Moore et al., 1985; Shackleton and Carlyon, 1994; Moore and Moore, 2003; Oxenham et al., 2004; Bernstein and Oxenham, 2005) and neuroscience (Cariani and Delgutte, 1996; Patterson et al., 2002; Bendor and Wang, 2005; Cedolin and Delgutte, 2005; Norman-Haignere et al., 2013). Yet despite a wealth of data, the underlying computations and constraints that determine pitch perception remain debated (de Cheveigné, 2005; Oxenham, 2013). In particular, controversy persists over the role of spike timing in the auditory nerve, which is exquisitely precise, but for which a physiological extraction mechanism has remained elusive (de Cheveigné and Pressnitzer, 2006; Moore et al., 2006; Oxenham et al., 2009; Verschooten et al., 2019). The role of cochlear frequency selectivity, which has also been proposed to constrain pitch discrimination, remains similarly debated (Cariani and Delgutte, 1996; Bernstein and Oxenham, 2003, 2006; Mehta and Oxenham, 2020). By contrast, little attention has been given to the possibility that pitch perception might instead or additionally be shaped by the constraints of estimating the F0 of natural sounds in natural environments.

One factor limiting the resolution of these debates is that previous models of pitch have generally not attained quantitatively accurate matches to human behavior (Licklider, 1951; Schouten et al., 1962; Goldstein, 1973; Wightman, 1973; Terhardt, 1979; Slaney and Lyon, 1990; Meddis and Hewitt, 1991; Meddis and O'Mard, 1997; Shamma and Klein, 2000; Bernstein and Oxenham, 2005; Laudanski et al., 2014; Ahmad et al., 2016; Barzelay et al., 2017). Moreover, because most previous models have been mechanistic rather than normative, they were not optimized for their task, and thus do not speak to the potential adaptation of pitch perception for particular types of sounds or peripheral neural codes. Here we used deep neural networks in the role traditionally occupied by ideal observers, optimizing them to extract pitch information from the rich peripheral neural representations of natural sounds. Deep neural networks have become the method of choice for pitch tracking in engineering applications (Lee and Ellis, 2012; Han and Wang, 2014; Kim et al., 2018), but have not been combined with realistic models of the peripheral auditory system, and have not been compared to human perception. We then tested the influence of peripheral auditory physiology and natural sound statistics on human pitch perception by manipulating them during model optimization. The results provide new evidence for the importance of peripheral phase-locking in human pitch perception. However, they also indicate that the properties of pitch perception reflect adaptation to natural sound statistics, in that systems optimized for alternative stimulus statistics deviate substantially from human-like behavior.

RESULTS

Optimizing DNNs to estimate F0 of natural sounds from cochlear representations

Training task and stimuli

We used supervised deep learning to build a model of pitch perception optimized for natural speech and music. Deep convolutional neural networks were trained to estimate the F0 of short (50 ms) segments of speech and musical instrument recordings, selected to have high periodicity and well-defined F0s. To emulate natural listening conditions, the speech and music clips were embedded in aperiodic background noise taken from YouTube soundtracks. The networks' task was to classify each stimulus into one of 700 F0 classes (log-spaced between 80 Hz and 1000 Hz, bin width = 1/16 semitones = 0.36% F0). We generated a dataset of 2.1 million stimuli. Networks were trained using 80% of this dataset and the remaining 20% was used as a validation set to measure the success of the optimization.

Peripheral auditory model

In our primary training condition, we hard-coded the input representation for our networks to be as faithful as possible to known peripheral auditory physiology. We used a detailed phenomenological model of the auditory nerve (Bruce et al., 2018) to simulate peripheral representations of each stimulus (Fig. 1A). The input representations to our network consisted of 100 simulated auditory nerve fibers. Each stimulus was represented as a 100-fiber by 1000-timestep array of instantaneous firing rates (sampled at 20 kHz).

An example simulated auditory nerve representation for a harmonic tone is shown in Fig. 1B. Theories of pitch have tended to gravitate toward one of the two axes of such representations: the frequency-to-place mapping along the cochlea’s length, or the time axis. However, it is visually apparent that the nerve representation of even this relatively simple sound is quite rich, with a variety of potential cues: phase-locking to individual frequencies, phase shifts between these phase-locked responses, peaks in the time-averaged response (the “excitation” pattern) for low-numbered harmonics, and phase-locking to the F0 for the higher-numbered harmonics. The models have access to all of this information. Through optimization for the training task, the models should learn to use whichever peripheral cues best allow them to extract F0.

DNN architecture search

The performance of an artificial neural network is influenced both by the particular weights that are learned during training and by the various parameters that define the architecture of the network (Yamins et al., 2014). To obtain a high-performing model, we performed a large-scale random architecture search. Each architecture consisted of a feedforward series of layers instantiating linear convolution, nonlinear rectification, normalization, and pooling operations. Within this family, we trained 400 networks varying in their number of layers, number of units per layer, extent of pooling between layers, and the size and shape of convolutional filters (Fig. 1C).

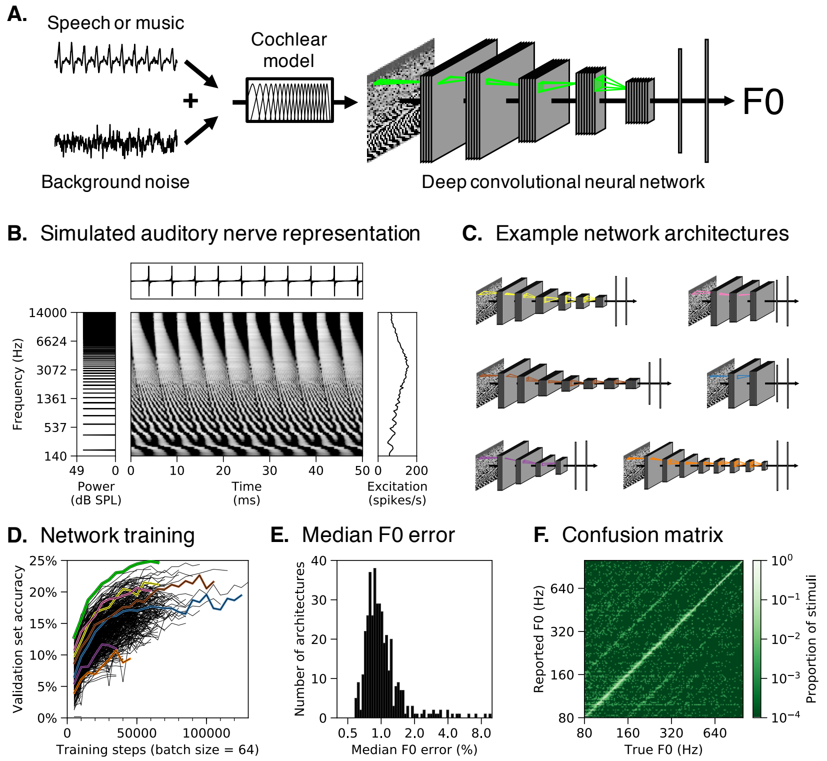


Figure 1. Pitch model overview. **(A)** Schematic of model structure. Deep neural networks were trained to estimate the F0 of speech and music sounds embedded in real-world background noise. Networks received simulated auditory nerve representations of acoustic stimuli as input. Green outlines depict the extent of example convolutional filter kernels in time and frequency (horizontal and vertical dimensions, respectively). **(B)** Simulated auditory nerve representation of a harmonic tone with a fundamental frequency (F0) of 200 Hz. The sound waveform is shown above and its power spectrum to the left. The waveform is periodic in time, with a period of 5ms. The spectrum is harmonic (i.e., containing multiples of the fundamental frequency). Network inputs were arrays of instantaneous auditory nerve firing rates (depicted in greyscale, with lighter hues indicating higher firing rates). Each row plots the firing rate of a frequency-tuned auditory nerve fiber, arranged in order of their place along the cochlea (with low frequencies at the bottom). Individual fibers phase-lock to low-numbered harmonics in the stimulus (lower portion of the nerve representation), or to the combination of high-numbered harmonics (upper portion). Time-averaged responses on the right show the pattern of nerve fiber excitation across the cochlear frequency axis (the "excitation pattern"). Low-numbered harmonics produce distinct peaks in the excitation pattern. **(C)** Schematics of six example neural network architectures trained to estimate F0. Network architectures varied in the number of layers, the number of units per layer, the extent of pooling between layers, and the size and shape of convolutional filter kernels **(D)** Summary of network architecture search. F0 classification performance on the validation set (noisy speech and instrument stimuli not seen during training) is shown as a function of training steps for all 400 networks trained. The highlighted curves correspond to the architectures depicted in A and C. The seemingly low overall accuracy reflects the fine-grained F0 bins we used. **(E)** Histogram of accuracy, expressed as the median F0 error on the validation set, for all trained networks (F0 error in percent is shown here because it is more interpretable than the classification accuracy, the absolute value of which is dependent on the width of the F0 bins). **(F)** Confusion matrix for the best-performing network (depicted in A) tested on the validation set.

The different architectures produced a broad distribution of training task performances (Fig. 1D). In absolute terms accuracy was good – the median error of the best-performing networks was well below 1% (Fig. 1E), which is on par with good human F0 discrimination thresholds (Houtsma and Smurzynski, 1990; Bernstein and Oxenham, 2005). The vast majority of misclassifications fell within bins neighboring the true F0 or at an integer number octaves away (Fig. 1F), as in human pitch-matching judgments (Moore et al., 1992; McDermott et al., 2008).

Characteristics of pitch perception emerge in DNNs optimized to estimate F0

Having obtained a model that can estimate F0 from natural sounds, we simulated a suite of well-known psychophysical experiments on the model to assess whether it replicated known properties of human pitch perception. Each experiment measures the effect of particular cues on pitch discrimination or estimation using synthetic tones (Fig. 2, left column), and produces an established result in human listeners (Fig. 2, center column). We tested the effect of these classic stimulus manipulations on our 10 best-performing network architectures. Given evidence for individual differences across different networks optimized for the same task (Mehrer et al., 2020), most figures feature results averaged across the 10 best networks identified in our architecture search (which we collectively refer to as 'the model'). Individual results for these networks are shown in Supplemental Fig. 1. We emphasize that none of the stimuli were included in the networks' training set, and that the model was not fit to match human results in any way.

As shown in Fig. 2, the model (right column) qualitatively and in most cases quantitatively replicates the result of each of the five different experiments in humans (center column). These results collectively suggest that the model relies on similar cues as the human pitch system.

Each of these results has an established interpretation relative to classical theories of pitch, and are described below for interested readers. However, understanding the details of each stimulus manipulation and corresponding perceptual effect is not critical to the larger message of the paper. The most important point is that the model produces human-like results for all of the experiments.

Dependence on low-numbered harmonics

One of the most robust findings in human pitch perception is that discrimination is more accurate for stimuli containing low-numbered harmonics (Fig. 2A, center, solid line) (Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994; Bernstein and Oxenham, 2003, 2005). This finding is often interpreted as evidence for the importance of place cues to pitch, which are only present for low-numbered harmonics (Fig. 1B, right). The model reproduced this effect, though the inflection point was somewhat lower than in human listeners: discrimination thresholds were low only for stimuli containing the fifth or lower harmonic (Fig. 2A, right, solid line).

Phase effects are limited to high-numbered harmonics

A second well-known result is that human perception is affected by harmonic phases only for high-numbered harmonics. This result is typically thought to indicate use of temporal fluctuations in a sound's envelope when cues for low-numbered harmonics are not available (Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994; Cariani and Delgutte, 1996). When harmonic phases are randomized, human discrimination thresholds are elevated for stimuli that lack low-numbered harmonics (Fig. 2A, center, dashed vs. solid line) (Bernstein and Oxenham, 2005). In addition, when odd and even harmonics are summed in sine and cosine phase, respectively ("alternating phase", a manipulation that doubles the number of peaks in the waveform's temporal envelope; Fig. 2B, left), listeners report the pitch to be twice as high as the corresponding sine-phase complex, but only for high-numbered harmonics (Fig. 2B, center) (Shackleton and Carlyon, 1994). The model replicates both effects (Fig. 2A&B, right), indicating that it uses similar temporal cues to pitch as humans, and in similar conditions.

Pitch shifts for shifted low-numbered harmonics

Third, frequency-shifted complex tones (in which all of the component frequencies have been shifted by the same number of Hz; Fig. 2C, left) produce linear shifts in the pitch reported by humans, but only if the tones contain low-numbered harmonics (Fig. 2C, center) (Moore and Moore, 2003). The model's F0 predictions for these stimuli resemble those measured from human listeners (Fig. 2C, right).

Fourth, shifting individual harmonics in a complex tone ("mistuning"; Fig. 2D, left) can also produce pitch shifts in humans under certain conditions (Moore et al., 1985): the mistuning must be small (effects are largest for 3-4% mistunings) and applied to a low-numbered harmonic (Fig. 2D, center). The model replicates this effect as well, although the size of the shift is smaller than that observed in humans (Fig. 2D, right).

Poor discrimination of transposed tones

A fifth result concerns "transposed tones" designed to instantiate the temporal cues from low frequencies at a higher-frequency place on the cochlea (Fig. 2E, left) (Oxenham et al., 2004). Despite producing similar temporal cues to pure tones (albeit at different places along the cochlea's frequency axis), transposed tones elicit weak pitch percepts in humans and thus yield higher discrimination thresholds than pure tones (Fig. 2E, center). This finding is taken to indicate that to the extent that temporal cues to pitch matter, they must occur at the correct place on the cochlea. The model reproduced this effect: discrimination thresholds were worse for transposed tones than they are for pure tones (Fig. 2E, right).

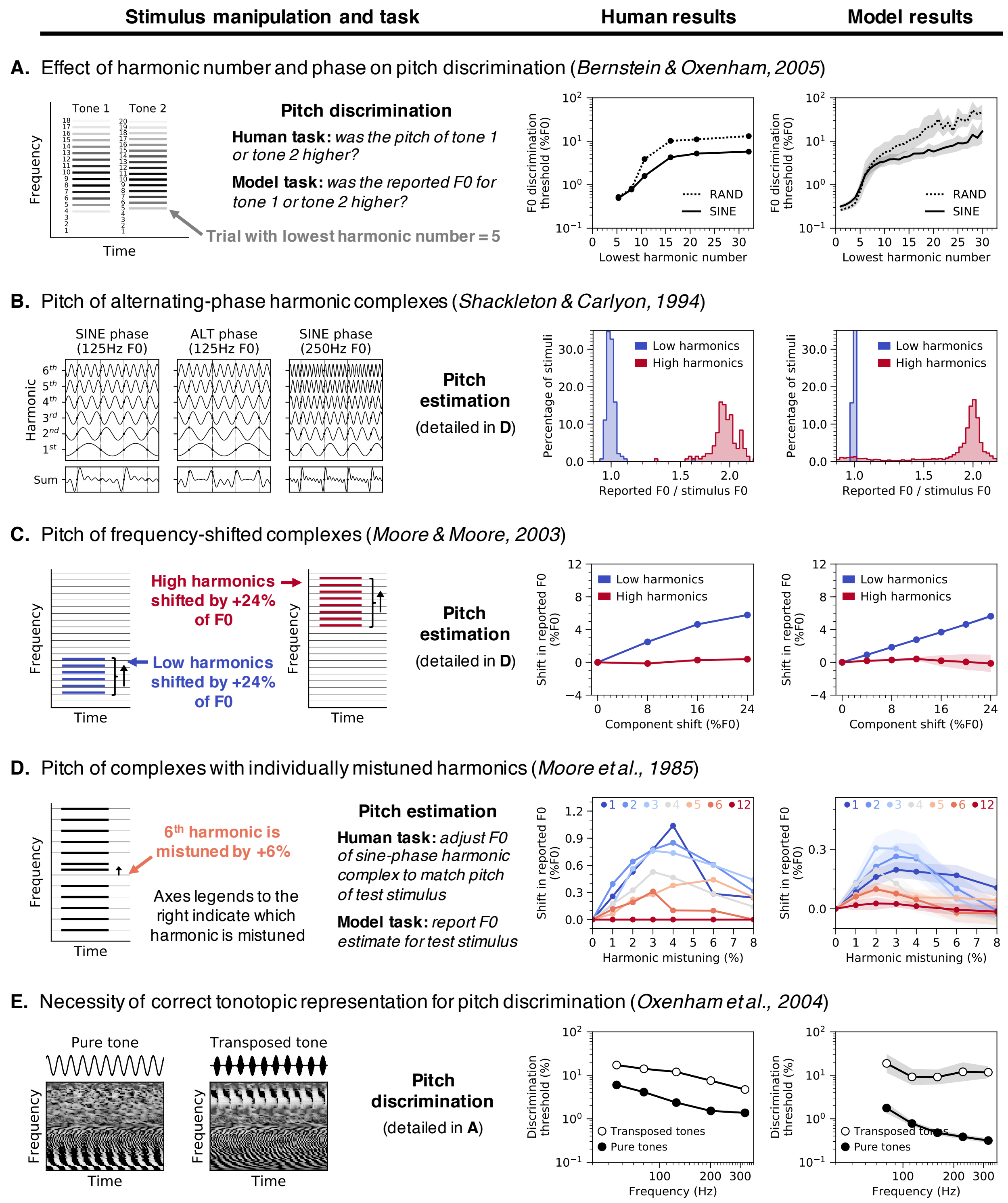


Figure 2. Pitch model validation: human and network psychophysics. Five classic experiments from the pitch psychoacoustics literature (**A-E**) were simulated on networks trained to estimate the F0 of natural sounds. Each row corresponds to a different experiment and contains (from left to right) a schematic of the experimental stimuli, results from human listeners (re-plotted from the original studies), and results from networks. Error bars indicate bootstrapped 95% confidence intervals around the mean of the 10 best network architectures ranked by F0 estimation performance on natural sounds (individual network results are shown in Supplemental Fig. 1). **(A)** F0 discrimination thresholds for bandpass synthetic tones, as a function of lowest harmonic number and phase. Human

listeners and networks made pitch judgments between pairs of sine-phase or random-phase harmonic tones with similar F0s. Stimuli were bandpass-filtered to control which harmonics were audible. **(B)** Perceived pitch of alternating-phase complex tones containing either low or high-numbered harmonics. Alternating-phase tones (i.e., with odd-numbered harmonics in sine phase and even-numbered harmonics in cosine phase) contain twice as many peaks in the waveform envelope as sine-phase tones with the same F0. Human listeners adjusted a sine-phase tone to match the pitch of the alternating-phase tone. Networks made F0 estimates for the alternating-phase tones directly. Histograms show distributions of pitch judgments as the ratio between the reported F0 and the stimulus F0. **(C)** Pitch of frequency-shifted complexes. Harmonic complexes (containing either low or high-numbered harmonics) were made inharmonic by shifting all component frequencies by the same number of Hz. Human listeners and networks reported the F0s they perceived for these stimuli (same experimental methods as in B). Shifts in the perceived F0 are shown as a function of the shift applied to the component frequencies. **(D)** Pitch of complexes with individually mistuned harmonics. Human listeners and networks reported the F0s they perceived for complex tones in which a single harmonic frequency has been shifted (same experimental methods as in B). Shifts in the perceived F0 are shown as a function of the mistuning applied to seven different harmonics within the tone (harmonic numbers indicated in different colors at top of graphs). Note that the y-axis limits are different in the human and model graphs – they exhibit qualitative but not quantitative similarity. **(E)** Frequency discrimination thresholds measured with pure tones and transposed tones. Transposed tones are high-frequency tones that are amplitude-modulated so as to instantiate the temporal cues from low-frequency pure tones at a higher-frequency place on the cochlea. Human and network listeners discriminated pairs of pure tones with similar frequencies and pairs of transposed tones with similar envelope frequencies.

DNNs with better F0 estimation exhibit more human-like behavior

To evaluate whether the human-model similarity evident in Fig. 2 depends on having optimized the model architecture for F0 estimation of natural sounds, we simulated the full suite of psychophysical experiments on each of our 400 trained networks. These 400 networks varied in how well they performed F0 estimation on the validation set (Fig. 1D&E). For each psychophysical experiment and network, we quantified the similarity between human and network results with a correlation coefficient. We then compared this human-model similarity metric to each network's performance on the validation set (Fig. 3A-E). The effect of optimization was more pronounced for some experiments than others. For four of the five experiments (Fig. 3A-D), there was a significant positive correlation between training task performance and human-model similarity ($p<0.001$ in each case). The transposed tones experiment (Fig. 3E) was the exception, as all networks similarly replicated the main human result regardless of their training task performance. To illustrate the effect of optimization for one experiment, Fig. 3F displays the average F0 discrimination thresholds for each of the worst, middle, and best 10% of networks (sorted by performance on the validation set) as a function of lowest harmonic number and phase. It is visually apparent that top-performing networks exhibit more similar psychophysical behavior to humans than worse-performing networks. See Supplemental Fig. 2 for analogous results for the other four experiments from Fig. 2. Overall, these results indicate that networks with better performance on the F0-estimation training task generally exhibit more human-like pitch behavior, consistent with the idea that these patterns of human behavior are byproducts of optimization under natural constraints.

Because the space of network architectures is large, it is a challenge to definitively associate particular network motifs with good performance and/or human-like behavior. However, one clear effect was evident in our results: very shallow networks both performed poorly on the training task and exhibited less similarity with human behavior. Of the 400 randomly-generated networks we considered, 54 contained only one convolutional layer. These 54 single-convolutional-layer networks produced lower validation set accuracies (z=7.31, p<0.001, Wilcoxon rank-sum test) and lower human-model similarity (with results pooled across all experiments; z=9.24, p<0.001, Wilcoxon rank-sum test) than the remaining 346 multi-convolutional-layer networks (Supplemental Fig. 3). Moreover, the top 40% of networks ranked according to overall human-model similarity consisted entirely of multi-convolutional-layer networks. These results provide evidence that deep networks (with multiple hierarchical stages of processing) better account for human pitch behavior than relatively shallow networks.

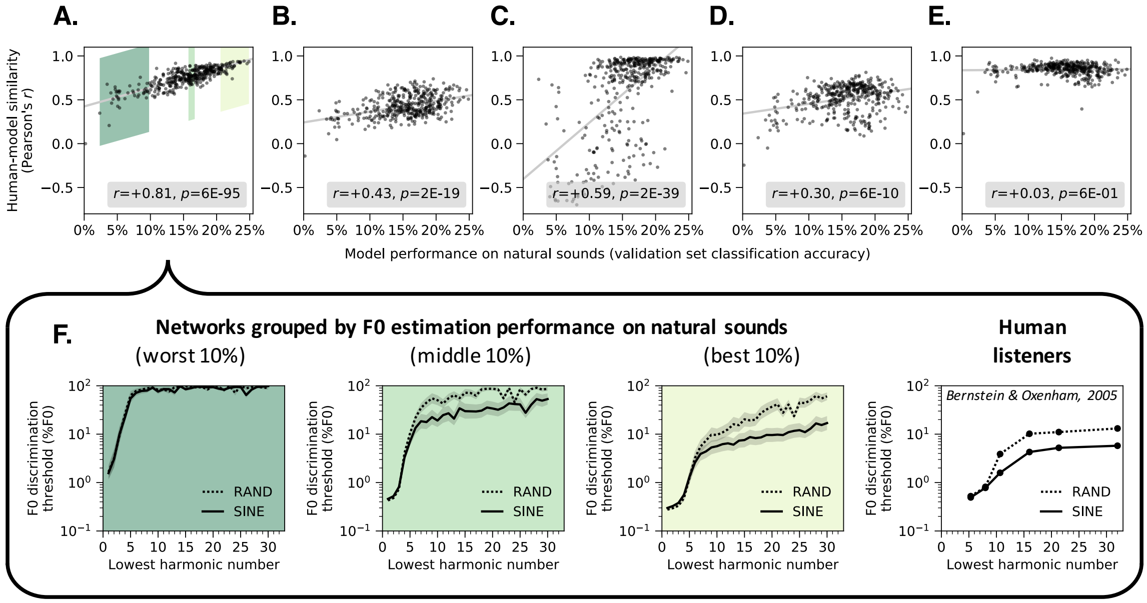


Figure 3. Network architectures producing better F0 estimation for natural sounds exhibit more human-like pitch behavior. **A-E** plot human-model similarity for all 400 architectures as a function of the accuracy of the trained architecture on the validation set (a set of stimuli distinct from the training dataset, but generated with the same procedure). The similarity between human and individual network results was quantified for each experiment as the correlation coefficient between analogous data points (see Methods). Pearson correlations between validation set accuracy and human-model similarity for each experiment are noted in the legends. Each graph (**A-E)** corresponds to one of the five main psychophysical experiments (see Fig. 2A-E): **(A)** F0 discrimination as a function of harmonic number and phase, **(B)** pitch estimation of alternating-phase stimuli, **(C)** pitch estimation of frequency-shifted complexes, **(D)** pitch estimation of complexes with individually mistuned harmonics, and **(E)** frequency discrimination with pure and transposed tones. **(F)** The results of the experiment from A (F0 discrimination thresholds as a function of lowest harmonic number and harmonic phase) measured from the 40 worst, middle, and best architectures ranked by F0 estimation performance on natural sounds (indicated with green-scale patches in A). Error bars indicate 95% confidence intervals via bootstrapping across the 40 networks. Human F0 discrimination thresholds from the same experiment are re-plotted for comparison.

Dependence of pitch behavior on the peripheral auditory representation

The results thus far demonstrate that a model optimized to extract F0 from cochlear representations of natural sounds accounts for many characteristics of human pitch perception. To gain insight into what underlies these characteristics, we next investigated how the model behavior depends on the peripheral representation. Decades of research has sought to determine the aspects of peripheral auditory representations that underlie pitch judgments, but experimental research has been limited by the difficulty of manipulating properties of peripheral representations, and the roles of place, timing, and other cues has thus remained uncertain. We took advantage of the ability to perform experiments on the model that are not possible in biology, training networks with peripheral representations that were altered in various ways. To streamline presentation, we focus on just the best-performing network architectures from our search (presenting results averaged across the top 10 networks, to ensure robustness) and on a single psychophysical result that we found to be particularly diagnostic: the effect of lowest harmonic number on F0 discrimination thresholds (Fig. 2A, solid line). Results for other experiments are generally congruent with the overall conclusions and are shown in supplemental figures.

Human-like behavior is critically dependent on phase-locking in the auditory nerve

To investigate the role of temporal coding in the auditory periphery, we trained networks with alternative upper limits of auditory nerve phase-locking. Phase-locking is limited by biophysical properties of inner hair cell transduction (Palmer and Russell, 1986), which are impractical to alter in vivo but which can easily be modified in silico by adjusting the simulated inner hair cell's lowpass filter (Bruce et al., 2018). We separately trained networks with inner hair cell lowpass cutoff frequencies of 50 Hz, 320 Hz, 1000 Hz, 3000 Hz (the nerve model's default value, matched to cat auditory nerve measurements, commonly presumed to roughly match that of the human auditory nerve), 6000 Hz, and 9000 Hz. With a cutoff frequency of 50 Hz, virtually all temporal structure in the peripheral representation of our stimuli is eliminated, meaning the network only has access to cues from the place of excitation along the cochlea (Fig. 4A). As the cutoff frequency is increased, the network has access to progressively finer-grained spike-timing information in addition to the place cues. The 10 best-performing networks from the architecture search were retrained separately with each of these altered cochlear transforms.

Reducing the upper limit of phase-locking qualitatively changed the model's psychophysical behavior and made it less human-like. As shown in Fig. 4B&C, F0 discrimination thresholds became worse, with the best threshold (the left-most data point, corresponding to a lowest harmonic number of 1) increasing as the cutoff was lowered (significantly worse for all three conditions: 1000 Hz, $t(18)= 4.39$, $p<0.001$; 320 Hz, $t(18)= 11.57$, $p<0.001$; 50 Hz, $t(18)= 9.30$, $p<0.001$; two-sample t-tests comparing to thresholds in the 3000 Hz condition). This in itself is not surprising, as it has long been known that phase locking enables better frequency discrimination than place information alone (Siebert, 1970; Heinz et al., 2001a, 2001b). However, thresholds also showed a different dependence on harmonic number as the phase-locking cutoff was lowered. Specifically, the transition from good to poor thresholds, here defined as the left-most point where thresholds exceeded 1%, was lower with degraded phase locking. This difference was significant for two of the three conditions (1000 Hz, $t(18)=5.15$, $p<0.001$; 50 Hz, $t(18)=10.10$, $p<0.001$; two-sample t-tests comparing to the 3000 Hz condition; the transition point was on average lower for the 320 Hz condition, but the results were more variable across architectures, and so the difference was not statistically significant). Increasing the cutoff to 6000 Hz or 9000 Hz had minimal effects on both of these features (Fig. 4C), suggesting that superhuman temporal resolution would not continue to improve pitch perception (at least as assessed here). Overall, these results suggest that auditory nerve phase-locking like that believed to be present in the human ear is critical for human-like pitch perception.

A common criticism of place-based pitch models is that they fail to account for the robustness of pitch across sound level, because cochlear excitation patterns saturate at high levels (Cedolin and Delgutte, 2005). Consistent with this idea, frequency discrimination thresholds (Fig. 4D) measured from networks with lower phase-locking cutoffs were less invariant to level than networks trained with normal spike-timing information (Fig. 4E, right). Thresholds measured from networks with limited phase-locking were progressively worse for louder tones, unlike those for humans (Fig. 4E, left) which remain good (well below 1%) at high sensation levels (Wier et al., 1977). This effect produced an interaction between the effect of stimulus level and the phase-locking cutoff on discrimination thresholds (F(13.80,149.08)=4.63, p<0.001, $\eta^2_{partial} = 0.30$), in addition to the main effect of the cutoff (F(5,54)=23.37, p<0.001, $\eta^2_{partial} = 0.68$; also evident in Fig. 4C). Similar effects were observed when thresholds were measured with complex tones (data not shown).

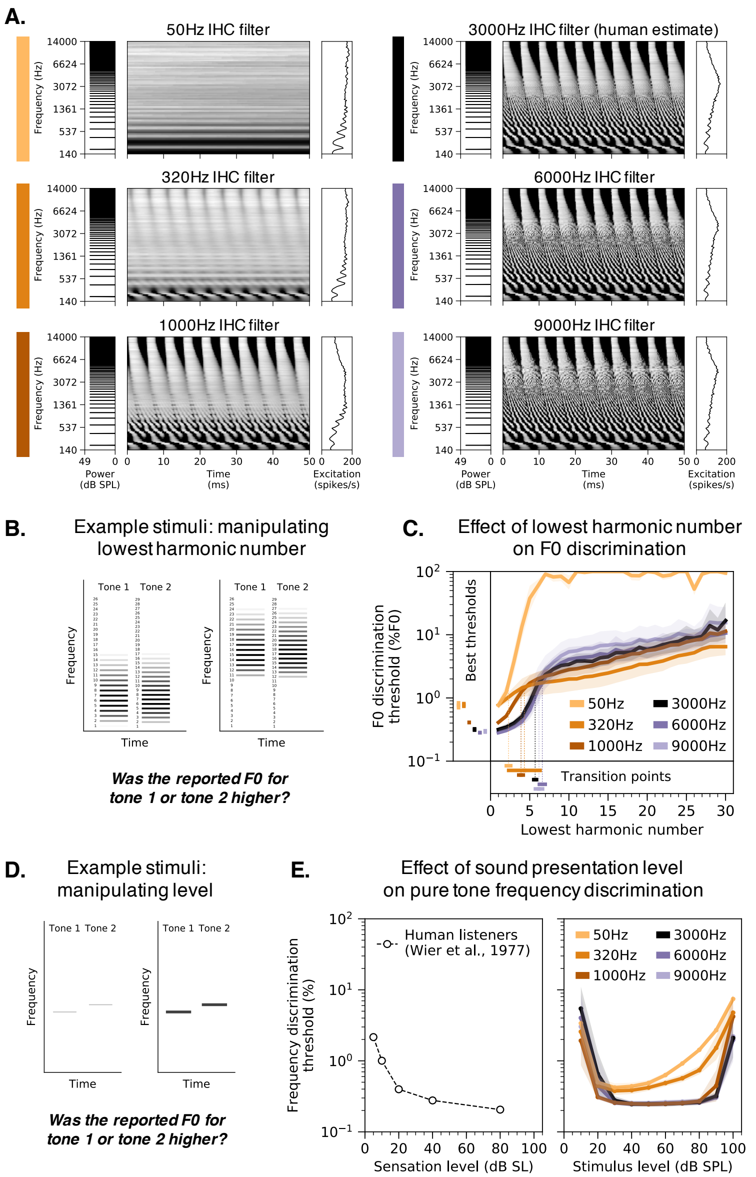


Figure 4. Pitch perception is impaired in networks optimized with degraded spike-timing information in the auditory nerve. **(A)** Simulated auditory nerve representations of the same stimulus (harmonic tone with 200 Hz F0) under six configurations of the peripheral auditory model. Configurations differed in the cutoff frequency of the inner hair cell lowpass filter, which sets the upper limit of auditory nerve phase-locking. The 3000 Hz setting is that normally used to model the human auditory system. As in Fig. 1A, each peripheral representation is flanked by the stimulus power spectrum and the time-averaged cochlear excitation pattern. **(B)** Schematic of stimuli used to measure F0 discrimination thresholds as a function of lowest harmonic number. Gray level denotes amplitude. Two example trials are shown, with two different lowest harmonic numbers. **(C)** F0 discrimination thresholds as a function of lowest harmonic number measured from networks trained and tested with each of the six peripheral model configurations depicted in A. The best thresholds and the transition points from good to poor thresholds (defined as the lowest harmonic number for which thresholds first exceeded 1%) are re-plotted to the left of and below the main axes, respectively. Error bars here and in E indicate 95% confidence intervals bootstrapped across the 10 best network architectures. **(D)** Schematic of stimuli used to measure frequency discrimination thresholds as a function of sound level. Gray level denotes amplitude. **(E)** Frequency discrimination thresholds as a function of sound level measured from human listeners (left) and from the same networks as C (right). Human thresholds, which are reported as a function of sensation level, are re-plotted from (Wier et al., 1977).

To control for the possibility that the poor performance of the networks trained with lower phase-locking cutoffs is due to the relatively small number of auditory nerve fibers used in the peripheral representation, we generated an alternative peripheral representation for the 50 Hz cutoff condition, with 1000 nerve fibers and 100 timesteps (sampled at 2 kHz). We then trained and tested the 10 best-performing networks from our architecture search on these representations (transposing the nerve fiber and time dimensions to maintain the input size and thus be able to use the same network architecture). Increasing the number of simulated auditory nerve fibers by a full order of magnitude modestly improved thresholds but did not qualitatively change the results: networks without high-fidelity temporal information still exhibited abnormal F0 discrimination behavior. The 50 Hz condition results in Fig. 4C&E are taken from the

1000 nerve fiber networks, as this seemed the most conservative comparison. Results for different numbers of nerve fibers are provided in Supplemental Fig. 4.

We simulated the full suite of psychophysical experiments on all networks with altered temporal resolution (Supplemental Fig. 5). Several other experimental results were also visibly different from those of humans in models with altered phase-locking cutoffs (in particular, the alternating-phase and mistuned harmonics experiments). Overall, the results indicate that normal human pitch perception depends on phase-locking up to 3000 Hz.

Human-like behavior is less dependent on cochlear filter bandwidths

The role of cochlear frequency tuning in pitch perception has also been the source of longstanding debates (Houtsma and Smurzynski, 1990; Carlyon and Shackleton, 1994; Arehart and Burns, 1999; Bernstein and Oxenham, 2003, 2006; Mehta and Oxenham, 2020). Classic “place” theories of pitch postulate that F0 is inferred from the peaks and valleys in the excitation pattern. Contrary to this idea, we found that simply eliminating all excitation pattern cues (by separately re-scaling each frequency channel in the peripheral representation to have the same time-averaged response) had almost no effect on network behavior (Supplemental Fig. 6). This result suggests that F0 estimation does not require the excitation pattern per se, but it remains possible it that might still be constrained by the frequency selectivity of the cochlea.

To investigate the perceptual effects of cochlear frequency tuning, we trained networks with altered tuning. We first scaled cochlear filter bandwidths to be two times narrower and two times broader than those estimated for human listeners (Shera et al., 2002). The effect of this manipulation is visually apparent in the width of pure tone frequency tuning curves measured from individual nerve fibers, as well as in the number of harmonics that produce distinct peaks in the cochlear excitation patterns (Fig. 5A).

We also modified the cochlear model to be linearly spaced (Fig. 5B), uniformly distributing the characteristic frequencies of the model nerve fibers along the frequency axis and equating their filter bandwidths (roughly matching the bandwidth to a standard human nerve fiber with a characteristic frequency of 400 Hz). Unlike a normal cochlea, which resolves only low-numbered harmonics, the linearly spaced alteration yielded a peripheral representation where all harmonics are equally resolved by the cochlear filters, providing another test of the role of frequency selectivity.

Contrary to the notion that cochlear frequency selectivity strongly constrains pitch discrimination, networks trained with different cochlear filter bandwidths exhibit relatively similar F0 discrimination behavior (Fig. 5C&D). Broadening filters by a factor of two had no significant effect on the best thresholds ($t(18)=0.40$, $p=0.69$, two-sample t-test comparing thresholds when lowest harmonic number = 1 to the human tuning condition). Narrowing filters by a factor of two yielded an improvement in best thresholds that was statistically significant ($t(18)=2.74$, $p<0.05$) but very small (0.27% compared to 0.32% for the networks with normal human tuning). Linearly spaced cochlear filters also yielded best thresholds that were not significantly different from those for normal human tuning (t(18)=1.88, p=0.07). In addition, the dependence of thresholds on harmonic number was fairly similar in all cases (Fig. 5C). The transition between good and poor thresholds consistently occurred around the sixth harmonic, irrespective of the cochlear filter bandwidths (the left-most point where thresholds exceeded 1% was not significantly different for any of the three altered tuning conditions: two times broader, t(18)=1.33, p=0.20; two times narrower, t(18)=1.00, p=0.33; linearly spaced, t(18)=0.37, p=0.71; two-sample t-tests comparing to the normal human tuning condition).

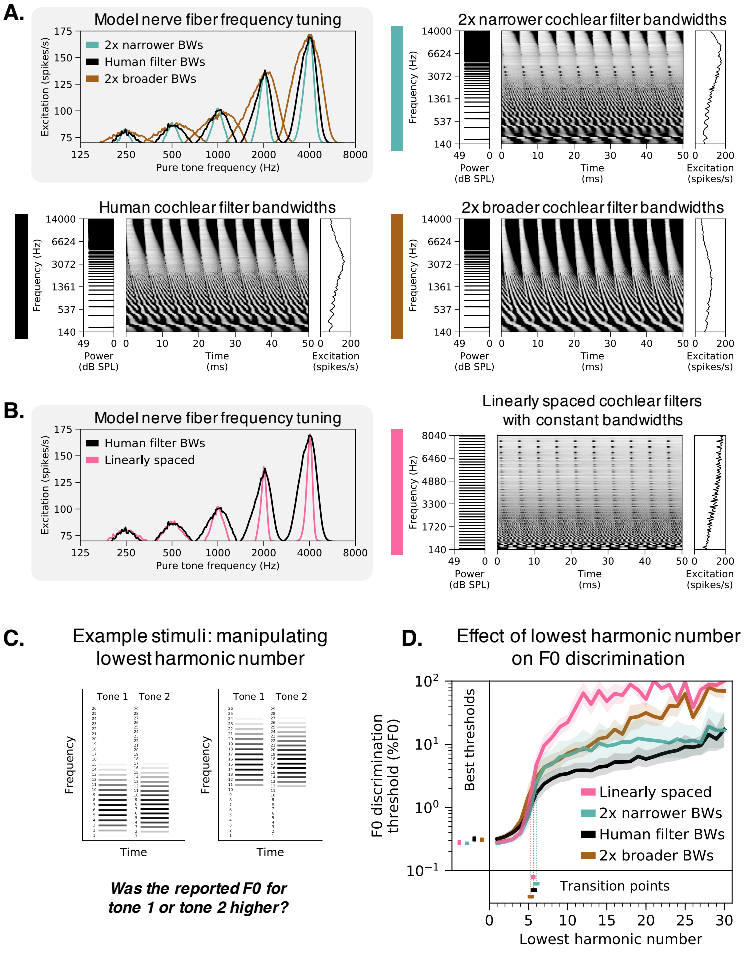


Figure 5. Cochlear frequency tuning of network inputs has relatively little effect on pitch perception. **(A)** Cochlear filter bandwidths were scaled to be two times narrower or two times broader than those estimated for normal-hearing humans. This manipulation is evident in the width of pure tone tuning curves measured from five individual nerve fibers per condition (upper left panel). Each peak corresponds to a different nerve fiber. Right and lower left panels show simulated auditory nerve representations of the same stimulus (harmonic tone with 200 Hz F0) for each bandwidth condition. Each peripheral representation is flanked by the stimulus power spectrum and the time-averaged auditory nerve excitation pattern. The excitation patterns are altered by changes in frequency selectivity, with coarser tuning yielding less pronounced peaks for individual harmonics, as expected. **(B)** The approximately log-spaced cochlear filters of the human ear were replaced with a set of linearly spaced filters with constant bandwidths. Pure tone tuning curves measured with linearly spaced filters are much sharper than those estimated for humans at higher frequencies (left panel). The right panel shows the simulated auditory nerve representation of the same stimulus from A with linearly spaced cochlear filters. In this condition, all harmonics are equally resolved by the cochlear filters and thus equally likely to produce peaks in the time-averaged excitation pattern. **(C)** Schematic of stimuli used to measure F0 discrimination thresholds. Gray level denotes amplitude. Two example trials are shown, with two different lowest harmonic numbers. **(D)** F0 discrimination thresholds as a function of lowest harmonic number, measured from networks trained and tested with each of the four peripheral model configurations depicted in A and B. The best thresholds and the transition points from good to poor thresholds (defined as the lowest harmonic number for which thresholds first exceeded 1%) are re-plotted to the left of and below the main axes, respectively. Error bars indicate 95% confidence intervals bootstrapped across the 10 best network architectures.

We also simulated the full suite of psychophysical experiments from Fig. 2 on networks with altered frequency tuning. Most experimental results were robust to peripheral frequency tuning (Supplemental Fig. 7). The main exception were the effects of harmonic phase, which were reduced for the linearly spaced models (correlations between human and model results for the phase randomization and alternating phase experiments were lower in the linearly spaced condition than in the human tuning condition; t(18)=3.13, p<0.01 and t(18)=6.50, p<0.001, respectively). These results are to be expected because the sharp tuning of the linearly spaced filters (Fig. 5B) results in less interaction between adjacent harmonics, which is believed to drive phase effects.

Dependence of pitch behavior on training set sound statistics

Networks trained on sounds with altered spectra exhibit altered behavior

In contrast to the roles of peripheral cues, which have been extensively debated throughout the history of hearing research, the role of natural sound statistics in pitch has been little discussed (Terhardt, 1974). To investigate how optimization for natural sound statistics may have shaped pitch perception, we fixed the cochlear representation to its normal human settings and instead manipulated the characteristics of the sounds on which networks were trained. One salient property of speech and instrument sounds is that they typically have more energy at low frequencies than high frequencies (Fig. 6A, left column, black line). To test if this lowpass characteristic shapes pitch behavior, we trained networks on highpass-filtered versions of the same stimuli (Fig. 6A, left column, orange line) and then measured their F0 discrimination thresholds (Fig. 6B). For comparison, we performed the same experiment with lowpass-filtered sounds.

Thresholds measured from networks optimized for unnatural highpass sounds exhibited a much weaker dependence on harmonic number than if optimized for natural sounds (Fig. 6C, left column). This difference produced an interaction between the effects of harmonic number and the training condition (F(2.16,38.85)=72.33, p<0.001, $\eta^2_{partial} = 0.80$). Mean thresholds remained very good (<1% F0) even when stimuli contain only high-numbered harmonics. By contrast the dependence on harmonic number was accentuated for lowpass-filtered stimuli, again producing an interaction between the effects of harmonic number and the training condition (F(4.25,76.42)=30.81, p<0.001, $\eta^2_{partial} = 0.63$).

We also simulated the full suite of psychophysical experiments on these networks (Supplemental Fig. 8). There are several other striking differences in the performance characteristics of networks trained on filtered natural sounds. In particular, networks optimized for highpass-filtered natural sounds exhibit better discrimination thresholds for transposed tones than pure tones (t(18)=14.08, p<0.001, two-sample t-test between pure tone and transposed tone thresholds averaged across frequency), a complete reversal of the human result. These results illustrate that the acuity of pitch is not strictly a function of the information available in the periphery – performance characteristics can depend strongly on the “environment” in which a system is optimized.

Network behavior can be flexibly modified by training on synthetic tones

To isolate the acoustic properties needed to reproduce human-like pitch behavior, we also trained networks on synthetic tones embedded in masking noise, with the spectral statistics of both tones and noise matched to those of the natural sound training set (Fig. 6A, middle column). Specifically, we fit multivariate Gaussians to the spectral envelopes of the speech/instrument sounds and the noise from the original training set, and synthesized stimuli with spectral envelopes sampled from these distributions. Although discrimination thresholds were overall somewhat better than when trained on natural sounds, the resulting network again exhibited human-like pitch characteristics (Fig. 6C, middle column, black line). Because the synthetic tones were constrained only by the mean and covariance of the spectral envelopes of our natural training data, the results suggest that such low-order spectral statistics capture much of the natural sound properties that matter for obtaining human-like pitch perception (see Supplemental Fig. 8 for results on the full suite of psychophysical experiments).

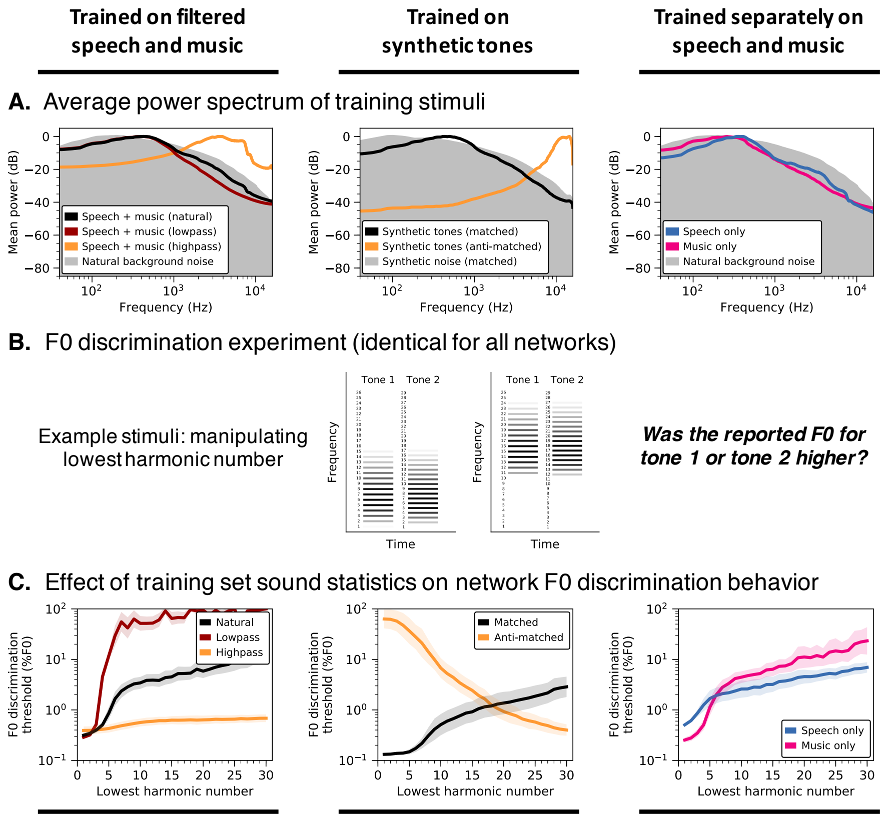


Figure 6. Pitch perception depends on training set sound statistics. **(A)** Average power spectrum of training stimuli under different training conditions. Networks were trained on datasets with lowpass- and highpass-filtered versions of the primary speech and music stimuli (column 1), as well as datasets of synthetic tones with spectral statistics either matched or anti-matched (see Methods) to those of the primary dataset (column 2), and datasets containing exclusively speech or music (column 3). Filtering indicated in column 1 was applied to the speech and music stimuli prior to their superposition on background noise. Grey shaded regions plot the average power spectrum of the background noise that pitch-evoking stimuli were embedded in. **(B)** Schematic of stimuli used to measure F0 discrimination thresholds as a function of lowest harmonic number. Two example trials are shown, with two different lowest harmonic numbers. **(C)** F0 discrimination thresholds as a function of lowest harmonic number, measured from networks trained on each dataset shown in A. Error bars indicate 95% confidence intervals bootstrapped across the 10 best network architectures.

For comparison, we also trained networks on synthetic tones with spectral statistics that deviate considerably from speech and instrument sounds. We generated these "anti-matched" synthetic tones by multiplying the mean of the fitted multivariate Gaussian by negative one (see Methods) and sampling spectral envelopes from the resulting distribution. Training on the resulting highpass synthetic tones (Fig. 6A, middle column, orange line) completely reversed the pattern of behavior seen in humans: discrimination thresholds were poor for stimuli containing low-numbered harmonics and good for stimuli containing only high-numbered harmonics (producing a negative correlation with human results: r=-0.98, p<0.001, Pearson correlation) (Fig. 6C, middle column, orange line). These results further illustrate that the dominance of low-numbered harmonics in human perception is not an inevitable consequence of cochlear transduction – good pitch perception is possible in domains where it is poor in humans, provided the system is trained to extract the relevant information.

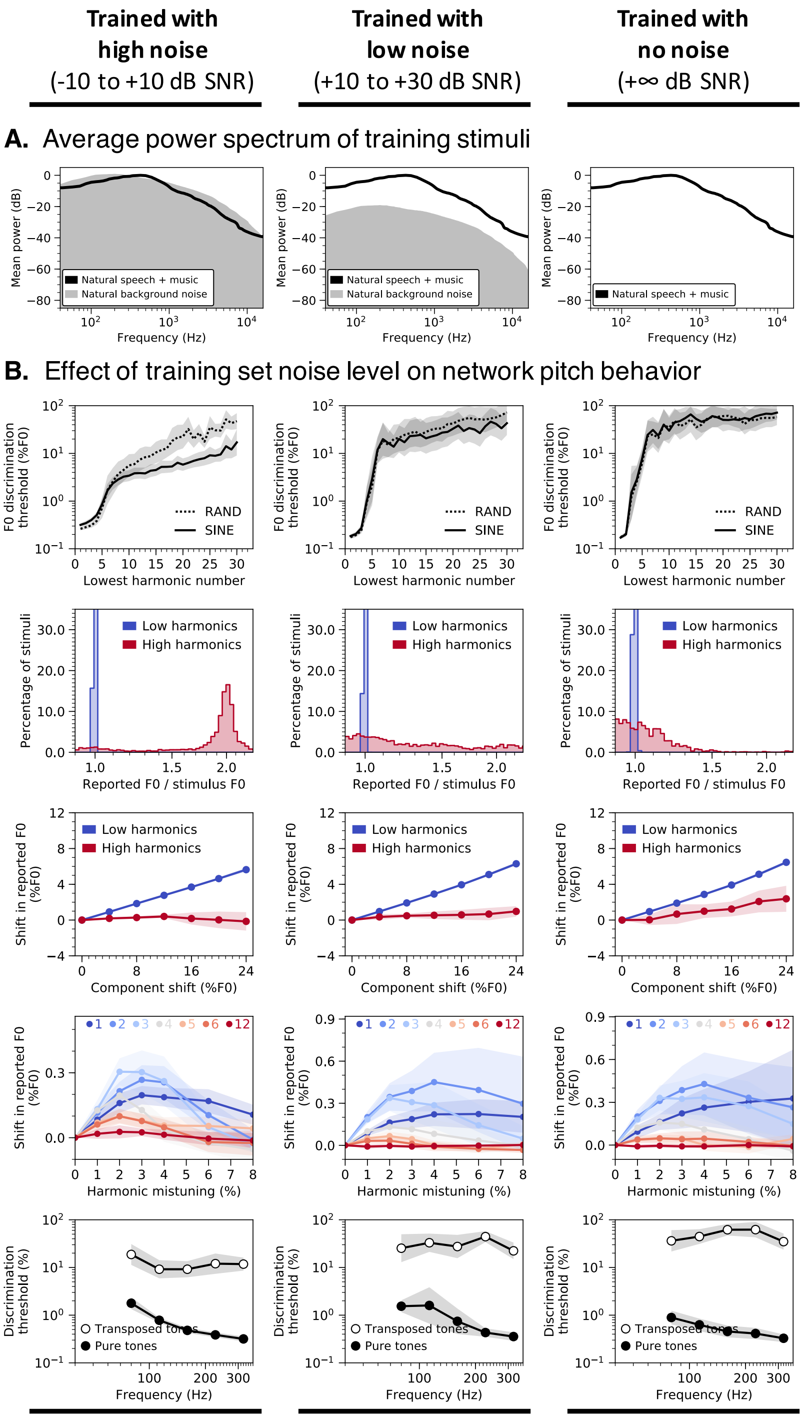
Networks trained on music exhibit better pitch acuity than networks trained on speech

We also trained networks separately using only speech or only music stimuli (Fig. 6A, right column). Consistent with the more accurate pitch discrimination found in human listeners with musical training (Kishon-Rabin et al., 2001; Micheyl et al., 2006; Besson et al., 2007; McDermott et al., 2010a), networks optimized specifically for music have lower discrimination thresholds for stimuli with low-numbered harmonics (Fig. 6C, right column; t(18)=9.73, p<0.001, two-sample t-test comparing left-most conditions – where the lowest harmonic number is 1, and which produce the best thresholds – for speech and music training). This result likely reflects the greater similarity of the psychophysical test stimuli to the music training stimuli compared to the speech stimuli, the latter of which are less perfectly periodic over the stimulus duration.

Networks trained without background noise lack the "missing fundamental" illusion

One of the core challenges of hearing is the ubiquity of background noise. To investigate how pitch behavior may have been shaped by the need to hear in noise, we varied the level of the background noise on which the speech and instrument excerpts were superimposed in our training set. Networks trained in relatively noisy environments (Fig. 7, left column) resembled human listeners in accurately inferring F0 even when the first harmonic was not physically present in the stimuli (thresholds for stimuli with lowest harmonic number between 2 and 5 were all under 1%). This effect, known as the "missing fundamental illusion", was progressively weakened in networks trained in lower levels of noise (higher SNRs) (Fig. 7, middle and right columns), with discrimination thresholds sharply elevated when the lowest harmonic number exceeded two (F(2,27)=6.79, p<0.01, $\eta^2_{partial} = 0.33$; main effect of training condition when comparing thresholds for lowest harmonic numbers between 2 and 5).

Networks trained in noiseless environments also deviated from human behavior when tested on alternating-phase (Fig. 7B, row 2) and frequency-shifted complexes (Fig. 7B, row 3), apparently ignoring temporal cues from high-numbered harmonics (correlations with human results were lower in each experiment; t(18)>4.41, p<0.001 for both comparisons between networks trained with high and no noise). Conversely, discrimination thresholds for pure tones (Fig. 7, row 5) remained good (below 1%), as though the networks learn to focus primarily on the first harmonic. Collectively, these results suggest the ability to extract F0 information from high-numbered harmonics in part reflects an adaptation for hearing in noise.



A. Average power spectrum of training stimuli

B. Effect of training set noise level on network pitch behavior

Figure 7. Key characteristics of human pitch behavior only emerge in noisy training conditions. **(A)** Average power spectrum of training stimuli. Networks were trained on speech and music stimuli embedded in three different levels of background noise: high (column 1), low (column 2), and none (column 3). **(B)** Effect of training set noise level on network behavior in all five main psychophysical experiments (see Fig. 2A-E): F0 discrimination as a function of harmonic number and phase (row 1), pitch estimation of alternating-phase stimuli (row 2), pitch estimation of frequency-shifted complexes (row 3), pitch estimation of complexes with individually mistuned harmonics (row 4), and frequency discrimination with pure and transposed tones (row 5). Error bars indicate 95% confidence intervals bootstrapped across the 10 best network architectures.

Network neurophysiology

Although our primary focus in this paper was to use DNNs to understand behavior in normative terms, the internal representations of DNNs have in some cases been related to stages of representation in the brain (Yamins et al., 2014; Guclu and van Gerven, 2015; Yamins and DiCarlo, 2016; Eickenberg et al., 2017; Kell et al., 2018). It was thus natural to wonder whether the internal representations of our model might exhibit established neural phenomena.

We simulated electrophysiology experiments on our best-performing network architecture by measuring the time-averaged rectified activation of each unit in each convolutional layer to a set of sine-phase harmonic tones (the same ones used to measure discrimination thresholds as a function of lowest harmonic number; Fig. 2A). Like pitch-selective neurons identified in primate auditory cortex (Bendor and Wang, 2005), we found network units that responded selectively to specific F0s, even across different harmonic compositions (Fig. 8A). In addition, F0 tuning was sharper deeper in the network (Fig. 8B).

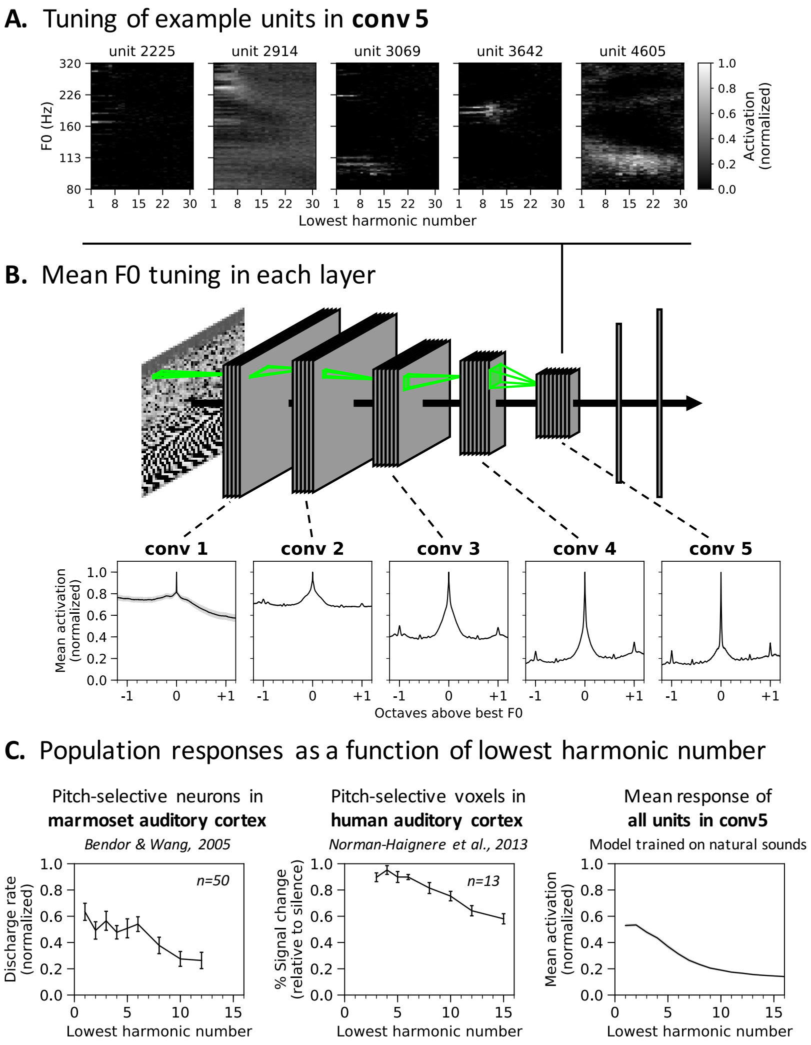


Figure 8. Network neurophysiology. Network activations were measured in response to complex tones varying in their lowest harmonic and F0 (using stimuli from Fig. 2A). **(A)** Time-averaged activations of five example units in the network's last convolutional layer, plotted as a function of lowest harmonic number and F0. **(B)** F0 tuning curves averaged across 3168 units (selected at random) in each convolutional layer. This number was chosen to equate the number of units being averaged in each layer (3168 is the number of units in the first layer). Prior to averaging, curves were aligned at the F0 that elicited the highest response. **(C)** Population responses of pitch-selective units in marmoset auditory cortex, human auditory cortex, and our network's last convolutional layer, plotted as a function of lowest harmonic number. Marmoset single-unit electrophysiological recordings were made from 3 animals and error bars indicate SEM across 50 neurons (re-plotted from (Bendor and Wang, 2005)). Human fMRI data were collected from 13 participants and error bars indicate within-subject SEM (re-plotted from (Norman-Haignere et al., 2013)). fMRI responses were measured from a functional region of interest defined by a contrast between harmonic tones and frequency-matched noise. Responses were measured in independent data (to avoid double dipping). Network responses were averaged across all units in the conv5 layer of the best-performing architecture (averaged across 10 instances of the architecture trained from different random initializations). Error bars for all network results indicate 95% confidence intervals bootstrapped across the 10 model instances.

To assess the population tuning to stimulus characteristics, we averaged the stimulus tuning curves across units in the network. Units in the deeper layers responded more strongly when stimuli contained low-numbered harmonics (Fig. 8C, right; main effect of lowest harmonic number on mean activation, F(29,261)=1341.39, p<0.001, $\eta_{partial}^2 = 0.99$). This result mirrors the response characteristics of pitch-selective neurons (measured with single-unit electrophysiology) in marmoset auditory cortex (Fig. 8C, left) (Bendor and Wang, 2005) and pitch-selective voxels (measured with fMRI) in human auditory cortex (Fig. 8C, center) (Norman-Haignere et al., 2013).

DISCUSSION

We developed a model of pitch perception by optimizing artificial neural networks to estimate the fundamental frequency of their acoustic input. The networks were trained on simulated auditory nerve representations of speech and music embedded in background noise. The best-performing networks closely replicated human pitch judgments in simulated psychophysical experiments despite never being trained on the synthetic psychophysical stimuli. To investigate which aspects of the auditory periphery and acoustical environment contribute to human-like pitch behavior, we optimized networks with altered cochleae and sound statistics. Lowering the upper-limit of phase-locking in the auditory nerve yielded models with behavior unlike that of humans: F0 discrimination was substantially worse than in humans and had a distinct dependence on stimulus characteristics. Model behavior was substantially less sensitive to changes in cochlear frequency tuning. However, the results were also strongly dependent on the sound statistics the model was optimized for. Optimizing for stimuli with unnatural spectra, or without concurrent background noise yielded behavior qualitatively different from that of humans. The results suggest that the characteristics of human pitch perception reflect the demands of estimating F0 of natural sounds, in natural conditions, given a human cochlea.

Relation to prior work

Our model innovates on prior work in pitch perception in two main respects. First, the model was optimized to achieve accurate pitch estimation in realistic conditions. By contrast, most previous pitch models have instantiated particular mechanistic or algorithmic hypotheses (Licklider, 1951; Schouten et al., 1962; Goldstein, 1973; Wightman, 1973; Terhardt, 1979; Slaney and Lyon, 1990; Meddis and Hewitt, 1991; Meddis and O'Mard, 1997; Shamma and Klein, 2000; Bernstein and Oxenham, 2005; Laudanski et al., 2014; Ahmad et al., 2016; Barzelay et al., 2017). Our model incorporated detailed simulations of the auditory nerve in its initial stages, but the rest of the model was free to implement any of a wide set of strategies that optimized performance. Optimization enabled us to test normative explanations of pitch perception (e.g., that it shows signatures of having been optimized for natural sound statistics and listening conditions) that have previously been neglected. Second, the model achieved reasonable quantitative matches to human pitch behavior. This match to behavior allowed strong tests of the role of different elements of peripheral coding in the auditory nerve. Prior work attempted to derive optimal decoders for pure tones (single frequencies) that operated on the auditory nerve (Siebert, 1970; Heinz et al., 2001a, 2001b), but was unable to assess pitch perception (i.e., F0 estimation) due to the added complexity of this task.

Both of these innovations were enabled by contemporary “deep” neural networks. For our purposes, such neural networks instantiate general-purpose functions that can be optimized to perform a training task. They are able to use any task-relevant information present in the sensory input, and avoid the need for hand-designed methods to extract such information. This generality is important for achieving good performance on real-world tasks. Hand-designed models, or simpler model classes, would likely not provide human-level performance. For instance, we found that very shallow networks both produced worse overall performance, and a poorer match to human behavior (Supplemental Fig. 3).

Normative insights into human pitch perception

Although mechanistic explanations of pitch perception have been widely discussed over many decades (Licklider, 1951; Schouten et al., 1962; Goldstein, 1973; Wightman, 1973; Terhardt, 1979; Slaney and Lyon, 1990; Meddis and Hewitt, 1991; Meddis and O’Mard, 1997; Shamma and Klein, 2000; Laudanski et al., 2014; Barzelay et al., 2017), there have been few attempts to explain pitch in normative terms. But like other aspects of perception, pitch is plausibly the outcome of an optimization process (realized through some combination of evolution and development) that produces good performance under natural conditions. We found evidence that these natural conditions have a large influence on the nature of pitch perception, in that human-like behavior emerged only in models optimized for naturalistic sounds heard in naturalistic conditions (with background noise).

In particular, the demands of extracting the F0 of natural sounds appear to explain one of the signature characteristics of human pitch perception: the dependence on low-numbered harmonics. This characteristic has traditionally been proposed to reflect limitations of cochlear filtering, with filter bandwidths determining the frequencies that can be resolved in a harmonic sound (Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994; Bernstein and Oxenham, 2003, 2006). However, we found that the dependence on harmonic number could be fully reversed for sufficiently unnatural sound training sets (Fig. 6C). Moreover, the dependence was remarkably robust to changes in cochlear filter bandwidths (Fig. 5C). These results suggest that pitch characteristics primarily reflect the constraints of natural sound statistics (specifically, lowpass power spectra) coupled with the high temporal fidelity of the auditory nerve. In the language of machine learning, discrimination thresholds appear to partly be a function of the match between the test stimuli and the training set (i.e., the sensory signals a perceptual system was optimized for). Our results suggest that this match is critical to explaining many of the well-known features of pitch perception.

A second large effect of the natural environment was evident when we eliminated background noise from the training set (Fig. 7). In such idealized listening conditions, the trained networks exhibited virtually no ability to extract F0 information from high-numbered harmonics. This result suggests that pitch is also in part a consequence of needing to hear in noise.

A third effect of the environment was evident in differences between models optimized solely for speech or music (Fig. 6C). Discrimination as measured in standard psychoacoustic tasks was better for models trained on musical instrument sounds than in models trained on speech sound. This result may help to explain known effects of musicianship on pitch discrimination (Besson et al., 2007; Kishon-Rabin et al., 2001; Micheyl et al., 2006; McDermott et al., 2010a). Together, these three results suggest that explanations of pitch perception cannot be separated from the natural environment.

The approach we propose here contrasts with prior work that derived optimal strategies for psychophysical tasks on synthetic stimuli (Durlach and Braida, 1969; Siebert, 1970; Colburn, 1973; Heinz et al., 2001a, 2001b; Micheyl et al., 2013). Although human listeners often improve on such tasks with practice, there is not much reason to expect them to approach optimal behavior for arbitrary tasks and stimuli (because these do not drive natural selection, or learning during development). By contrast, it is plausible that humans are near-optimal for important tasks in the natural environment, and that the consequences of this optimization will be evident in patterns of psychophysical performance, as we found here.

Evidence for the role of phase-locking in perception

Although the phase-locking of spikes in the auditory nerve is known to be important for computing interaural time differences in spatial hearing (Joris et al., 1998), its role in other aspects of perception is controversial (Kaernbach and Demany, 1998; Smith et al., 2002; Lorenzi et al., 2006; Moore, 2008; Grothe et al., 2010; Verschooten et al., 2019). Psychophysical evidence for the importance of phase-locking has remained circumstantial (Attneave and Olson, 1971; Javel and Mott, 1988; Semal and Demany, 1990; Jacoby et al., 2019) given the absence of direct measurements of phase-locking in humans along with the inability to manipulate it experimentally.

Our model provides strong evidence for the importance of phase-locking in pitch perception, and further suggests that the limits of phase-locking in humans are similar to those suggested from animal models. Absolute discrimination thresholds were substantially worse if the limit of phase-locking was decreased, as might be expected given that the models had less information to work with (Siebert, 1970). However, the dependence of thresholds on harmonic number also deviated markedly from that observed in humans (Fig. 4C). In particular, pitch perception for missing-fundamental complexes was impaired. Discrimination also became much less invariant to level (Fig. 4E). By contrast, increasing the limit of phase-locking to higher fidelity than is normally proposed had relatively little effect, suggesting that the posited limits of phase-locking are necessary and sufficient to account for human behavior, at least in this domain.

These conclusions were enabled by combining a realistic model of the auditory periphery with task-optimized neural networks.

Relation to classical theories of pitch

Debates over pitch mechanisms have historically been couched in terms of the two axes of the cochlear representation: place and time. Place models analyze the signature of harmonic frequency spectra in the excitation pattern along the length of the cochlea (Goldstein, 1973; Terhardt, 1979), whereas temporal models quantify signatures of periodicity in temporal patterns of spikes (Licklider, 1951; Meddis and Hewitt, 1991; Meddis and O'Mard, 1997; Cariani, 1999). Our model makes no distinction between place and time per se, using whatever information in the cochlear representation is useful for the training task. However, we were able to assess its dependence on the resolution of the cochlea in place and time by altering the properties of the simulated cochlea. These manipulations provided evidence that fine-grained peripheral timing is critical for normal pitch perception, and that fine-grained place-based frequency tuning is less so (Fig. 4). Some degree of cochlear frequency selectivity is likely critical to enabling phase-locking to low-numbered harmonics, but such effects evidently do not depend sensitively on tuning bandwidth. In addition, the model also reproduced perceptual effects in which listeners cannot use temporal cues occurring in an atypical place on the cochlea (Fig. 2E). The strong dependence of the model's behavior on the training set statistics suggests that it learns the cues that are present normally in natural sounds, which may be difficult to summarize in words.

Limitations

Our model is consistent with most available pitch perception data, but it is not perfect. For instance, the inflection point in the graph of Fig. 2A occurs at a somewhat lower harmonic number in the model than in humans. Given the evidence presented here that pitch perception reflects the stimulus statistics a system is optimized for, some discrepancies might be expected from the training set, which (due to the limitations of available corpora) consisted entirely of speech and musical instrument sounds, and omitted other types of natural sounds that are periodic in time. The range of F0s we trained on was similarly limited by available audio data sets, and prevents us from making predictions about the perception of very high frequencies (Oxenham et al., 2011). The level distribution was also not matched in a principled way to the natural world.

The cochlear model we used, although state-of-the-art and relatively well validated, is imperfect (e.g., peripheral representations consisted of firing rates rather than spikes), and could limit the extent to which this approach can replicate human behavior. We also note that the ear itself is the product of evolution and thus likely itself reflects properties of the natural environment (Lewicki, 2002). We chose to train models on a fixed representation of the ear in part to address longstanding debates over the role of established features of peripheral neural coding on pitch perception. We view this approach as sensible on the grounds that the evolution of the cochlea was plausibly influenced by many different natural behaviors, such that it is more appropriately treated as a constraint on a model of pitch rather than a model stage to be derived along with the rest of the model. However, one could in principle optimize peripheral representations for tasks, and this could yield additional insights. One could also in principle incorporate additional stages of peripheral physiology, which might similarly provide constraints on pitch perception.

Our model shares many of the commonly-noted limitations of deep neural networks as models of the brain (Lake et al., 2017; Richards et al., 2019; Lindsay, 2020). Our optimization procedure is not a model of biological learning and/or evolution, but rather provides a way to obtain a system that is optimized for the training conditions given a particular peripheral representation of sound. Biological organisms are almost certainly not learning to estimate F0 from thousands of explicitly labeled examples, and in the case of pitch may leverage their vocal ability to produce harmonic stimuli to hone their perceptual mechanisms. These differences could cause the behavior of biological systems to deviate from optimized neural networks in some ways.

The neural network architectures we used here are also far from fully consistent with biology, being only a coarse approximation to neural networks in the brain. Although numerous similarities have been documented between trained neural network representations and brain representations (Yamins et al., 2014; Guclu and van Gerven, 2015; Eickenberg et al., 2017; Kell et al., 2018), and although we saw some such similarities ourselves in the network's activations (Fig. 8C), the inconsistencies with biology could in principle lead to behavioral differences compared to humans.

And although our approach is inspired by classical ideal observer models, the model class and optimization methods likely bias the solutions to some extent, and are not provably optimal like classic ideal observer models. Nonetheless, the relatively good match to available data suggests that the optimization is sufficiently successful as to be useful for our purposes.

Future directions

The model we developed here performs a single task – that of estimating the F0 of a short sound. Human pitch behavior is often substantially more complex, in part because information is conveyed by how the F0 changes over time, as in prosody (Tang et al., 2017) or melody (Dowling and Fujitani, 1971). This "relative pitch" often appears to involve a comparison of F0s estimated at different time points (i.e., the way that we simulated such judgments with our model). But in other cases relative pitch involves comparisons of the spectrum rather than the F0 (McPherson and McDermott, 2018, 2020). Relative pitch judgments are also often biased by changes in the timbre of a sound (Borchert et al., 2011; Allen and Oxenham, 2014), for reasons that are not well understood. The framework used here could help to develop normative understanding of such effects, by incorporating additional relative pitch tasks and assessing the behavioral characteristics that result.

Pitch is also believed to be used in the service of more complicated auditory tasks, such as voice recognition (Latinus and Belin, 2011; McPherson and McDermott, 2018; Lavan et al., 2019) or the perception of stress in spoken language (Shattuck-Hufnagel and Turk, 1996; Cutler et al., 1997). Moreover, we often must estimate the F0 of a sound amid other sounds that have their own F0s. F0 is used to track target voices in such situations (Darwin et al., 2003; Woods and McDermott, 2015; Popham et al., 2018), helping to solve the cocktail party problem, and to distinguish multiple melodic lines in polyphonic music (Pressnitzer et al., 2011). Representations of harmonic sounds also appear intimately related to musical harmony (Terhardt, 1974; McDermott et al., 2010b). Training models on more complicated tasks involving speech or music and then interrogating representations of F0 could help generate hypotheses about the extent to which there may be specialized pitch mechanisms for different aspects of audition. Deep neural networks that perform more complex pitch tasks might also exhibit multiple stages of pitch representations that could provide insight into putative hierarchical stages of auditory cortex (Patterson et al., 2002; Bizley et al., 2013; Bizley and Cohen, 2013; Kell et al., 2018).

The approach we used here has natural extensions to understanding other aspects of hearing (Lorenzi et al., 2006), in which similar questions about the roles of peripheral cues have remained unresolved. Our methods could also be extended to investigate hearing impairment, which can be simulated with alterations to standard models of the cochlea (Heinz et al., 2001c; Zilany and Bruce, 2006) and which often entails particular types of deficits in pitch perception (Arehart and Burns, 1999; Bernstein and Oxenham, 2006). Prostheses such as cochlear implants are another natural application of task-optimized modeling. Current implants restore some aspects of hearing relatively well, but pitch perception is not one of them (Gfeller et al., 2007). Models optimized with different types of simulated cochlear implant input could clarify the patterns of behavior to expect under different modes of electrical stimulation.

There is also growing evidence for species differences in pitch perception (Osmanski et al., 2013; Shofner and Chaney, 2013; Walker et al., 2019). Our approach could be used to relate species differences in perception to species differences in the cochlea (Joris et al., 2011) or to differences in the acoustic environment and/or tasks a species may be optimized for. More generally, the results here illustrate how supervised machine learning enables normative analysis in domains where traditional ideal observers are intractable, an approach that is broadly applicable outside of pitch and audition.

METHODS

Natural sounds training dataset

The main training set consisted of 50ms excerpts of speech and musical instruments. This duration was chosen to produce accurate pitch perception in human listeners (White and Plack, 1998), but short enough that the F0 would be relatively stable even in natural sounds such as speech that have time-varying F0s. The F0 label for a training example was estimated from a “clean” speech or music excerpt. These excerpts were then superimposed on natural background noise.

Speech and music clips

We used STRAIGHT (Kawahara et al., 2008) to compute time-varying F0 and periodicity traces for sounds in several large corpora of recorded speech and instrumental music: Spoken Wikipedia Corpora (SWC) (Köhn et al., 2016), Wall Street Journal (WSJ) (Paul and Baker, 1992), CMU Kids Corpus (Eskenazi et al., 1997), CSLU Kids Speech (Shobaki et al., 2007), NSynth (Engel et al., 2017), and RWC (Goto et al., 2003). STRAIGHT provides accurate estimates of the F0 provided the background noise is low, as it was in each of the corpora. Musical instrument recordings were notes from the chromatic scale, and thus were spaced roughly in semitones. To ensure that sounds would span a continuous range of F0s, we randomly pitch-shifted each instrumental music recording by a small amount (up to ±3% F0 via resampling).

Source libraries were constructed for each corpus by extracting all highly periodic (time-averaged periodicity level > 0.8) and non-overlapping 50ms segments from each recording. We then generated our natural sounds dataset by sampling segments with replacement from these source libraries to uniformly populate 700 log-spaced F0 bins between 80 Hz and 1000 Hz (bin width = 1/16 semitones = 0.36% F0). 50ms segments were assigned to bins according to their time-averaged F0. The resulting training dataset consisted of 3000 clips per F0 bin for a total of 2.1 million exemplars. The relative contribution of each corpus to the final dataset was constrained both by the number of segments per F0 bin available in each source library (the higher the F0, the harder it is to find speech clips) and our effort to use audio from many different speakers, instruments, and corpora. The composition we settled on is:

- F0 bins between 80 Hz and 320 Hz
    - 50% instrumental music (1000 NSynth and 500 RWC clips per bin),
    - 50% adult speech (1000 SWC and 500 WSJ clips per bin)
- F0 bins between 320 Hz and 450 Hz
    - 50% instrumental music (1000 NSynth and 500 RWC clips per bin)
    - 50% child speech (750 CSLU and 750 CMU clips per bin)
- F0 bins between 450 Hz and 1000 Hz
    - 100% instrumental music (2500 NSynth and 500 RWC clips per bin)

Background noise

To make the F0 estimation task more difficult and to simulate naturalistic listening conditions, each stimulus in the training dataset was embedded in natural background noise. Noise source clips were taken from a subset of the AudioSet corpus (Gemmeke et al., 2017), screened to remove nonstationary sounds (e.g., speech or music). The screening procedure involved measuring auditory texture statistics (envelope means, correlations, and modulation power in and across cochlear frequency channels) (McDermott and Simoncelli, 2011) from all recordings, and discarding segments over which these statistics were not stable in time, as in previous studies (McWalter and McDermott, 2018; Kell and McDermott, 2019b). To ensure the F0 estimation task remained well-defined for the noisy stimuli, background noise clips were also screened for periodicity by computing their autocorrelation functions. Noise clips with peaks greater than 0.8 at lags greater than 1ms in their normalized autocorrelation function were excluded.

The signal-to-noise ratio for each training example was drawn uniformly between -10 dB and +10 dB. Overall stimulus presentation levels were drawn uniformly between 30 dB SPL and 90 dB SPL. All training stimuli were sampled at 32 kHz.

Peripheral auditory model

The Bruce et al. (2018) auditory nerve model was used to simulate the peripheral auditory representation of every stimulus. This model was chosen because it captures many of the complex response properties of auditory nerve fibers and has been extensively validated against electrophysiological data from cats (Zilany and Bruce, 2006; Zilany et al., 2009, 2014; Bruce et al., 2018). Stages of peripheral signal processing in the model include: a fixed middle-ear filter, a nonlinear cochlear filter bank to simulate level-dependent frequency tuning of the basilar membrane, inner and outer hair cell transduction functions, and a synaptic vesicle release/re-docking model of the inner hair cell to auditory nerve fiber synapse. Although the model's responses have only been directly compared to recordings made in nonhuman animals, certain model parameters have been inferred for humans (such as the bandwidths of cochlear filters) on the basis of behavioral and otoacoustic measurements (Shera et al., 2002).

Because the majority of auditory nerve fibers, especially those linked to feedforward projections to higher auditory centers, have high spontaneous firing rates (Liberman, 1991; Carney, 2018), we used exclusively high spontaneous rate fibers (70 spikes/s) in most of our experiments. To control for the possibility that spontaneous auditory nerve fiber activity could influence pitch behavior (for instance, at conversational speech levels, firing rates of high spontaneous rate fibers are typically saturated, which may degrade excitation pattern cues to F0) we trained and tested the 10 best-performing networks from the architecture search using exclusively low spontaneous rate fibers (0.1 spikes/s). The average results for these networks are shown in Supplemental Fig. 9. We found that psychophysical behavior was qualitatively unaffected by nerve fiber spontaneous rate. These results suggested to us that high spontaneous rate fibers were sufficient to yield human-like pitch behavior, so we excluded low spontaneous rate fibers from all other analyses.

In most cases the input to the neural network models consisted of the instantaneous firing rate responses of 100 auditory nerve fibers with characteristic frequencies spaced uniformly on an ERB-number scale (Glasberg and Moore, 1990) between 125 Hz and 14000 Hz. Firing rates were used to approximate the information that would be available in a moderate group of spiking nerve fibers receiving input from the same inner hair cell. The use of 100 frequency channels primarily reflects computational constraints (CPU time for simulating peripheral representations, storage costs, and GPU memory for training), but we note that this number is similar to that used in other auditory models with cochlear front-ends (Chi et al., 2005). In one training condition, we confirmed that increasing the number of channels by a factor of 10 had little effect (Supplemental Figs. 4 and 5), and given that 100 channels was sufficient to obtain model thresholds on par with those of humans, it appears that there is little benefit to additional channels for the task we studied.

To prevent the stimuli being dominated by sound onset/offset effects, each stimulus was padded with 100ms of the original waveform before being passed through the nerve model. The resulting 150ms auditory nerve responses were resampled to 20 kHz and the middle 50 ms was excerpted, leaving a 100-fiber by 1000-timestep array of instantaneous firing rates.

Source code for the Bruce et al. (2018) auditory nerve model is publically available (https://www.ece.mcmaster.ca/~ibruce/zbcANmodel/zbcANmodel.htm). We developed a Python wrapper around the model, which supports flexible manipulation of cochlear filter bandwidths and the upper limit of phase-locking (code available upon request).

Deep neural network models

The 100-by-1000 simulated auditory nerve representations were passed into convolutional neural networks (CNNs), each consisting of a series of feedforward layers. These layers were hierarchically organized and instantiated a number of relatively simple operations: linear convolution, pointwise nonlinear rectification, weighted average pooling, batch normalization, linear transformation, dropout regularization, and softmax classification.

The last layer of each network performed F0 classification. We opted to use classification with narrow F0 bins rather than regression in order to soften the assumption that output F0 distributions should be unimodal for a given stimulus. For example, an octave error would incur a very large penalty under standard regression loss functions (e.g., L1 or L2), which measure the distance between the predicted and target F0. Classification loss functions, such as the softmax cross-entropy used here, penalize all misclassifications equally. In preliminary work, we found classification networks were empirically easier to train than regression networks and yielded smaller median F0 errors.

The precision of the network's F0 estimate is limited by the bin width of the output layer (and by the precision of the training set labels). We chose a bin width of 1/16 semitones (0.36%). We found empirically that the median F0 estimation error increased for bins wider than this value, and did not improve for narrower bins (Supplemental Fig. 10A). This could reflect the limits of the F0 labels the network was trained on. As it happened, with this bin width it was possible to attain discrimination thresholds for synthetic tones that were on par with the best thresholds typically measured in human listeners (~0.1-0.4%) (Hoekstra, 1979; Houtsma and Smurzynski, 1990) for some model architectures and auditory nerve settings. Discrimination thresholds were worse for wider classification bins (Supplemental Fig. 10B). We otherwise observed no qualitative change in the pattern of psychophysical results as the bin width was changed. The bin width might thus be considered analogous to decision noise that is sometimes added to models to match human performance, though our choice of bin width appears near-optimal for the dataset we worked with. We note that discrimination thresholds for synthetic tones were also plausibly limited by the similarity of the tones to the training data.

Definitions of constituent neural network operations

Convolutional layer: A convolutional layer implements the linear convolution of a bank of N_k two-dimensional filter kernels with an input X. Convolution performs the same operation at each point in the input, which for the 2D auditory nerve representations we used entails convolving in both time and frequency. Convolution in time is natural for models of sensory systems as their input has translation-invariant temporal statistics. Because translation invariance does not hold in frequency, convolution in frequency is less obviously natural. However, many types of sound signals are well described by approximate translation invariance in local neighborhoods of the frequency domain, and auditory models can often be described as imposing convolution in frequency (Dau et al., 1997; Chi et al., 2005). Moreover, imposing convolution greatly reduces the number of parameters to be learned, and neural network models often train more readily when convolution in frequency is used, suggesting that it is a useful form of model regularization.

The input is a three-dimensional array with shape $[M_f, M_{t,}, M_k]$. For the first convolutional layer in our networks, the input shape was $[100, 1000, 1]$, corresponding to 100 frequency bins (nerve fibers), 1000 time bins, and a placeholder 1 in the filter kernel dimension.

A convolutional layer is defined by five parameters:

1. h : height of the convolutional filter kernels (number of filter taps in the frequency dimension)
2. w : width of the convolutional filter kernels (number of filter taps in the time dimension)
3. N_k : number of different convolutional filter kernels
4. W : Trainable weights for each of the N_k filter kernels; W has shape $[h, w, M_k, N_k]$
5. B : Trainable bias vector with shape $[N_k]$

The output of the convolutional layer Y has shape $[N_f, N_t, N_k]$ and is given by:

$$Y[n_f, n_t, n_k] = B[n_k] + \sum_{i=1,\ j=1,\ m_k=1}^{h,\ w,\ M_k} W[i, j, m_k n_k] \odot X[n_f + i - \lfloor h/2 \rfloor, n_t + j - \lfloor w/2 \rfloor, m_k]$$

where $\odot$ denotes pointwise multiplication and $\lfloor \cdot / \cdot \rfloor$ denotes integer division. Convolutional layers all used a stride of 1 (i.e., non-strided convolution) and "valid" padding, meaning filters were only applied at positions where every element of the kernel overlaps the input. Due to this boundary handling, the frequency and time dimensions of the output were smaller than those of the input: $N_f = M_f - h + 1$ and $N_t = M_t - w + 1$.

Pointwise nonlinear rectification: To learn a nonlinear function, a neural network must contain nonlinear operations. We incorporate nonlinearity via the rectified linear unit (ReLU) activation function, which is applied pointwise to every element x in some input X:

$$ReLU(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Weighted average pooling: Pooling operations reduce the dimensionality of inputs by aggregating information across adjacent frequency and time bins. To reduce aliasing in our networks (which would otherwise occur from downsampling without first lowpass-filtering), we used weighted average pooling with Hanning windows (Hénaff and Simoncelli, 2016; Feather et al., 2019). This pooling operation was implemented as the strided convolution of a two-dimensional (frequency-by-time) Hanning filter kernel H with an input X:

$$Y = H *_{s_f, s_t} X$$

where $*$ denotes convolution and s_f and s_t indicate the stride length in frequency and time, respectively. The Hanning window H had a stride-dependent shape $[h_f, h_t]$, where

$$h_f = \begin{cases} 1 & s_f = 1 \\ 4 \cdot s_f & s_f > 1 \end{cases} \quad and \quad h_t = \begin{cases} 1 & s_t = 1 \\ 4 \cdot s_t & s_t > 1 \end{cases}$$

For an input X with shape $[N_f, N_t, N_k]$, the shape of the output Y is $[N_f/s_f, N_t/s_t, N_k]$. Note that when either s_f or s_t is set to 1, there is no pooling along the corresponding dimension.

Batch normalization: Batch normalization is an operation that normalizes its inputs in a pointwise manner using running statistics computed from every batch of training data. Normalizing activations between layers greatly improves DNN training efficiency by reducing the risk of exploding and vanishing gradients: small changes to network parameters in one layer are less likely to be amplified in successive layers if they are separately normalized (Ioffe and Szegedy, 2015). For every batch of inputs B during training, the pointwise batch mean (μ_B) and batch variance (σ_B^2) are computed and then used to normalize each input $X_b \in B$:

$$X_{b,normalized} = \gamma\left(\frac{X_b - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}\right) + \beta$$

where all operations are applied pointwise, $\epsilon = 0.001$ to prevent division by zero, and γ and β are learnable scale and offset parameters. Throughout training, single-batch statistics are used to update the running mean (μ_{total}) and variance (σ_{total}^2). During evaluation mode, $X_{b,normalized}$ is computed using μ_{total} and σ_{total}^2 in place of μ_B and σ_B^2.

Fully-connected layer: A fully-connected (or dense) layer applies a linear transformation to its input without any notion of localized frequency or time. An input X with shape $[N_f, N_t, N_k]$, is first reshaped to a vector X_{flat} with shape $[N_f \cdot N_t \cdot N_k]$. Then, X_{flat} is linearly transformed to give an output Y with shape $[N_{out}]$:

$$Y_{out}[n_{out}] = B[n_{out}] + \sum_{n_{in}=1}^{N_f \cdot N_t \cdot N_{ch}} W[n_{out}, n_{in}] \cdot X_{flat}[n_{in}]$$

where B is a bias vector with shape $[N_{out}]$ and W is a weight matrix with shape $[N_{out}, N_{in}]$. The values of B and W are learned during the optimization procedure.

Dropout regularization: The dropout operation receives as input a vector X with shape $[N_{in}]$ and randomly selects a fraction (r) of its values to set to zero. The remaining values are scaled by $1/(1-r)$, so the expected sum over all outputs is equal to the expected sum over all inputs. The $r \cdot N_{in}$ positions in X that get set to zero are chosen at random for every new batch of data. Dropout is commonly used to reduce overfitting in artificial neural networks (Srivastava et al., 2014). It can be thought of as a form of model averaging across the exponentially many sub-networks generated by zeroing-out different combinations of units. All of our networks contained exactly one dropout operation immediately preceding the final fully-connected layer. We used a dropout rate of 50% during both training and evaluation.

Softmax classifier: The final operation of every network is a softmax activation function, which receives as input a vector X of length $N_{classes}$ (equal to the number of output classes; 700 in our case). The input vector is passed through a normalized exponential function to produce a vector Y of the same length:

$$Y[n_{out}] = \frac{\exp\left(X[n_{out}]\right)}{\sum_{n_{classes}=1}^{N_{classes}} \exp(X[n_{classes}])}$$

The values of the output vector are all greater than zero and sum to one. Y is interpreted as a probability distribution over F0 classes for the given input sound.

Model optimization

Architecture search

All of our CNN architectures had the general form of one to eight convolutional layers plus one to two fully-connected layers. Each convolutional layer was always immediately followed by three successive operations: ReLU activation function, weighed average pooling, and batch normalization. Fully-connected layers were always situated at the end of the network, after the last convolution-ReLU-pooling-normalization block. The final fully-connected layer was always immediately followed by the softmax classifier. For architectures with two fully-connected layers, the first fully-connected layer was followed by a ReLU activation function and a batch normalization operation. In our analyses, we sometimes grouped networks by their number of convolutional layers (e.g., single vs. multi-convolutional-layer networks; Supplemental Fig. 3) regardless of the number of fully-connected layers. When we refer to network "activations" in a given convolutional layer (Fig. 8), we always mean the outputs of the ReLU activation function immediately following that convolutional layer.

Within the family of models considered, we generated 400 distinct CNN architectures by randomly sampling from a large space of hyperparameters. The number of convolutional layers was first uniformly drawn from 1 to 8. Within each layer, the number and dimensions of convolutional filter kernels were then sampled based on the size of the layer's input. The number of filter kernels in the first layer was either 16, 32, or 64 (each sampled with probability=1/3). The number of kernels in each successive layer could either increase by a factor of 2 (probability=1/2), stay the same (probability=1/3), or decrease by a factor of 2 (probability=1/6) relative to the previous layer. Frequency dimensions of the filter kernels were integers sampled uniformly between 1 and $N_f/2$, where N_f is the frequency dimension of the layer's input. Time dimensions of the filter kernels were integers sampled uniformly between $N_t/20$ and $N_t/2$, where N_t is the time dimension of the layer's input. These sampling ranges tended to produce rectangular filters (longer in the time dimension than the frequency dimension), especially in the early layers. We felt this was a reasonable design choice given the rectangular dimensions of the input (100-by-1000, frequency-by-time). To limit the memory footprint of the generated CNNs, we imposed 16 and 1024 as lower and upper bounds on the number of kernels in a single layer and capped the frequency × time area of convolutional filter kernels at 256.

The stride lengths for the weighted average pooling operations after each convolutional layer were also sampled from distributions. Pooling stride lengths were drawn uniformly between 1 and 4 for the frequency dimension and 1 and 8 for the time dimension. The existence (probability=1/2) and size (128, 256, 512, or 1024 units) of a penultimate fully-connected layer were also randomly sampled. The final fully-connected layer always contained 700 units to support classification into the 700 F0 bins.

Network training

All 400 network architectures were trained to classify F0 of our natural sounds dataset via stochastic gradient descent with gradients computed via back-propagation. We used a batch size of 64 and the ADAM optimizer with a learning rate of 0.0001 (Kingma and Ba, 2014). Network weights were trained using 80% of the dataset and the remaining 20% was held-out as a validation set. Performance on the validation set was measured every 5000 training steps and, to reduce overfitting, training was stopped once classification accuracy stopped increasing by at least 0.5% every 5000 training steps. Training was also stopped for networks that failed to achieve 5% classification accuracy after 10000 training steps. Each network was able to reach these early-stopping criteria in less than 48 hours when trained on a single NVIDIA Tesla V100 GPU.

To ensure conclusions were not based on the idiosyncrasies of any one particular DNN architecture, we selected the 10 architectures that produced the highest validation set accuracies to use as our experimental subjects (collectively referred to as 'the model'). We re-trained all 10 architectures for each manipulation of the peripheral auditory model (Figs. 4 and 5) and the training set sound statistics (Figs. 6 and 7). The 10 different network architectures are described in Supplemental Table 1. Tensorflow code to implement these models, including their trained weights, will be made available on the McDermott Lab website (http://mcdermottlab.mit.edu).

Network psychophysics

To investigate network pitch behavior, we simulated a set of classic psychophysical experiments on all trained networks. The general procedure was to (1) pass each experimental stimulus through a network, (2) compute F0 discrimination thresholds or shifts in the "perceived" F0 (depending on the experiment) from network predictions, and (3) compare network results to published data from human listeners tested on the same stimulus manipulations. We selected five psychophysical experiments, denoted A through E in the following sections to align with Fig. 2, which contains schematics of the stimulus manipulations in each experiment. We attempted to reproduce stimuli from these studies as closely as possible, though some modifications were necessary (e.g., all stimuli were truncated to 50 ms for our networks). Because the cost of running experiments on networks is negligible, networks were tested on many more (by 1 to 3 orders of magnitude) stimuli than were human participants. Human data from the original studies was obtained either directly from the original authors (experiments A-B) or by extracting data points from published figures (experiments C-E) using Engauge Digitizer (Mitchell et al.).

Experiment A: effect of harmonic number and phase on pitch discrimination

Experiment A reproduced the stimulus manipulation of Bernstein and Oxenham (2005) to measure F0 discrimination thresholds as a function of lowest harmonic number and phase.

Stimuli: Stimuli were harmonic complex tones, bandpass-filtered and embedded in masking noise to control the lowest audible harmonic, and whose harmonics were in sine or random phase. In the original study, the bandpass filter was kept fixed while the F0 was roved to set the lowest harmonic number. Here, to measure thresholds at many combinations of F0 and lowest harmonic number, we roved both the F0 and the location of the filter. We took the 4th-order Butterworth filter (2500 to 3500 Hz -3 dB passband) described in the original study and translated its frequency response along the frequency axis to set the lowest audible harmonic for a given stimulus. Before filtering, the level of each individual harmonic was set to 48.3 dB SPL, which corresponds to 15 dB above the masked thresholds of the original study's normal-hearing participants. After filtering, harmonic tones were embedded in modified uniform masking noise (Bernstein and Oxenham, 2003), which has a spectrum that is flat (15 dB/Hz SPL) below 600 Hz and rolls off at 2 dB/octave above 600 Hz. This noise was designed to ensure that only harmonics within the filter's -15 dB passband are presented above participants' masked audibility thresholds.

Human experiment: The previously published human F0 discrimination thresholds were measured from 5 normal-hearing participants (3 female) between the ages of 18 and 21 years old, all self-described amateur musicians with at least 5 years of experience (Bernstein and Oxenham, 2005). Each participant completed 4 adaptive tracks per condition.

Model experiment: The F0 discrimination experiment we ran on each network had 600 conditions corresponding to all combinations of 2 harmonic phases (sine or random), 30 lowest harmonic numbers ($n_{low} = 1, 2, 3 \dots 30$), and 10 reference F0s ($F_{0,ref}$) spaced uniformly on a logarithmic scale between 100 and 300 Hz. Within each condition, the network was evaluated on 121 stimuli with slightly different F0s (within ±6% of $F_{0,ref}$) but the same bandpass filter. The filter was positioned such that the low frequency cutoff of its -15 dB passband was equal to $n_{low} \times F_{0,ref}$. On the grounds that human listeners likely employ a strong prior that stimuli should have fairly similar F0s within single trials of a pitch discrimination experiment, we limited network F0 predictions to fall within a one-octave range (centered at $F_{0,ref}$). We simulated a two-alternative forced choice paradigm by making all 7260 possible pairwise comparisons between the 121 stimuli. In each trial, we asked if the network predicted a higher F0 for the stimulus in the pair with the higher F0 (i.e., if the network correctly identified which of two stimuli had a higher F0). A small random noise term was used to break ties when the network predicted the same F0 for both stimuli. F0 discrimination judgments across trials were then used to construct a psychometric function plotting the percentage of correct trials as a function of %F0 difference between two stimuli. We combined psychometric functions across the 10 reference F0s by pooling trials with the same harmonic phase and lowest harmonic number. Network thresholds were thus based on 1210 stimuli (72600 pairwise F0 discriminations) per condition. Normal cumulative distribution functions were fit to the 60 (2 phase conditions x 30 lowest harmonic numbers) resulting psychometric functions. To match human F0 discrimination thresholds, which were measured with a 2-down-1-up adaptive algorithm (Levitt, 1971), we defined the network F0 discrimination threshold as the F0 difference (in percent, capped at 100%) that yielded 70.7% of trials correct.

Human-model comparison: Bernstein and Oxenham (2005) reported very similar F0 discrimination thresholds for two different spectral conditions ("low spectrum" with 2500 to 3500 Hz filter passband and "high spectrum" with 5000 to 7000 Hz filter passband).

To simplify presentation and because our network experiment measured average thresholds across a wide range of bandpass filter positions, here we report their human data averaged across spectral condition.

We quantified the similarity between human and network F0 discrimination thresholds as the correlation between vectors of analogous data points. The network vector contained 60 F0 discrimination thresholds, one for each combination of phase and lowest harmonic number. To get a human vector with 60 analogous F0 discrimination thresholds, we a) linearly interpolated the human data between lowest harmonic numbers and b) extrapolated that F0 discrimination thresholds are constant for lowest harmonic numbers between 1 and 5 (supported by other published data (Bernstein and Oxenham, 2003; Norman-Haignere et al., 2013)). We then computed the Pearson correlation coefficient between log-transformed vectors of human and network thresholds.

Experiment B: pitch of alternating-phase harmonic complexes

Experiment B reproduced the stimulus manipulation of Shackleton and Carlyon (1994) to test if our networks exhibit pitch-doubling for alternating-phase harmonic stimuli.

Stimuli: Stimuli consisted of consecutive harmonics (each presented at 50 dB SPL) summed together in alternating sine/cosine phase: odd-numbered harmonics in sine phase (0° offset between frequency components) and even-numbered harmonics in cosine phase (90° offset, such that components align at their peaks). As in Experiment A, these harmonic tones were bandpass-filtered and embedded in masking noise to control which harmonics were audible. The original study used pink noise and analog filters. Here, we used modified uniform masking noise and digital Butterworth filters (designed to approximate the original passbands). We generated stimuli with three different 4th-order Butterworth filters specified by their -3 dB passbands: 125 to 625 Hz ("low harmonics"), 1375 to 1875 Hz ("mid harmonics"), and 3900 to 5400 Hz ("high harmonics"). The exact harmonic numbers that are audible in each of these passbands depends on the F0. The original study used stimuli with F0s near 62.5, 125, and 250 Hz (sometimes offset by ±4% from the nominal F0 to avoid stereotyped responses). The 62.5 Hz condition was excluded here because our networks never saw F0s below 80 Hz during training. We generated 354 stimuli with F0s near 125 Hz (120-130 Hz) and 250 Hz (240-260 Hz), in both cases uniformly sampled on a logarithmic scale, for each filter condition (2124 stimuli in total).

Human experiment: In the original experiment of Shackleton and Carlyon (1994), participants adjusted the F0 of a sine-phase control tone to match the pitch of a given alternating-phase test stimulus. The matched F0 thus gives the perceived F0 for the test stimulus. The previously published human data were obtained from 8 normal-hearing listeners who had a wide range of musical experience. Each participant made 18 pitch matches per condition.

Model experiment: To simulate the human paradigm in our model, we simply took the network's F0 prediction (within a 3-octave range centered at the stimulus F0) for the "perceived" F0 of the alternating-phase test stimulus. For each stimulus, we computed the ratio of the predicted F0 to the stimulus F0. Histograms of these frequency ratios (bin width = 2%) were generated for each of the 6 conditions (3 filter conditions × 2 nominal F0s). To simplify presentation, histograms are only shown for 2 conditions: "low harmonics" and "high harmonics", both with F0s near 125 Hz.

Human-model comparison: Shackleton and Carlyon (1994) constructed histograms from their pitch matching data, pooling responses across participants (144 pitch matches per histogram). We quantified the similarity between human and network responses by measuring linear correlations between human and network histograms for the same condition. Human histograms were first re-binned to have the same 2% bin width as network histograms. Pearson correlation coefficients were computed separately for each of the 6 conditions and then averaged across conditions to give a single number quantifying human-network similarity.

Experiment C: pitch of frequency-shifted complexes

Experiment C reproduced the stimulus manipulation of Moore and Moore (2003) to test if our networks exhibited pitch shifts for frequency-shifted complexes.

Stimuli: Stimuli were modifications of harmonic complex tones with consecutive harmonic frequencies in cosine phase. We imposed three different F0-dependent spectral envelopes -- as described by Moore and Moore (2003) -- on the stimuli. The first, which we termed the "low harmonics" spectral envelope had a flat 3-harmonic wide passband centered at the 5th harmonic. The second (termed "mid harmonics") had a flat 5-harmonic wide passband centered at the 11th harmonic. The third (termed "high harmonics") had a flat 5-harmonic wide passband centered at the 16th harmonic. All three of these spectral envelopes had sloping regions flanking the flat passband. Amplitudes (relative to the flat passband) at a given frequency F in the sloping regions were always given by $(10^x - 1)/9$ where $x = 1 - |(F - F_e)/1.5F0|$ and F_e is the edge of the flat region. The amplitude was set to zero for $x \leq 0$.

For a given F0 and fixed spectral envelope, we made stimuli inharmonic by shifting every component frequency by a common offset in Hz specified as a percentage of the F0. As a concrete example, consider a stimulus with F0 = 100 Hz and the "low harmonics" spectral envelope. This stimulus contains nonzero energy at 200, 300, 400, 500, 600, and 700 Hz. Frequency-shifting this harmonic tone by +8% of the F0 results in an inharmonic tone with energy at 208, 308, 408, 508, 608, and 708 Hz. For each of the three spectral envelopes, we generated stimuli with component shifts of +0, +4, +8, +12, +16, +20, and +24 %F0. For each combination of spectral envelope and component shift, we generated stimuli with 3917 nominal F0s spaced log-uniformly between 80 and 480 Hz (83391 stimuli in total). Stimuli were presented at overall levels of 70 dB SPL to match the original study.

Human experiment: Moore and Moore (2003) used a pitch matching paradigm to allow listeners to report the perceived F0s for frequency-shifted complex tones. 5 normal-hearing listeners (all musically trained) between the ages of 19 and 31 years old participated in the study. Each participant made 108 pitch matches.

Model experiment: For the model experiment, we again took network F0 predictions for the 83391 frequency-shifted complexes as the "perceived" F0s. F0 predictions were limited to a one-octave range centered at the target F0 (the F0 of the stimulus before frequency-shifting). We summarize these values as shifts in the predicted F0, which are given by $(F0_{predicted} - F0_{target})/F0_{target}$. These shifts are reported as the median across all tested F0s and plotted as a function of component shift and spectral envelope. To simplify presentation, results are only shown for two spectral envelopes, "low harmonics" and "high harmonics".

Human-model comparison: Moore and Moore (2003) reported quantitatively similar patterns of pitch shifts for the three F0s tested (100, 200, and 400 Hz). To simplify presentation and because we used many more F0s in the network experiment, here we present their human data averaged across F0 conditions. We quantified the similarity between human and network pitch shifts as the Pearson correlation coefficient between vectors of analogous data points. The network vector contained 21 median shifts, one for each combination of spectral envelope and component shift. To obtain a human vector with 21 analogous pitch shifts, we linearly interpolated the human data between component shifts.

Experiment D: pitch of complexes with individually mistuned harmonics

Experiment D reproduced the stimulus manipulation of Moore et al. (1985) to test if our networks exhibit pitch shifts for complexes with individually mistuned harmonics.

Stimuli: Stimuli were modifications of harmonic complex tones containing 12 equal-amplitude harmonics (60 dB SPL per component) in sine phase. We generated such tones with F0s near 100 Hz, 200 Hz, and 400 Hz (178 F0s uniformly spaced on a logarithmic scale within ±4% of each nominal F0). Stimuli were then made inharmonic by shifting the frequency of a single component at a time. We applied +0, +1, +2, +3, +4, +6, and +8 % frequency shifts to each of the following harmonic numbers: 1, 2, 3, 4, 5, 6, and 12. In total there were 178 stimuli in each of the 63 conditions (3 nominal F0s × 7 component shifts × 7 harmonic numbers).

Human experiment: Moore et al. (1985) used a pitch-matching paradigm in which participants adjusted the F0 of a comparison tone to match the perceived pitch of the complex with the mistuned harmonic. Three participants (all highly experienced in psychoacoustic tasks) completed the experiment. Participants each made 10 pitch matches per condition tested.

Model experiment: For the model experiment, we used the procedure described for Experiment C to measure shifts in the network's predicted F0 for all 11214 stimuli. Shifts were averaged across similar F0s (within ±4% of the same nominal F0) and reported as a function of component shift and harmonic number. To simplify presentation, results are only shown for F0s near 200 Hz. Results were similar for F0s near 100 and 400 Hz.

Human-model comparison: We compared the network's pattern of pitch shifts to those averaged across the three participants from Moore et al. (1985). Human-model similarity was again quantified as the Pearson correlation coefficient between vectors of analogous data points. The network vector contained 63 mean shift values corresponding to the 63 conditions. Though Moore et al. (1985) did not report pitch shifts for the 12th harmonic, they explicitly stated they were unable to measure significant shifts when harmonics above the 6th were shifted. We thus inferred pitch shifts were always equal to zero for the 12th harmonic when compiling the vector of 63 analogous pitch shifts. We included this condition because some networks exhibited pitch shifts for high-numbered harmonics and we wanted our similarity metric to be sensitive to this deviation from human behavior.

Experiment E: necessity of correct tonotopic representation for pitch discrimination

Experiment E measured network discrimination thresholds for pure tones and transposed tones as described by Oxenham et al. (2004).

Stimuli: Transposed tones were generated by multiplying a half-wave rectified low-frequency sinusoid (the "envelope") with a high frequency sinusoid (the "carrier"). Before multiplication, the "envelope" was lowpass filtered (4th order Butterworth filter) with a cutoff frequency equal to 20% of the carrier frequency. To match the original study, we used carrier frequencies of 4000, 6350, and 10080 Hz. For each carrier frequency, we generated 6144 transposed tones with envelope frequencies spaced uniformly on a logarithmic scale between 80 and 320 Hz. We also generated 6144 pure tones with frequencies spanning the same range. All stimuli were presented at 70 dB SPL and embedded in the same modified uniform masking noise as Experiment A. The original study embedded only the transposed tones in lowpass-filtered noise to mask distortion products. To ensure that the noise would not produce differences in the model's performance for the two types of stimuli, we included it for pure tones as well.

Human experiment: Oxenham et al. (2004) reported discrimination thresholds for these same 4 conditions (transposed tones with 3 different carrier frequencies + pure tones) at 5 reference frequencies between 55 and 320 Hz. Data was collected from 4 young (<30 years old) adult participants who had at least 1 hour of training on the frequency discrimination task. Discrimination thresholds were based on 3 adaptive tracks per participant per condition.

Model experiment: The procedure for measuring network discrimination thresholds for pure tones was analogous to the one used in Experiment A. We first took network F0 predictions (within a one-octave range centered at the stimulus frequency) for all 6144 stimuli. We then simulated a two-alternative forced choice paradigm by making pairwise comparisons between predictions for stimuli with similar frequencies (within 2.7

semitones of 5 "reference frequencies" spaced log-uniformly between 80 and 320 Hz). For each pair of stimuli, we asked if the network correctly predicted a higher F0 for the stimulus with the higher frequency. From all trials at a given reference frequency, we constructed a psychometric function plotting the percentage of correct trials as a function of percent frequency difference between the two stimuli. Normal cumulative distribution functions were fit to each psychometric function and thresholds were defined as the percent frequency difference (capped at 100%) that yielded 70.7% correct. Each threshold was based on 233586 pairwise discriminations made between 684 stimuli. The procedure for measuring thresholds with transposed tones was identical, except that the correct answer was determined by the envelope frequency rather than the carrier frequency. Thresholds were measured separately for transposed tones with different carrier frequencies. To simplify presentation, we show transposed tone thresholds averaged across carrier frequencies (results were similar for different carrier frequencies).

Human-model comparison: We again quantified human-network similarity as the Pearson correlation coefficient between vectors of analogous log-transformed discrimination thresholds. Both vectors contained 20 discrimination thresholds corresponding to 5 reference frequencies × 4 stimulus classes (transposed tones with 3 different carrier frequencies + pure tones). Human thresholds were linearly interpolated to estimate thresholds at the same reference frequencies used for networks. This step was necessary because our networks were not trained to make F0 predictions below 80 Hz.

Additional experiment: effect of stimulus level on frequency discrimination

To investigate how phase-locking in the periphery contributes to the level-robustness of pitch perception, we measured pure tone frequency discrimination thresholds from our networks as a function of stimulus level (Fig. 4D).

Stimuli: We generated pure tones with 6144 frequencies spaced uniformly on a logarithmic scale between 200 and 800 Hz. Tones were embedded in the same modified uniform masking noise as Experiment A. The signal-to-noise ratio was fixed at 20 dB and the overall stimulus levels were varied between 10 and 100 dB SPL in increments of 10 dB.

Human experiment: Wier et al. (1977) reported frequency discrimination thresholds for pure tones in low-level broadband noise as a function of frequency and sensation level (i.e., the amount by which the stimulus is above its detection threshold). Thresholds were measured from four participants with at least 20 hours of training on the frequency discrimination task. Participants completed four or five 2-down-1-up adaptive tracks of 100 trials per condition. Stimuli were presented at five different sensation levels: 5, 10, 20, 40, and 80 dB relative to masked thresholds in 0 dB spectrum level noise (broadband, lowpass-filtered at 10000 Hz). We averaged the reported thresholds across four test frequencies (200, 400, 600, and 800 Hz) and re-plotted them as a function of sensation level in Fig. 4E.

Model experiment: We used the same procedure used in Experiments A and E to measure frequency discrimination thresholds as a function of stimulus presentation level. The simulated frequency discrimination experiment considered all possible pairings of stimuli with similar frequencies (within 2.7 semitones). Reported discrimination thresholds were pooled across all tested frequencies (200 to 800 Hz).

Human-model comparison: Because the human results were reported in terms of sensation level rather than SPL, we did not compute a quantitative measure of the match between model and human results, and instead plot the results side-by-side for qualitative comparison.

Auditory nerve manipulations

The general procedure for investigating the dependence of network behavior on aspects of the auditory nerve representation was to (1) modify the auditory nerve model, (2) retrain networks (starting from a random initialization) on modified auditory nerve representations of the same natural sounds dataset, and (3) simulate psychophysical experiments on trained networks using modified auditory nerve representations of the same test stimuli. We used this approach to investigate the dependence of network pitch behavior on both temporal and “place” information in the auditory nerve representation.

Manipulating fine timing information in the auditory nerve

We modified the upper frequency limit of phase-locking in the auditory nerve by adjusting the cutoff frequency of the inner hair cell lowpass filter within the auditory nerve model. By default, the lowpass characteristics of the inner hair cell's membrane potential are modeled as a 7th order filter with a cutoff frequency of 3000 Hz (Bruce et al., 2018). We trained and tested networks with this cutoff frequency set to 50, 250, 1000, 3000, 6000, and 9000 Hz. In each of these cases, the sampling rate of the peripheral representation used as input to the networks was 20 kHz so that spike-timing information would not be limited by the Nyquist frequency.

When the inner hair cell cutoff frequency is set to 50 Hz, virtually all temporal information in the short-duration stimuli we used was eliminated, leaving only place information (Fig. 4A). To control for the possibility that the performance characteristics of networks trained on such representations could be limited by the number of model nerve fibers (set to 100 for most of our experiments), we repeated this manipulation with 1000 auditory nerve fibers (characteristic frequencies again spaced uniformly on an ERB-number scale between 125 Hz and 14000 Hz). To keep the network architecture constant, we reduced the sampling rate to 2 kHz (which preserved all stimulus-related information due to the lowpass filter settings), yielding peripheral representations that were 1000-fiber by 100-timestep arrays of instantaneous firing rates. We then simply transposed the nerve fiber and time dimensions so that networks still operated on 100-by-1000 inputs. Note that by transposing the input representation, we effectively changed the orientation of the convolutional filter kernels. Kernels that were previously long in the time dimension and short in the nerve fiber dimension became short in the time dimension and long in the nerve fiber dimension. We saw this as desirable as it allowed us to rule out the additional possibility that the performance characteristics of networks with lower limits of phase-locking were due to convolutional kernel shapes that were optimized for input representations with high temporal fidelity and thus perhaps less suited for extracting place information (which requires pooling information across nerve fibers).

To more closely examine how the performance with degraded phase-locking (i.e., the 50 Hz inner hair cell cutoff frequency condition) might be limited by the number of model nerve fibers, we also generated peripheral representations with either 100, 250, or 500 nerve fibers (with characteristics frequencies uniformly spaced on an ERB-number scale between 125 Hz and 14000 Hz in each case). To keep the network's input size fixed at 100-by-1000 (necessary to use the same network architecture), we transposed the input array, again using 100 timesteps instead of 1000 (sampled at 2 kHz), and upsampling the frequency (nerve fiber) dimension to 1000 via linear interpolation. In this way the input dimensionality was preserved across conditions even though the information was limited by the desired number of nerve fibers. Median %F0 error on the validation set and discrimination thresholds and were measured for networks trained and tested with each of these peripheral representations (Supplemental Fig. 4).

Manipulating cochlear filter bandwidths

Cochlear filter bandwidths in the auditory nerve model were set based on estimates of human frequency tuning from otoacoustic and behavioral experiments (Shera et al., 2002). We modified the frequency tuning to be two times narrower and two times broader than these human estimates by scaling the filter bandwidths by 0.5 and 2.0, respectively.

To investigate the importance of logarithmic filtering in the cochlea, we also generated a peripheral representation with linearly spaced cochlear filters. The characteristic frequencies of 100 model nerve fibers were linearly spaced between 125 Hz and 8125 Hz and the 10-dB-down bandwidth of each cochlear filter was set to 80 Hz. This bandwidth (which is approximately equal to that of a human model fiber with 400 Hz characteristic frequency) was chosen to be as narrow as possible without introducing frequency “gaps” between adjacent cochlear filters.

To verify that these manipulations had the anticipated effects, we measured frequency tuning curves of simulated nerve fibers with characteristic frequencies of 250, 500, 1000, 2000, 4000 Hz (Fig. 5A&B). Mean firing rate responses were computed for each fiber to pure tones with frequencies between 125 Hz and 8000 Hz presented at 50 dB SPL.

Sound statistics manipulations

The general procedure for investigating the dependence of network behavior on sound statistics was to (1) modify the sounds in training dataset, (2) retrain networks (starting from a random initialization) on auditory nerve representations of the modified training dataset, and (3) simulate psychophysical experiments on trained networks, always using the same test stimuli.

Training on filtered natural sounds

We generated lowpass and highpass versions of our natural sounds training dataset by applying randomly-generated lowpass or highpass Butterworth filters to every speech and instrument sound excerpt. For the lowpass-filtered dataset, 3-dB-down filter cutoff frequencies were drawn uniformly on a logarithmic scale between 500 and 5000 Hz.. For the highpass-filtered dataset, cutoff frequencies were drawn uniformly from a logarithmic scale between 1000 and 10000 Hz. The order of each filter was drawn uniformly from 1 to 5 and all filters were applied twice, once forward and once backwards, to eliminate phase shifts. Filtered speech and instrument sounds were then combined with the unmodified background noise signals used in the original dataset (SNRs drawn uniformly from -10 to +10 dB).

Training on spectrally matched and anti-matched synthetic tones

To investigate the extent to which network pitch behavior could be explained by low-level spectral statistics of our natural sounds dataset, we generated a dataset of 2.1 million synthetic stimuli with spectral statistics matched to those measured from our primary dataset. STRAIGHT (Kawahara et al., 2008) was used to measure the spectral envelope (by averaging the estimated filter spectrogram across time) of every speech and instrument sound in our dataset. We then measured the mean and covariance of the first 13 Mel-frequency cepstral coefficients (MFCCs), defining a multivariate Gaussian. We sampled new spectral envelopes from this distribution by drawing MFCC coefficients and inverting them to produce a spectral envelope. These enveloped were imposed (via multiplication in the frequency domain) on harmonic complex tones with F0s sampled to uniformly populate the 700 log-spaced F0 bins in the network's classification layer. Before envelope imposition, tones initially contained all harmonics up to 16 kHz in cosine phase, with equal amplitudes.

To generate a synthetic dataset with spectral statistics that deviate considerably from those measured from our primary dataset, we simply multiplied the mean of the fitted multivariate Gaussian (a vector of 13 MFCCs) by negative one, which inverts the mean spectral envelope. Spectral envelopes sampled from the distribution defined by the negated mean (and unaltered covariance matrix) were imposed on 2.1 million harmonic complex tones to generate an "anti-matched" synthetic tones dataset.

Both the matched and anti-matched synthetic tones were embedded in synthetic noise spectrally matched to the background noise in our primary natural sounds dataset. The procedure for synthesizing spectrally-matched noise was analogous to the one used to generate spectrally-matched tones, except that we estimated the envelope with the power spectrum. We measured the power spectrum of every background noise clip in our primary dataset, computed the mean and covariance of the first 13 MFCCs, and imposed spectral envelopes sampled from the resulting multivariate Gaussian on white noise via multiplication in the frequency domain. Synthetic tones and noise were combined with SNRs drawn uniformly from -10 to +10 dB and overall stimulus presentation levels were drawn uniformly from 30 to 90 dB SPL.

Training on speech and music separately

We generated speech-only and music-only training datasets by selectively sampling from the same source libraries used to populate the combined dataset. Due to the lack of speech clips in our source libraries with high F0s, we decided to limit both datasets to F0s between 80 and 450 Hz (spanning 480 of 700 F0 bins). This ensured that differences between networks trained on speech or music would not be due to differences in the F0 range. The composition of the speech-only dataset was:

- F0 bins between 80 Hz and 320 Hz
    - 100% adult speech (2000 SWC and 1000 WSJ clips per bin)
- F0 bins between 320 Hz and 450 Hz
    - 100% child speech (1500 CSLU and 1500 CMU clips per bin)

The composition of our music-only dataset was:

- F0 bins between 80 Hz and 450 Hz
    - 100% instrumental music (2000 NSynth and 1000 RWC clips per bin)

Stimuli in both datasets were added to texture-like background noise clips sampled from the same sources used for the combined dataset (SNRs drawn uniformly from -10 to +10 dB).

Training on natural sounds with reduced background noise

To train networks in a low-noise environment, we regenerated our natural sounds training dataset with SNRs drawn uniformly from +10 to +30 dB rather than -10 to +10 dB. For the noiseless case, we entirely omitted the addition of background noise to the speech and instruments sounds before training. To ensure F0 discrimination thresholds measured from networks trained with reduced background noise would not be limited by masking noise in the psychophysical stimuli, we evaluated these networks on noiseless versions of the psychophysical stimuli (Experiments A and E). The amplitudes of harmonics components typically masked by noise in these stimuli (i.e., harmonics inaudible to human listeners in the original studies) were set to zero in the noiseless stimuli. When these networks were evaluated on psychophysical stimuli that did include masking noise, F0 discrimination behavior was qualitatively similar, but absolute thresholds were elevated relative to networks that were trained on the -10 to +10 dB SNR dataset.

Network neurophysiology

We simulated electrophysiological recordings and functional imaging experiments on our trained networks by examining the internal activations of networks in response to stimuli. We treated units in the convolutional layers as model "neurons" and looked at their overall responses by averaging ReLU activations across time. We measured the network’s tuning properties using the harmonics tones from Experiment A (in sine-phase – a total of 46080 stimuli: 30 lowest harmonic numbers × 1536 F0s spaced logarithmically between 80 and 320 Hz). For each unit, we constructed a two-dimensional tuning matrix with shape $[384, 30]$ (examples are shown in Fig. 8A). Rows correspond to the 384 F0 classification bins spanning 80 to 320 Hz. Columns correspond to lowest harmonic numbers 1 through 30. Each entry of the tuning matrix contained the mean response to 4 stimuli with the same lowest harmonic number and slightly different F0s (within 1/16 semitones). We constructed F0 tuning curves for each unit by taking the mean of the tuning matrix across the columns. Tuning curves of each unit were normalized by dividing by the maximum response.

To measure the average sharpness of F0 tuning in a given layer (Fig. 8B), we randomly selected 3168 units in the layer and took the mean of their F0 tuning curves after aligning them by their best F0s. This subsampling procedure ensured that apparent differences in F0 tuning sharpness between layers are not simply due to differences in the number of units per layer.

To measure population responses as a function of lowest harmonic number (Fig. 8C), we first identified the best F0 (i.e., the F0 producing the largest normalized mean response across all lowest harmonic numbers) for each of the 4608 units in the fifth convolutional layer. We then constructed lowest harmonic number tuning curves by taking responses to stimuli at the best F0 with lowest harmonic numbers 1 to 30. These tuning curves were averaged across units to give the population response.

We qualitatively compared the network's population response to those of pitch-selective neurons in marmoset auditory cortex (Bendor and Wang, 2005) and pitch-selective voxels in human auditory cortex (Norman-Haignere et al., 2013). Bendor and Wang used single-unit electrophysiology to measure the spiking rates of 50 pitch-selective neurons (from 3 marmosets) in response to complexes containing 3 to 9 consecutive harmonics in either cosine or Schroeder phase. Recordings were repeated a minimum of 10 times per stimulus condition. Norman-Haignere and colleagues measured changes in the BOLD signal with fMRI in response to bandpass-filtered sine-phase harmonic complex tones. The 12 participants (4 male, 8 female, ages 21-28) were non-musicians with normal hearing. Recordings were repeated 7-8 times per stimulus condition. Pitch-selective voxels were defined as those whose responses were larger for complex tones than for frequency-matched noise. We re-plotted data extracted from figures in both published studies using Engauge Digitizer (Mitchell et al.).

Statistics

Analyses of human-model comparison metrics

Human-model comparison metrics were computed separately for each psychophysical experiment (as described in the "Network Psychophysics" section) and for each of the 400 networks trained in our architecture search. To test if networks with better performance on the F0-estimation training task produce better matches to human psychophysical behavior, we computed the Pearson correlation between validation set accuracies and human-model comparison metrics for each experiment (Fig. 3, right-most column).

To test if "deep" networks (defined here as networks with more than one convolutional layer) tended to produce better performance on the training task than the networks with just one convolutional layer (Supplemental Fig. 2), we performed a Wilcoxon rank-sum test comparing the validation set accuracies of the 54 single-convolutional-layer networks to those of the other 346 networks. To test if the "deep" networks tended to produce better matches to human behavior we performed a Wilcoxon rank-sum test on the human-model similarity metrics. To obtain a single human-model similarity score per network for this test, we pooled metrics across the five main psychophysical experiments. This was accomplished by first rank-ordering the human-model similarity metrics of all networks within experiments and then averaging ranks across experiments.

To assess the statistical significance of changes in these human-model similarity metrics when networks were optimized for altered cochleae or sound statistics, we compared metrics measured from 10 networks trained per condition. The networks had 10 different architectures, corresponding to the 10 best-performing architectures identified in our search (Supplemental Table 1). We performed two-sample t-tests (each sample containing results from each of 10 networks) to compare human-model comparison metrics between training conditions. Because human-model similarity metrics were bounded between -1 and 1, we passed the metrics through an inverse normal cumulative distribution function before performing the t-tests.

Analyses of variance

We performed analyses of variance (ANOVAs) on log-transformed F0 discrimination thresholds to help satisfy the assumptions of equal variance and normality (normality was evaluated by eye). Mixed model ANOVAs were performed with training conditions (peripheral model and training set manipulations) as between-subject factors and psychophysical stimulus parameters (lowest harmonic number, stimulus presentation level, and signal-to-noise ratio) as within-subject factors. The specific pairings of these different factors were: stimulus presentation level vs. auditory nerve phase-locking cutoff (Fig. 4E), lowest harmonic number vs. training set spectral statistics (Fig. 6C), and lowest harmonic number vs. training set noise level (Fig. 7). We also performed a repeated-measures ANOVA to test for a main effect of lowest harmonic number on population responses in the network's last convolutional (Fig. 8C). F-statistics, p-values, and $\eta^2_{partial}$ are reported for main effects and interactions of interest. Greenhouse-Geisser corrections were applied in all cases where Mauchly's test indicated the assumption of sphericity had been violated.

Analyses of F0 discrimination behavior

One of the key signatures of human pitch perception is that listeners are very good at making fine F0 discriminations (thresholds typically below 1%) if and only if stimuli contain low-numbered harmonics. F0 discrimination thresholds increase by an order of magnitude for stimuli containing only higher-numbered harmonics (Houtsma and Smurzynski, 1990; Bernstein and Oxenham, 2003). To assess the effect of altered cochlear input or training sound statistics, we thus focused on two measures: first, the absolute F0 discrimination acuity of our model when all low-numbered harmonics were present, and second, the harmonic number at which discrimination ability transitioned from good to poor. In each case we used two-sample t-tests, comparing either the F0 discrimination thresholds (log-transformed) for tones containing the first harmonic, or the lowest harmonic number where thresholds first exceeded 1%. In each case we compared results for networks with different auditory nerve models or training sets.

ACKNOWLEDGMENTS

We thank Alex Durango and Jenelle Feather for providing the AudioSet background noise stimuli, Jenelle Feather and Andrew Francl for contributions to a shared codebase used in this project, John Cohn for assistance with computing resources, Josh Bernstein, Andrew Oxenham, Trevor Shackleton, and Bob Carlyon for sharing human psychophysics data and stimuli, Sam Norman-Haignere and the McDermott lab for helpful feedback on the manuscript, and Ian Bruce, Laurel Carney, Bertrand Delgutte, Oded Barzelay, and Hideki Kawahara for helpful discussions. Work supported by NSF grant BCS-1634050 and NIH grant R01DC017970.

REFERENCES

Ahmad, N., Higgins, I., Walker, K.M.M., and Stringer, S.M. (2016). Harmonic training and the formation of pitch representation in a neural network model of the auditory brain. Front. Comput. Neurosci. *10*.

Allen, E.J., and Oxenham, A.J. (2014). Symmetric interactions and interference between pitch and timbre. J. Acoust. Soc. Am. *135*, 1371–1379.

Arehart, K.H., and Burns, E.M. (1999). A comparison of monotic and dichotic complex-tone pitch perception in listeners with hearing loss. J. Acoust. Soc. Am. *106*, 993–997.

Attneave, F., and Olson, R.K. (1971). Pitch as a medium: a new approach to psychophysical scaling. Am. J. Psychol. *84*, 147–166.

Barzelay, O., Furst, M., and Barak, O. (2017). A new approach to model pitch perception using sparse coding. PLOS Comput. Biol. *13*, e1005338.

Bendor, D., and Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. Nature *436*, 1161–1165.

Bernstein, J.G., and Oxenham, A.J. (2003). Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number? J. Acoust. Soc. Am. *113*, 3323–3334.

Bernstein, J.G.W., and Oxenham, A.J. (2005). An autocorrelation model with place dependence to account for the effect of harmonic number on fundamental frequency discrimination. J. Acoust. Soc. Am. *117*, 3816–3831.

Bernstein, J.G.W., and Oxenham, A.J. (2006). The relationship between frequency selectivity and pitch discrimination: Sensorineural hearing loss. J. Acoust. Soc. Am. *120*, 3929–3945.

Besson, M., Schon, D., Moreno, S., Santos, A., and Magne, C. (2007). Influence of musical expertise and musical training on pitch processing in music and language. Restor. Neurol. Neurosci. *25*, 13.

Bizley, J.K., and Cohen, Y.E. (2013). The what, where and how of auditory-object perception. Nat. Rev. Neurosci. *14*, 693–707.

Bizley, J.K., Walker, K.M.M., Nodal, F.R., King, A.J., and Schnupp, J.W.H. (2013). Auditory cortex represents both pitch judgments and the corresponding acoustic cues. Curr. Biol. *23*, 620–625.

Borchert, E.M.O., Micheyl, C., and Oxenham, A.J. (2011). Perceptual grouping affects pitch judgments across time and frequency. J. Exp. Psychol. Hum. Percept. Perform. *37*, 257–269.

Bruce, I.C., Erfani, Y., and Zilany, M.S.A. (2018). A phenomenological model of the synapse between the inner hair cell and auditory nerve: implications of limited neurotransmitter release sites. Hear. Res. *360*, 40–54.

Burge, J., and Geisler, W.S. (2011). Optimal defocus estimation in individual natural images. Proc. Natl. Acad. Sci. *108*, 16849–16854.

Cariani, P. (1999). Temporal coding of periodicity pitch in the auditory system: an overview (Hindawi).

Cariani, P.A., and Delgutte, B. (1996). Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. J. Neurophysiol. *76*, 1698–1716.

Carlyon, R.P., and Shackleton, T.M. (1994). Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms? J. Acoust. Soc. Am. *95*, 3541–3554.

Carney, L.H. (2018). Supra-threshold hearing and fluctuation profiles: implications for sensorineural and hidden hearing loss. J. Assoc. Res. Otolaryngol. *19*, 331–352.

Cedolin, L., and Delgutte, B. (2005). Pitch of complex tones: rate-place and interspike interval representations in the auditory nerve. J. Neurophysiol. *94*, 347–362.

de Cheveigné, A. (2005). Pitch Perception Models. In Pitch: Neural Coding and Perception, C.J. Plack, R.R. Fay, A.J. Oxenham, and A.N. Popper, eds. (New York, NY: Springer), pp. 169–233.

de Cheveigné, A., and Pressnitzer, D. (2006). The case of the missing delay lines: Synthetic delays obtained by cross-channel phase interaction. J. Acoust. Soc. Am. *119*, 3908–3918.

Chi, T., Ru, P., and Shamma, S.A. (2005). Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. *118*, 887–906.

Colburn, H.S. (1973). Theory of binaural interaction based on auditory-nerve data. I. General strategy and preliminary results on interaural discrimination. J. Acoust. Soc. Am. *54*, 1458–1470.

Cutler, A., Dahan, D., and van Donselaar, W. (1997). Prosody in the comprehension of spoken language: a literature review. Lang. Speech *40*, 141–201.

Darwin, C.J., Brungart, D.S., and Simpson, B.D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. J. Acoust. Soc. Am. *114*, 2913–2922.

Dau, T., Kollmeier, B., and Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. J. Acoust. Soc. Am. *102*, 2906–2919.

Dowling, W.J., and Fujitani, D.S. (1971). Contour, interval, and pitch recognition in memory for melodies. J. Acoust. Soc. Am. *49*, 524–531.

Durlach, N.I., and Braida, L.D. (1969). Intensity perception. I. Preliminary theory of intensity resolution. J. Acoust. Soc. Am. *46*, 372–383.

Eickenberg, M., Gramfort, A., Varoquaux, G., and Thirion, B. (2017). Seeing it all: convolutional network layers map the function of the human visual system. NeuroImage *152*, 184–194.

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Eck, D., Simonyan, K., and Norouzi, M. (2017). Neural audio synthesis of musical notes with WaveNet autoencoders. ArXiv170401279 Cs.

Eskenazi, M., Mostow, J., and Graff, D. (1997). The CMU kids speech corpus. Linguist. Data Consort.

Feather, J., Durango, A., Gonzalez, R., and McDermott, J. (2019). Metamers of neural networks reveal divergence from human perceptual systems. In Advances in Neural Information Processing Systems, pp. 10078–10089.

Geisler, W.S. (2011). Contributions of ideal observer theory to vision research. Vision Res. *51*, 771–781.

Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., and Ritter, M. (2017). Audio Set: an ontology and human-labeled dataset for audio events. In Proc. IEEE ICASSP 2017, (New Orleans, LA), pp. 776–780.

Gfeller, K., Turner, C., Oleson, J., Zhang, X., Gantz, B., Froman, R., and Olszewski, C. (2007). Accuracy of cochlear implant recipients on pitch perception, melody recognition, and speech reception in noise: Ear Hear. *28*, 412–423.

Girshick, A.R., Landy, M.S., and Simoncelli, E.P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. Nat. Neurosci. *14*, 926–932.

Glasberg, B.R., and Moore, B.C.J. (1990). Derivation of auditory filter shapes from notched-noise data. Hear. Res. *47*, 103–138.

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

Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. (2003). RWC music database: music genre database and musical instrument sound database.

Grothe, B., Pecka, M., and McAlpine, D. (2010). Mechanisms of sound localization in mammals. Physiol. Rev. *90*, 983–1012.

Guclu, U., and van Gerven, M.A.J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. *35*, 10005–10014.

Han, K., and Wang, D. (2014). Neural network based pitch tracking in very noisy speech. IEEEACM Trans. Audio Speech Lang. Process. *22*, 2158–2168.

Heinz, M.G., Colburn, H.S., and Carney, L.H. (2001a). Evaluating auditory performance limits: I. One-parameter discrimination using a computational model for the auditory nerve. Neural Comput. *13*, 2273–2316.

Heinz, M.G., Colburn, H.S., and Carney, L.H. (2001b). Evaluating auditory performance limits: II. One-parameter discrimination with random-level variation. Neural Comput. *13*, 2317–2338.

Heinz, M.G., Zhang, X., Bruce, I.C., and Carney, L.H. (2001c). Auditory nerve model for predicting performance limits of normal and impaired listeners. Acoust. Res. Lett. Online *2*, 91–96.

Hénaff, O.J., and Simoncelli, E.P. (2016). Geodesics of learned representations. ArXiv151106394 Cs.

Hildebrand, J.G., and Shepherd, G.M. (1997). Mechanisms of olfactory discrimination: converging evidence for common principles across phyla. Annu. Rev. Neurosci. *20*, 595–631.

Hoekstra, A. (1979). Frequency discrimination and frequency analysis in hearing. [S.n.].

Houtsma, A.J.M., and Smurzynski, J. (1990). Pitch identification and discrimination for complex tones with many harmonics. J. Acoust. Soc. Am. *87*, 304–310.

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. ArXiv150203167 Cs.

Jacoby, N., Undurraga, E.A., McPherson, M.J., Valdés, J., Ossandón, T., and McDermott, J.H. (2019). Universal and non-universal features of musical pitch perception revealed by singing. Curr. Biol. *29*, 3229-3243.e12.

Javel, E., and Mott, J.B. (1988). Physiological and psychophysical correlates of temporal processes in hearing. Hear. Res. *34*, 275–294.

Joris, P.X., Smith, P.H., and Yin, T.C. (1998). Coincidence detection in the auditory system: 50 years after Jeffress. Neuron *21*, 1235–1238.

Joris, P.X., Bergevin, C., Kalluri, R., Laughlin, M.M., Michelet, P., Heijden, M. van der, and Shera, C.A. (2011). Frequency selectivity in Old-World monkeys corroborates sharp cochlear tuning in humans. Proc. Natl. Acad. Sci. *108*, 17516–17520.

Jozwik, K.M., Kriegeskorte, N., Storrs, K.R., and Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. Front. Psychol. *8*.

Kaernbach, C., and Demany, L. (1998). Psychophysical evidence against the autocorrelation theory of auditory temporal processing. J. Acoust. Soc. Am. *104*, 2298–2306.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. 2008 IEEE Int. Conf. Acoust. Speech Signal Process. 3933–3936.

Kell, A.J., and McDermott, J.H. (2019a). Deep neural network models of sensory systems: windows onto the role of task constraints. Curr. Opin. Neurobiol. *55*, 121–132.

Kell, A.J.E., and McDermott, J.H. (2019b). Invariance to background noise as a signature of non-primary auditory cortex. Nat. Commun. *10*, 3958.

Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S.V., and McDermott, J.H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron *98*, 630-644.e16.

Kim, J.W., Salamon, J., Li, P., and Bello, J.P. (2018). CREPE: a convolutional representation for pitch estimation. ArXiv180206182 Cs Eess Stat.

Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. ArXiv14126980 Cs.

Kishon-Rabin, L., Amir, O., Vexler, Y., and Zaltz, Y. (2001). Pitch discrimination: are professional musicians better than non-musicians? J. Basic Clin. Physiol. Pharmacol. *12*.

Köhn, A., Stegen, F., and Baumann, T. (2016). Mining the Spoken Wikipedia for speech data and beyond. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), (Portorož, Slovenia: European Language Resources Association (ELRA)), pp. 4644–4647.

Lake, B.M., Ullman, T.D., Tenenbaum, J.B., and Gershman, S.J. (2017). Building machines that learn and think like people. Behav. Brain Sci. *40*.

Latinus, M., and Belin, P. (2011). Human voice perception. Curr. Biol. *21*, R143–R145.

Laudanski, J., Zheng, Y., and Brette, R. (2014). A structural theory of pitch. ENeuro *1*.

Lavan, N., Knight, S., and McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. Nat. Commun. *10*, 2404.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.

Lee, B.S., and Ellis, D.P.W. (2012). Noise robust pitch tracking by subband autocorrelation classification. 4.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. J. Acoust. Soc. Am. *49*, 467–477.

Lewicki, M.S. (2002). Efficient coding of natural sounds. Nat. Neurosci. *5*, 356–363.

Liberman, M.C. (1991). Central projections of auditory-nerve fibers of differing spontaneous rate. I. Anteroventral cochlear nucleus. J. Comp. Neurol. *313*, 240–258.

Licklider, J.C.R. (1951). A duplex theory of pitch perception. Experientia *7*, 128–134.

Lindsay, G. (2020). Convolutional neural networks as a model of the visual system: past, present, and future. J. Cogn. Neurosci.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B.C.J. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. Proc. Natl. Acad. Sci. *103*, 18866–18869.

McDermott, J.H., and Simoncelli, E.P. (2011). Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron *71*, 926–940.

McDermott, J.H., Lehr, A.J., and Oxenham, A.J. (2008). Is relative pitch specific to pitch? Psychol. Sci. *19*, 1263–1271.

McDermott, J.H., Keebler, M.V., Micheyl, C., and Oxenham, A.J. (2010a). Musical intervals and relative pitch: frequency resolution, not interval resolution, is special. J. Acoust. Soc. Am. *128*, 1943–1951.

McDermott, J.H., Lehr, A.J., and Oxenham, A.J. (2010b). Individual differences reveal the basis of consonance. Curr. Biol. *20*, 1035–1041.

McPherson, M.J., and McDermott, J.H. (2018). Diversity in pitch perception revealed by task dependence. Nat. Hum. Behav. *2*, 52–66.

McPherson, M.J., and McDermott, J.H. (2020). Efficient codes for memory determine pitch representations (Neuroscience).

McWalter, R., and McDermott, J.H. (2018). Adaptive and selective time averaging of auditory scenes. Curr. Biol. *28*, 1405-1418.e10.

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

Meddis, R., and O’Mard, L. (1997). A unitary model of pitch perception. J. Acoust. Soc. Am. *102*, 1811–1820.

Mehrer, J., Spoerer, C.J., Kriegeskorte, N., and Kietzmann, T.C. (2020). Individual differences among deep neural network models. BioRxiv 2020.01.08.898288.

Mehta, A.H., and Oxenham, A.J. (2020). Effect of lowest harmonic rank on fundamental-frequency difference limens varies with fundamental frequency. J. Acoust. Soc. Am. *147*, 2314–2322.

Micheyl, C., Delhommeau, K., Perrot, X., and Oxenham, A.J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. Hear. Res. *219*, 36–47.

Micheyl, C., Schrater, P.R., and Oxenham, A.J. (2013). Auditory frequency and intensity discrimination explained using a cortical population rate code. PLOS Comput. Biol. *9*, e1003336.

Mitchell, M., Muftakhidinov, B., and Winchen, T. Engauge Digitizer Software.

Moore, B.C.J. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. J. Assoc. Res. Otolaryngol. *9*, 399–406.

Moore, G.A., and Moore, B.C.J. (2003). Perception of the low pitch of frequency-shifted complexes. J. Acoust. Soc. Am. *113*, 977–985.

Moore, B.C.J., Glasberg, B.R., and Peters, R.W. (1985). Relative dominance of individual partials in determining the pitch of complex tones. J. Acoust. Soc. Am. *77*, 1853–1860.

Moore, B.C.J., Glasberg, B.R., and Proctor, G.M. (1992). Accuracy of pitch matching for pure tones and for complex tones with overlapping or nonoverlapping harmonics. J. Acoust. Soc. Am. *91*, 3443–3450.

Moore, B.C.J., Glasberg, B.R., Flanagan, H.J., and Adams, J. (2006). Frequency discrimination of complex tones; assessing the role of component resolvability and temporal fine structure. J. Acoust. Soc. Am. *119*, 480–490.

Norman-Haignere, S., Kanwisher, N., and McDermott, J.H. (2013). Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. J. Neurosci. *33*, 19451–19469.

Osmanski, M.S., Song, X., and Wang, X. (2013). The role of harmonic resolvability in pitch perception in a vocal nonhuman primate, the common marmoset (Callithrix jacchus). J. Neurosci. *33*, 9161–9168.

Oxenham, A.J. (2013). Revisiting place and temporal theories of pitch. Acoust. Sci. Technol. Ed. Acoust. Soc. Jpn. *34*, 388–396.

Oxenham, A.J., Bernstein, J.G.W., and Penagos, H. (2004). Correct tonotopic representation is necessary for complex pitch perception. Proc. Natl. Acad. Sci. *101*, 1421–1425.

Oxenham, A.J., Micheyl, C., and Keebler, M.V. (2009). Can temporal fine structure represent the fundamental frequency of unresolved harmonics? J. Acoust. Soc. Am. *125*, 2189–2199.

Oxenham, A.J., Micheyl, C., Keebler, M.V., Loper, A., and Santurette, S. (2011). Pitch perception beyond the traditional existence region of pitch. Proc. Natl. Acad. Sci. *108*, 7629–7634.

Palmer, A.R., and Russell, I.J. (1986). Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. Hear. Res. *24*, 1–15.

Patterson, R.D., Uppenkamp, S., Johnsrude, I.S., and Griffiths, T.D. (2002). The processing of temporal pitch and melody information in auditory cortex. Neuron *36*, 767–776.

Paul, D.B., and Baker, J.M. (1992). The design for the Wall Street Journal-based CSR corpus. In Proceedings of the Workshop on Speech and Natural Language, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 357–362.

Plack, C.J., and Oxenham, A.J. (2005). The Psychophysics of Pitch. In Pitch: Neural Coding and Perception, C.J. Plack, R.R. Fay, A.J. Oxenham, and A.N. Popper, eds. (New York, NY: Springer), pp. 7–55.

Popham, S., Boebinger, D., Ellis, D.P.W., Kawahara, H., and McDermott, J.H. (2018). Inharmonic speech reveals the role of harmonicity in the cocktail party problem. Nat. Commun. *9*, 2122.

Pressnitzer, D., Suied, C., and Shamma, S. (2011). Auditory scene analysis: the sweet music of ambiguity. Front. Hum. Neurosci. *5*.

Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. Nat. Neurosci. *22*, 1761–1770.

Rose, J.E., Brugge, J.F., Anderson, D.J., and Hind, J.E. (1967). Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. J. Neurophysiol. *30*, 769–793.

Schouten, J.F., Ritsma, R.J., and Cardozo, B.L. (1962). Pitch of the residue. J. Acoust. Soc. Am. *34*, 1418–1424.

Semal, C., and Demany, L. (1990). The upper limit of musical pitch. Music Percept. *8*, 165–175.

Shackleton, T.M., and Carlyon, R.P. (1994). The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. J. Acoust. Soc. Am. *95*, 3529–3540.

Shamma, S., and Klein, D. (2000). The case of the missing pitch templates: How harmonic templates emerge in the early auditory system. J. Acoust. Soc. Am. *107*, 2631–2644.

Shattuck-Hufnagel, S., and Turk, A.E. (1996). A prosody tutorial for investigators of auditory sentence processing. J. Psycholinguist. Res. *25*, 193–247.

Shera, C.A., Guinan, J.J., and Oxenham, A.J. (2002). Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. Proc. Natl. Acad. Sci. *99*, 3318–3323.

Shobaki, K., Hosom, J.-P., and Cole, R. (2007). CSLU: Kids' speech version 1.1. Linguist. Data Consort.

Shofner, W.P., and Chaney, M. (2013). Processing pitch in a nonhuman mammal (Chinchilla laniger). J. Comp. Psychol. *127*, 142–153.

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

Slaney, M., and Lyon, R.F. (1990). A perceptual pitch detector. In International Conference on Acoustics, Speech, and Signal Processing, pp. 357–360 vol.1.

Smith, Z.M., Delgutte, B., and Oxenham, A.J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. Nature *416*, 87–90.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. *15*, 1929–1958.

Tang, C., Hamilton, L.S., and Chang, E.F. (2017). Intonational speech prosody encoding in the human auditory cortex. Science *357*, 797–801.

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

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

Verschooten, E., Shamma, S., Oxenham, A.J., Moore, B.C.J., Joris, P.X., Heinz, M.G., and Plack, C.J. (2019). The upper frequency limit for the use of phase locking to code temporal fine structure in humans: A compilation of viewpoints. Hear. Res. *377*, 109–121.

Walker, K.M., Gonzalez, R., Kang, J.Z., McDermott, J.H., and King, A.J. (2019). Across-species differences in pitch perception are consistent with differences in cochlear filtering. ELife *8*, e41626.

Wandell, B.A. (1995). Foundations of Vision (Sinauer Associates).

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., and Tanaka, K. (2018). Illusory motion reproduced by deep neural networks trained for prediction. Front. Psychol. *9*.

Weiss, Y., Simoncelli, E.P., and Adelson, E.H. (2002). Motion illusions as optimal percepts. Nat. Neurosci. *5*, 598–604.

White, L.J., and Plack, C.J. (1998). Temporal processing of the pitch of complex tones. J. Acoust. Soc. Am. *103*, 2051–2063.

Wier, C.C., Jesteadt, W., and Green, D.M. (1977). Frequency discrimination as a function of frequency and sensation level. J. Acoust. Soc. Am. *61*, 178–184.

Wightman, F.L. (1973). The pattern-transformation model of pitch. J. Acoust. Soc. Am. *54*, 407–416.

Woods, K.J.P., and McDermott, J.H. (2015). Attentive tracking of sound sources. Curr. Biol. *25*, 2238–2246.

Yamins, D.L.K., and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. *19*, 356–365.

Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. *111*, 8619–8624.

Yin, R.K. (1969). Looking at upside-down faces. J. Exp. Psychol. *81*, 141–145.

Zilany, M.S.A., and Bruce, I.C. (2006). Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. J. Acoust. Soc. Am. *120*, 1446–1466.

Zilany, M.S.A., Bruce, I.C., Nelson, P.C., and Carney, L.H. (2009). A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics. J. Acoust. Soc. Am. *126*, 2390–2412.

Zilany, M.S.A., Bruce, I.C., and Carney, L.H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. J. Acoust. Soc. Am. *135*, 283–286.