Task-optimized models of human hearing link perception and neural coding

by

Mark R. Saddler

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2024

Authored by: Mark R. Saddler
Department of Brain and Cognitive Sciences
October 4, 2023

Certified by: Josh H. McDermott
Associate Professor, Thesis Supervisor

Accepted by: Mark T. Harnett
Graduate Officer, Department of Brain and Cognitive Sciences

Task-optimized models of human hearing link perception and neural coding

by

Mark R. Saddler

Submitted to the Department of Brain and Cognitive Sciences
on October 4, 2023 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

ABSTRACT

Hearing allows organisms to derive information about the world from sound. The ears convert pressure waves into patterns of neural activity which the brain can use to make powerful inferences about the source. While extensive modeling efforts in the past few decades have resulted in well-established computational descriptions of peripheral auditory coding, comparatively less is known about how this neural code supports complex auditory behavior. Humans with normal hearing are remarkably adept at recognizing and localizing sounds in noisy environments with multiple competing sources. However, these abilities are fragile and are greatly compromised in listeners with hearing loss or cochlear implants, often leading to frustration and social isolation. Current assistive devices largely fail to aid impaired listeners in noisy environments, and the development of more effective devices is currently limited by an incomplete understanding of which features of neural coding underlie perception. This thesis develops computational models for explicitly linking specific features of peripheral auditory processing and perception. In a series of three studies, we optimized deep artificial neural network models to perform real-world hearing tasks using simulated auditory nerve input. The first study outlined a framework for optimizing models under different conditions to test how perception is constrained by our ears and acoustic environment in the domain of pitch perception. The second study extended the framework to examine the widely debated perceptual role of auditory nerve spike timing in hearing more broadly. The third study explored a practical application of task-optimized models by leveraging intermediate model representations as a perceptual metric for speech enhancement. Collectively, the results link aspects of hearing to environmental and neural coding constraints, illustrating the utility of artificial networks to reveal underpinnings of behavior.

Thesis supervisor: Josh H. McDermott
Title: Associate Professor

Acknowledgments

First and foremost, I thank Josh McDermott. I am incredibly grateful for his limitless patience, availability, and dedication to mentoring. It's been truly fantastic to go through grad school with an advisor as consistent and invested in my success as Josh.

I next thank my committee whose guidance has been invaluable to this thesis: Bertrand Delgutte for his encouragement and expertise, Jim DiCarlo for his rigor and motivation, Michale Fee for his insight and clarity. I am fortunate to have benefited so directly from their leadership in their departments and in their fields. I am also sincerely grateful to my department, particularly the wonderful staffs of McGovern and BCS Headquarters and all those who tirelessly keep OpenMind afloat.

I thank each member of the McDermott Lab for their kindness, generosity, and advice. It has been a joy to work alongside them these last six years and I could not imagine a better environment in which to pursue a PhD. I particularly thank Jarrod Hicks for his friendship throughout the entire journey as well as Andrew Francl and Jenelle Feather for their mentorship. I thank my collaborators over the years: Ray Gonzalez, Bryan Medina, Bryan Garcia, Annesya Banerjee, Malinda McPherson, Ian Griffith, and Annika Magaro. Working with each of you was a highlight of my PhD. To those I did not directly work with, I thank you for the many lab meetings, lunches, conversations, commiserations, and laughs that nonetheless contributed to this thesis.

I am immensely grateful to those who introduced me to research and inspired me to pursue this PhD: Laela Sayigh, Wim van Drongelen, Elizabeth Kovar, Simone Baumann-Pickering, Laura Hale, and Shrek Chalasani.

Lastly, I thank my friends and family. The communities of MIT's Baker House and Cambridge's pick-up soccer scene for making me feel welcome. My siblings Hannah, Neil, and Vince for their distractions and for keeping me humble. My partner Cassie for her welcome distractions and support. My mom and dad for their love and unwavering support from the very beginning.

Contents

Title page **1**

Abstract **3**

Acknowledgments **5**

List of Figures **13**

List of Tables **15**

1 Introduction **17**

1.1 Motivation 17

1.2 Organization of thesis 18

2 Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception **20**

2.1 Introduction 21

2.2 Results 23

2.2.1 Training task and stimuli 23

2.2.2 Peripheral auditory model 23

2.2.3 Neural network architecture search 23

2.2.4 Characteristics of pitch perception emerge in optimized DNNs 25

2.2.5 Dependence on low-numbered harmonics 25

2.2.6 Phase effects are limited to high-numbered harmonics 27

2.2.7 Pitch shifts for shifted low-numbered harmonics 28

2.2.8 Poor discrimination of transposed tones 28

2.2.9 DNNs with better F0 estimation show more human-like behavior . . 28

2.2.10 Human-like behavior requires a biologically-constrained cochlea . . . 30

2.2.11 Dependence of pitch behavior on the cochlea 30

2.2.12 Human-like behavior depends critically on phase locking 32
2.2.13 Human-like behavior depends less on cochlear filter bandwidths . . . 35
2.2.14 Dependence of pitch behavior on training set sound statistics 38
2.2.15 Altered training set spectra produce altered behavior 38
2.2.16 Natural spectral statistics account for human-like behavior 40
2.2.17 Music-trained networks exhibit better pitch acuity 40
2.2.18 Training set noise required for "missing fundamental" illusion 41
2.2.19 Network neurophysiology 41
2.3 Discussion 46
2.4 Methods 51
2.4.1 Natural sounds training dataset - overview 51
2.4.2 Speech and music training excerpts 51
2.4.3 Background noise for training data 52
2.4.4 Peripheral auditory model 52
2.4.5 Deep neural network models - overview 53
2.4.6 Definitions of constituent neural network operations 54
2.4.7 Model optimization - architecture search 57
2.4.8 Model optimization - network training 58
2.4.9 Network psychophysics - overview 59
2.4.10 Experiment A: effect of harmonic number and phase on pitch discrimination 59
2.4.11 Experiment B: pitch of alternating-phase harmonic complexes 61
2.4.12 Experiment C: pitch of frequency-shifted complexes 63
2.4.13 Experiment D: pitch of complexes with individually mistuned harmonics 64
2.4.14 Experiment E: frequency discrimination with pure and transposed tones 65
2.4.15 Effect of stimulus level on frequency discrimination 67
2.4.16 Auditory nerve manipulations - overview 68
2.4.17 Replacing the hardwired cochlear model with learnable filters 68
2.4.18 Manipulating fine timing information in the auditory nerve 69
2.4.19 Eliminating place cues by flattening the excitation pattern 70
2.4.20 Manipulating cochlear filter bandwidths 70
2.4.21 Sound statistics manipulations - overview 71
2.4.22 Training on filtered natural sounds 71
2.4.23 Training on spectrally matched and anti-matched synthetic tones . . 71
2.4.24 Training on speech and music separately 72
2.4.25 Training on natural sounds with reduced background noise 73

2.4.26 Network neurophysiology . . . 73
2.4.27 Statistics - analysis of human data . . . 75
2.4.28 Statistics - analysis of human-model comparison metrics . . . 75
2.4.29 Statistics - analysis of best thresholds and transition points . . . 76
2.4.30 Statistics - ANOVAs on discrimination thresholds . . . 76
2.5 Supplementary information . . . 78
2.5.1 Data availability . . . 78
2.5.2 Code availability . . . 78
2.5.3 Acknowledgments . . . 78
2.5.4 Author contributions . . . 78
2.5.5 Supplementary figures . . . 79

3 The role of temporal coding in hearing: evidence from task-optimization 92
3.1 Introduction . . . 93
3.2 Results . . . 95
3.2.1 Auditory nerve model stage . . . 95
3.2.2 Temporal coding manipulation . . . 95
3.2.3 Simplified cochlear model stage . . . 97
3.2.4 Artificial neural network model stages and training . . . 97
3.2.5 Model tasks and roadmap . . . 98
3.2.6 Model optimization – sound localization . . . 99
3.2.7 Degraded temporal coding impairs sound localization . . . 99
3.2.8 Sound localization psychoacoustics . . . 101
3.2.9 Auditory nerve phase locking is critical for ITD-based sound localization 101
3.2.10 Azimuth dependence of human localization requires phase locking . . . 102
3.2.11 Human ITD sensitivity is not limited only by auditory nerve phase locking . . . 103
3.2.12 Models without access to phase locking exhibit inhuman spectral cue dependence . . . 105
3.2.13 Not all localization phenomena are inextricably linked to phase-locked spike timing . . . 107
3.2.14 Quantitative measures of human-model similarity – sound localization 107
3.2.15 Model optimization – pitch perception . . . 107
3.2.16 Human-like pitch perception depends critically on auditory nerve phase locking . . . 109
3.2.17 Quantitative measures of human-model similarity – pitch perception . 110

3.2.18 Model optimization – voice and word recognition . . . 110
3.2.19 Phase locking to the fundamental improves voice recognition in noise 110
3.2.20 The dependence of human voice recognition on absolute pitch requires phase locking . . . 112
3.2.21 Phase locking offers little benefit for word recognition in real-world noise 113
3.2.22 Phase locking is needed to account for vocoding effects in human word recognition . . . 115
3.2.23 Quantitative measures of human-model similarity – voice and word recognition . . . 116
3.2.24 Replication with simplified cochlear model . . . 116
3.3 Discussion . . . 117
3.3.1 Relation to prior experimental work . . . 118
3.3.2 Relation to prior modeling work . . . 118
3.3.3 Limitations . . . 119
3.3.4 Future directions . . . 120
3.4 Methods . . . 122
3.4.1 Peripheral auditory model . . . 122
3.4.2 Phase locking manipulation . . . 123
3.4.3 Simplified cochlear model . . . 123
3.4.4 Artificial neural network architectures . . . 124
3.4.5 Localization network architecture modification . . . 124
3.4.6 Transposed pitch network architectures . . . 125
3.4.7 Model optimization – overview . . . 126
3.4.8 Model optimization – sound localization . . . 126
3.4.9 Model optimization – pitch perception . . . 127
3.4.10 Model optimization – voice and word recognition . . . 128
3.4.11 Model evaluation – overview . . . 129
3.4.12 Localization model evaluation – sound localization in noise . . . 129
3.4.13 Localization model evaluation – psychoacoustics . . . 130
3.4.14 Minimum audible angle vs. azimuth . . . 130
3.4.15 ITD / ILD cue weighting . . . 131
3.4.16 ITD lateralization vs. frequency . . . 132
3.4.17 Effect of changing ears . . . 133
3.4.18 Effect of smoothing spectral cues . . . 134
3.4.19 Median plane spectral cues . . . 135
3.4.20 Precedence effect . . . 135

3.4.21 Bandwidth dependency of localization . . . 136
3.4.22 Pitch model evaluation – pitch discrimination in noise . . . 137
3.4.23 Pitch model evaluation – psychoacoustics . . . 138
3.4.24 Effect of harmonic number and phase on pitch discrimination . . . 138
3.4.25 Pitch of alternating-phase harmonic complexes . . . 140
3.4.26 Pitch of frequency-shifted complexes . . . 141
3.4.27 Pitch of complexes with individually mistuned harmonics . . . 142
3.4.28 Pitch discrimination with pure and transposed tones . . . 143
3.4.29 Effect of level on pure tone frequency discrimination . . . 145
3.4.30 Speech model evaluation – voice and word recognition in noise . . . 146
3.4.31 Voice discrimination in noise . . . 147
3.4.32 Voice and word recognition with pitch-altered speech . . . 148
3.4.33 Word recognition with tone-vocoded speech . . . 149
3.4.34 Statistics . . . 151
3.5 Supplementary Information . . . 152

4 Speech denoising with auditory models 164
4.1 Introduction . . . 165
4.2 Methods . . . 165
4.2.1 Recognition networks . . . 166
4.2.2 Audio transforms . . . 167
4.2.3 Evaluation . . . 168
4.3 Results . . . 170
4.3.1 Deep feature losses yield improved denoising . . . 170
4.3.2 Learned vs. random deep features . . . 170
4.3.3 Comparison to previous deep feature systems . . . 171
4.3.4 Effect of task used to train deep features . . . 171
4.3.5 Cochlear model losses match deep feature losses . . . 171
4.3.6 Effect of filter bank characteristics . . . 172
4.3.7 Objective metrics . . . 173
4.4 Discussion . . . 173
4.5 Acknowledgements . . . 174

5 Conclusion 175
5.1 Contributions . . . 175
5.2 Future directions . . . 177

References **179**

List of Figures

2.1 Pitch model overview. 24
2.2 Pitch model validation: human and neural network psychophysics. 26
2.3 Network architectures producing better F0 estimation for natural sounds exhibit more human-like pitch behavior. 29
2.4 Networks trained to estimate F0 directly from sound waveforms exhibit less human-like pitch behavior. 31
2.5 Pitch perception is impaired in networks optimized with degraded spike timing in the auditory nerve. 33
2.6 Cochlear frequency tuning has relatively little effect on pitch perception. . . 36
2.7 Pitch perception depends on training set sound statistics. 39
2.8 Key characteristics of human pitch behavior only emerge in noisy training conditions. 42
2.9 Network neurophysiology. 43
2.10 Pitch behavior of the 10 best network architectures ranked by F0 estimation performance on natural sounds. 80
2.11 Network architectures producing better F0 estimation for natural sounds exhibit more human-like pitch behavior. 81
2.12 Deep networks better account for human psychophysical behavior than networks with just one convolutional layer. 82
2.13 Effect of number of auditory nerve fibers on F0 estimation error and discrimination thresholds measured from networks without spike-timing information. 83
2.14 Effects of phase locking cutoff on network pitch behavior. 84
2.15 Networks do not rely on place cues to F0 in the excitation pattern to produce human-like pitch behavior. 85
2.16 Effects of altered cochlear frequency selectivity on network pitch behavior. . 86
2.17 Effects of training set sound statistics on network pitch behavior. 87

2.18 F0 discrimination thresholds as a function of lowest harmonic number, measured from networks trained separately on speech-only and music-only datasets. 88
2.19 Effects of auditory nerve fiber type (high vs. low spontaneous rate) on network pitch behavior. 89
2.20 Effects of F0 classification bin width on F0 estimation error and discrimination thresholds. 90

3.1 Overview. 96
3.2 Sound localization is impaired in models with degraded auditory nerve spike timing. 100
3.3 Upper frequency limit of ITD sensitivity. 103
3.4 Localization models with degraded phase locking rely on spectral cues to judge azimuth as well as elevation. 106
3.5 Models with degraded auditory nerve spike-timing exhibit less human-like pitch perception. 108
3.6 Models with degraded auditory nerve spike-timing exhibit less human-like voice recognition. 111
3.7 Models with degraded temporal coding achieve human-level word recognition in noise but fail to account for psychoacoustic phenomena. 114
3.8 Effect of phase locking manipulation on all localization experiments. 153
3.9 Effect of phase locking manipulation on all pitch experiments. 154
3.10 Models optimized separately for voice and word recognition – effect of phase locking manipulation on all speech experiments. 155
3.11 Simplified cochlear model – effect of phase locking manipulation on all localization experiments. 156
3.12 Simplified cochlear model – effect of phase locking manipulation on all pitch experiments. 157
3.13 Simplified cochlear model – effect of phase locking manipulation on all speech experiments. 158
3.14 Simplified cochlear model – human-model similarity scores. 159
3.15 Model voice and word recognition with inharmonic speech in noise. 160

4.1 Schematic of audio transform training. 166
4.2 Rated naturalness vs. SNR for speech processed by Wave-U-Nets trained on deep feature losses. 170
4.3 Rated naturalness vs. SNR for speech processed by Wave-U-Nets trained on cochlear model losses. 172

List of Tables

2.1 Details of 10 best network architectures. 91

3.1 Neural network architectures for sound localization models. 161
3.2 Neural network architectures for pitch models. 162
3.3 Neural network architectures for voice and word recognition models. 163

4.1 Experiment 1 results. 169
4.2 Experiment 2 results. 171

Chapter 1

Introduction

1.1 Motivation

Sound pressure waves in the environment wiggle our ear drums back and forth. The cochlea transduces these mechanical vibrations into electrical signals, which are transmitted to the brainstem via the auditory nerve. The signal processing of these peripheral auditory stages has been extensively characterized [1], [2]. Decades of neurophysiological and computational research has resulted in quantitative models of the ear that faithfully capture the spiking responses of auditory nerve fibers to arbitrary sound stimuli [3]–[5]. Comparatively less is understood about how patterns of auditory nerve spikes give rise to auditory perception.

Humans with normal hearing can recognize and localize sources, make sense of musical melodies, and hold conservations in noisy environments with multiple competing sources. These remarkable perceptual abilities are fragile and are greatly compromised in listeners with hearing loss or cochlear implants, often leading to frustration and social isolation [6]. Sensorineural hearing loss – an umbrella term for impairments in the transduction of sound energy to electrical signals or in their transmission to brain – is a ubiquitous and growing public health issue, with 1 in every 10 people projected to have disabling hearing loss by 2050 [7]. Current assistive devices restore some aspects of hearing but largely fail to provide benefit in the noisy environments where listeners are most impaired. One factor limiting the development of more effective devices is our incomplete understanding of how peripheral auditory coding relates to perception [8]. Though much is known separately about the physiological and perceptual consequences of sensorineural hearing loss, quantitatively linking the two has long posed a challenge.

Classic computational approaches from signal detection and ideal observer theory have made headway in this direction; however, they have been limited to very constrained perceptual tasks such as simple frequency and level discrimination [9], [10]. These methods involve

deriving statistically optimal solutions to perceptual tasks given the information available in a neural representation. Deriving task solutions for different neural representations and comparing the behavior of resulting models to humans can provide insight into which aspects of neural coding are important for human perception [11]. While these classic methods have been fruitful, they quickly become intractable for more complicated, real-world perceptual tasks. It is not feasible to derive statistically optimal solutions to many of the tasks that hearing-impaired listeners struggle with, such as speech recognition in noise.

The emergence of deep learning models presents an opportunity to resurrect these classic approaches for real-world perceptual problems. Deep artificial neural networks are highly expressive mathematical functions that can be optimized to perform complex tasks. Networks optimized to classify natural images and sounds have emerged as leading candidate models of sensory systems in vision [12] and hearing [13]. They provide state-of-the-art matches to human behavior [13]–[18] and predictions of neural activity [13], [19]–[21]. This thesis demonstrates how networks can also be used to uncover the environmental and neural coding factors that determine perception. In a series of studies, we optimized deep artificial neural networks to perform real-world hearing tasks using input from a detailed model of the auditory nerve. The resulting models replicated many characteristics of human hearing behavior. To analyze why, we separately optimized models under different constraints. We investigated links between perception and neural coding by optimizing models with altered cochlear input.

1.2 Organization of thesis

The first study [17] focused on pitch perception, which is among the best-characterized aspects of human hearing. Despite a wealth of available behavioral data, the underlying computations and constraints that determine pitch perception have long been debated [22], [23]. Particular controversy persists over the relative importance of auditory nerve spike timing and cochlear frequency selectivity [24]–[28]. 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 [29]. We investigated the issue with artificial neural networks optimized to estimate fundamental frequency (F0), the perceptual correlate of pitch. Our networks replicated many characteristics of human pitch perception, but only when optimized for naturalistic sounds and cochleae with high temporal fidelity. The results suggest pitch perception is critically shaped by the constraints of natural environments in addition to those of the cochlea.

The second study extends the approach of the first to investigate the role of temporal

coding in hearing more broadly. It is well established that spike timing is exquisitely precise in the auditory nerve, where action potentials phase lock to the individual pressure oscillations of sound waveforms up to 3 to 5 kHz [30]–[32]. The perceptual role of this precise temporal coding remains controversial because physiological mechanisms for extracting information from it have yet to be discovered despite significant efforts to identify them [27], [33]. Nonetheless, the information encoded in relatively high frequency phase locking is widely suspected to be critical for sound localization and is often proposed to underlie hearing in noise [34]–[36]. We investigated the issue with artificial neural networks optimized for sound localization, pitch estimation, and speech perception in noise. We separately optimized models with four different auditory nerve phase locking limits and compared them to humans across 21 behavioral experiments. Models with access to precise auditory nerve spike timing comprehensively accounted for human behavior. Models with degraded temporal coding could not, providing new evidence for the role of phase locking in hearing. The results link neural coding to real-world perception and clarify conditions in which prostheses that fail to restore high-fidelity temporal coding (e.g., contemporary cochlear implants) could in principle restore near-normal hearing.

The third study [37] explores a practical application of task-optimized hearing models by leveraging intermediate model representations as a perceptual metric for speech enhancement. Contemporary speech enhancement predominantly relies on audio transforms optimized for waveform-based loss functions [38]–[45]. We investigated whether the learned intermediate representations from task-optimized auditory models might serve as a more perceptually aligned loss function. We trained a speech enhancement system to reconstruct clean speech waveforms from noisy speech waveforms with different loss functions. The baseline system minimized distances between model outputs and clean speech targets in waveform space. Our proposed systems minimized distances between outputs and targets in representational space from different auditory models [46]. We found these auditory model-based loss functions yielded better speech denoising than waveform-based losses, but most of the benefit could be attributed to the model's cochlear representation rather than the task-optimized features. The results suggest learned deep network features do not yet improve upon simple cochlear filter bank features for speech denoising.

Chapter 2

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

Mark R. Saddler, Ray Gonzalez, Josh H. McDermott

Abstract

Perception is thought to be shaped by the environments for which organisms are optimized. These influences are difficult to test in biological organisms but may be revealed by machine perceptual systems optimized under different conditions. We investigated environmental and physiological influences on pitch perception, whose properties are commonly linked to peripheral neural coding limits. We first trained artificial neural networks to estimate fundamental frequency from biologically faithful cochlear representations of natural sounds. The best-performing networks replicated many characteristics of human pitch judgments. To probe the origins of these characteristics, we then 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 naturalistic 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 artificial neural networks to reveal underpinnings of behavior.

2.1 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 [47]. Olfactory discrimination is similarly constrained by the receptor classes in the nose [48]. 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 [49].

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 [32], but the role of this information in hearing remains controversial [50]–[52]. 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 [11]. 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 [10], [53]–[55]. 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 [56]. While DNNs are unlikely to fully achieve optimal performance, they might reveal the effects of optimizing a system under particular constraints [18], [57]. Previous work has documented similarities between human and network behavior for neural networks trained on vision or hearing tasks [13], [14], [19]. 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) [22]. Pitch is believed to enable a wide range of auditory-driven behaviors, such as voice and melody recognition [58], and has been the subject of a long history of work in psychology [59]–[63] and neuroscience [24], [26], [64], [65]. Yet despite a wealth of data, the underlying computations and constraints that determine pitch perception remain debated [22]. In particular, controversy persists over the role of spike timing in the auditory nerve, for which a physiological extraction mechanism has remained elusive [27], [33]. The role of cochlear frequency selectivity, which has also been proposed to constrain pitch discrimination, remains similarly debated [24], [66]. 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 resolution of these debates is that previous models of pitch have generally not attained quantitatively accurate matches to human behavior [63], [67]–[74]. Moreover, because most previous models have been mechanistic rather than normative, they do not speak to the potential adaptation of pitch perception to particular types of sounds or peripheral neural codes. Here we used DNNs in the role traditionally occupied by ideal observers, optimizing them to extract pitch information from peripheral neural representations of natural sounds. DNNs have become the method of choice for pitch tracking in engineering applications [75], 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.

2.2 Results

2.2.1 Training task and stimuli

We used supervised deep learning to build a model of pitch perception optimized for natural speech and music. DNNs were trained to estimate the F0 of short (50ms) 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.

2.2.2 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 [5] to simulate peripheral representations of each stimulus (Fig. 3.1A). The input representations to our networks 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. 3.1B. 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 DNN models have access to all of this information. Through optimization for the training task, the DNNs should learn to use whichever peripheral cues best allow them to extract F0.

2.2.3 Neural network 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

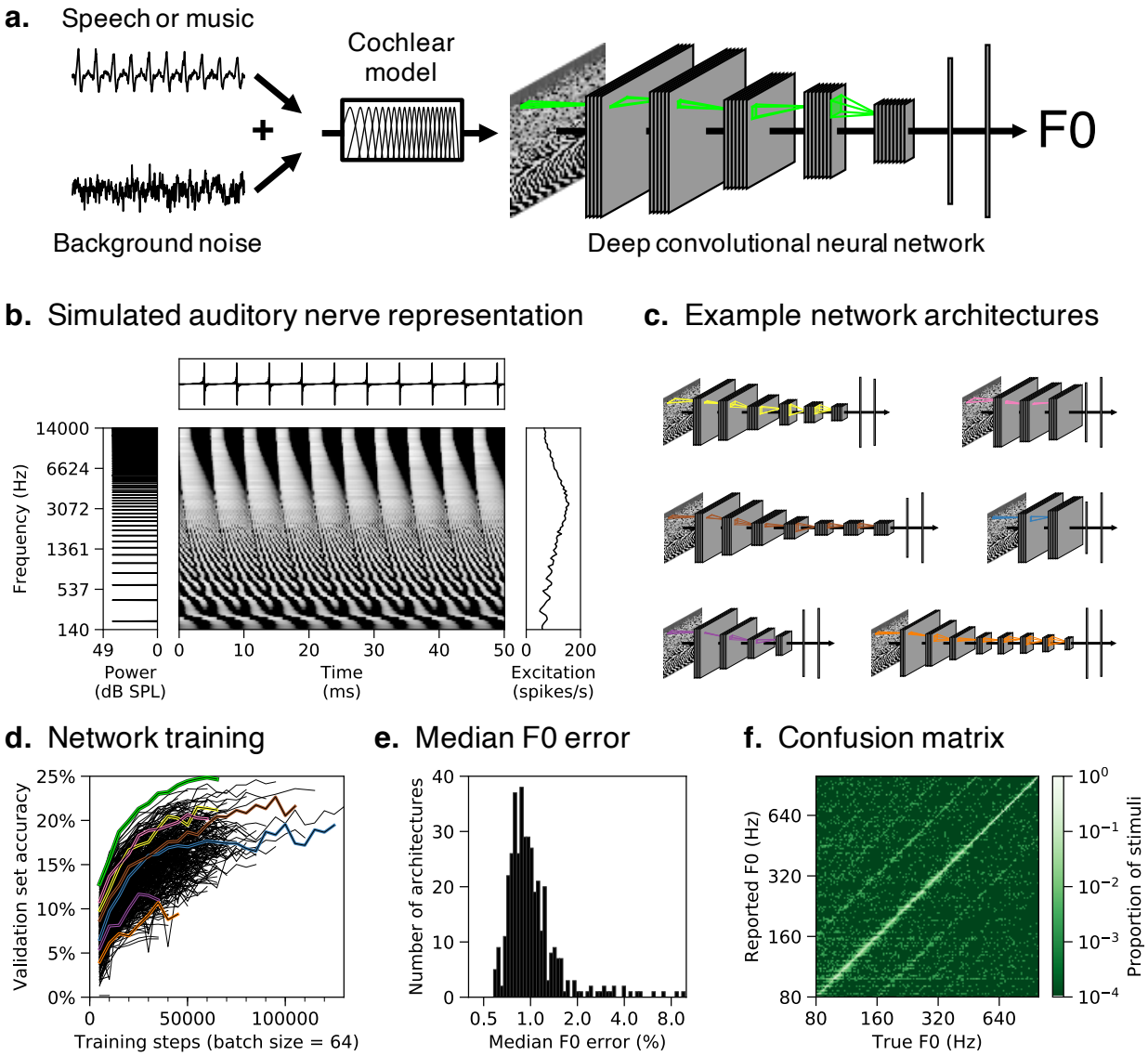

Figure 2.1: Pitch model overview.

(A) Schematic of model structure. DNNs 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 DNN 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 relatively 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 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.

of the network [19]. 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. 3.1C).

The different architectures produced a broad distribution of training task performances (Fig. 3.1D). In absolute terms accuracy was good – the median error was well below 1% (Fig. 3.1E), which is on par with good human F0 discrimination thresholds [63], [76]. The vast majority of misclassifications fell within bins neighboring the true F0 or at an integer number octaves away (Fig. 3.1F), as in human pitch-matching judgments [77].

2.2.4 Characteristics of pitch perception emerge in optimized DNNs

Having obtained a model that can estimate F0 from natural sounds, we simulated a suite of well-known psychophysical experiments to assess whether the model replicated known properties of human pitch perception. Each experiment measures the effect of particular cues on pitch discrimination or estimation using synthetic tones (Fig. 3.2, left column), and produces an established result in human listeners (Fig. 3.2, center column). We tested the effect of these stimulus manipulations on our 10 best-performing network architectures. Given evidence for individual differences across different networks optimized for the same task [78], most figures feature results averaged across the 10 best networks identified in our architecture search (which we collectively refer to as "the model"). Averaging across an ensemble of networks effectively allows us to marginalize over architectural hyperparameters and provide uncertainty estimates for our model's results [79], [80]. Individual results for the 10 networks are shown in Supplementary Fig. 3.8.

As shown in Fig. 3.2, the model (right column) qualitatively and in most cases quantitatively replicates the result of each of the five different experiments in humans (center column). 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. These results collectively suggest that the model relies on similar cues as the human pitch system. We describe these results in turn.

2.2.5 Dependence on low-numbered harmonics

First, human pitch discrimination is more accurate for stimuli containing low-numbered harmonics (Fig. 3.2A, center, solid line) [60], [63], [76], [81]. This finding is often interpreted

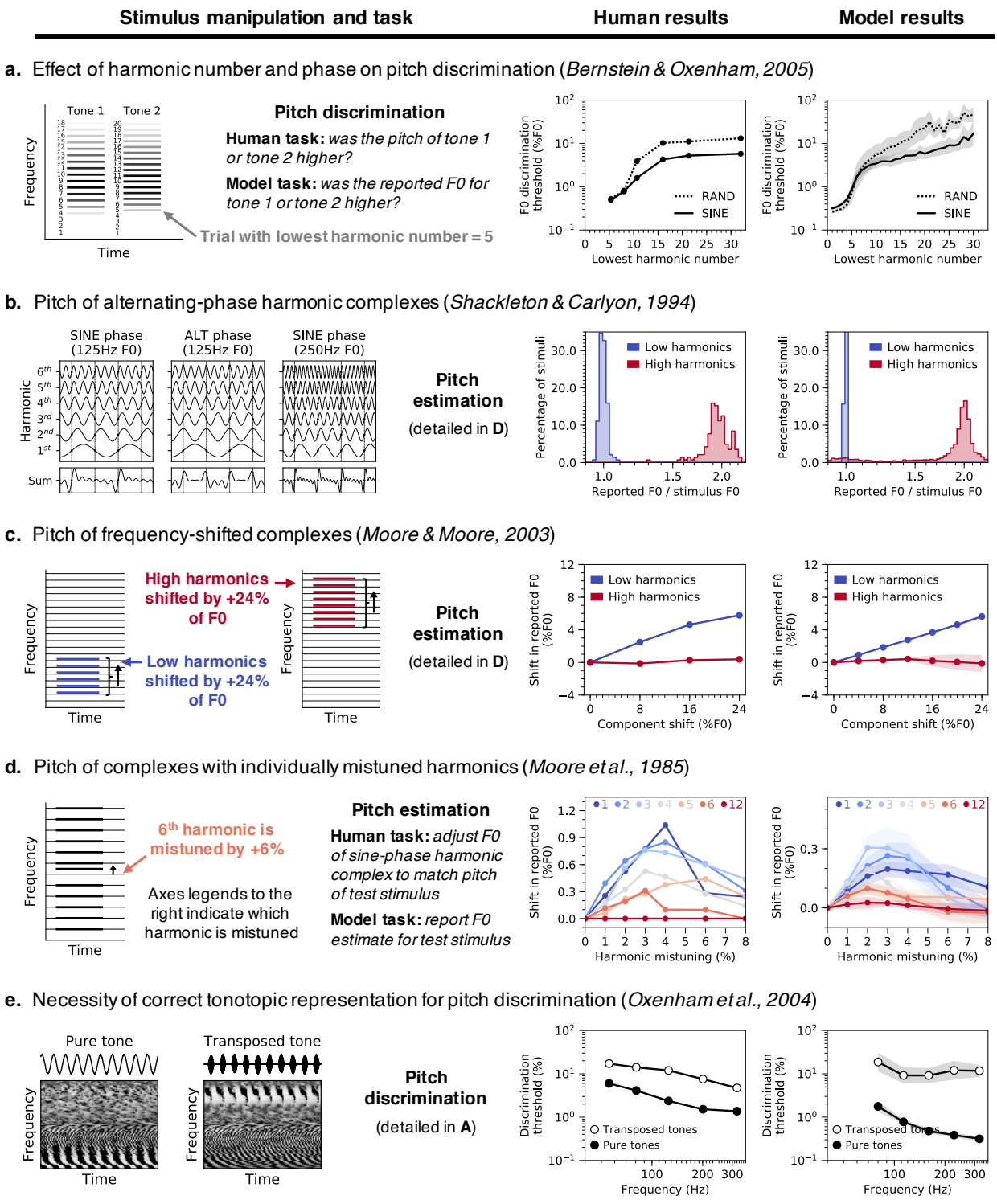

Figure 2.2: Pitch model validation: human and neural network psychophysics.

Five classic experiments from the pitch psychoacoustics literature (A-E) were simulated on neural 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 Supplementary Fig. 3.8). (A) F0 discrimination thresholds for bandpass synthetic tones, as a function of lowest harmonic number and phase. Human listeners and networks discriminated 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 was 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. This could be because the networks are better able to isolate the contribution of the harmonic to the F0, whereas human listeners may sometimes erroneously be biased by the harmonic itself. (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.

as evidence for the importance of "place" cues to pitch, which are only present for low-numbered harmonics (Fig. 3.1B, 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. 3.2A, right, solid line).

2.2.6 Phase effects are limited to high-numbered harmonics

Second, human perception is affected by harmonic phases only for high-numbered harmonics. When harmonic phases are randomized, human discrimination thresholds are elevated for stimuli that lack low-numbered harmonics (Fig. 3.2A, center, dashed vs. solid line) [63]. 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. 3.2B, left), listeners report the pitch to be twice as high as the corresponding sine-phase complex, but only for high-numbered harmonics (Fig. 3.2B, center) [60]. These results are typically thought to indicate use of temporal fluctuations in a sound's envelope when cues for low-numbered harmonics are not available [24], [60], [76]. The model replicates both effects (Fig. 3.2A&B, right), indicating that it uses similar temporal cues to pitch as humans, and in similar conditions.

2.2.7 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. 3.2C, left) produce linear shifts in the pitch reported by humans, but only if the tones contain low-numbered harmonics (Fig. 3.2C, center) [61]. The model's F0 predictions for these stimuli resemble those measured from human listeners (Fig. 3.2C, right).

Fourth, shifting individual harmonics in a complex tone ("mistuning"; Fig. 3.2D, left) can also produce pitch shifts in humans under certain conditions [59]: the mistuning must be small (effects are largest for 3-4% mistunings) and applied to a low-numbered harmonic (Fig. 3.2D, center). The model replicates this effect as well, although the size of the shift is smaller than that observed in humans (Fig. 3.2D, right).

2.2.8 Poor discrimination of transposed tones

Fifth, "transposed tones" designed to instantiate the temporal cues from low frequencies at a higher-frequency place on the cochlea (Fig. 3.2E, left) elicit weak pitch percepts in humans and thus yield higher discrimination thresholds than pure tones (Fig. 3.2E, center) [62]. This finding is taken to indicate that to the extent that temporal cues to pitch matter perceptually, 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. 3.2E, right).

2.2.9 DNNs with better F0 estimation show more human-like behavior

To evaluate whether the human-model similarity evident in Fig. 3.2 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 estimated F0 for the validation set (Fig. 3.1D&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 to each network's performance on the validation set (Fig. 3.3A-E).

For four of the five experiments (Fig. 3.3A-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. 3.3E) was the exception, as all networks similarly replicated the main human result regardless of their training task performance. We suspect

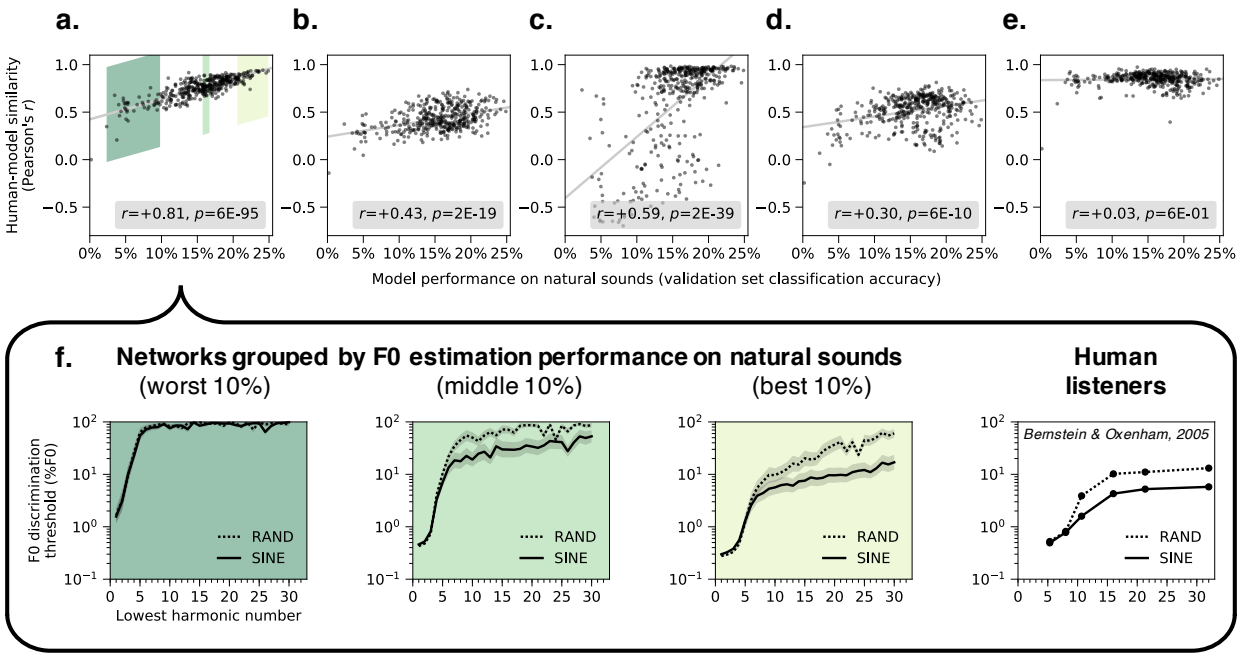

Figure 2.3: Network architectures producing better F0 estimation for natural sounds exhibit more human-like pitch behavior.

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 model 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. 3.2A-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 patches in A). Lines plot means across the 40 networks. 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.

this is because transposed tones cause patterns of peripheral stimulation that rarely occur for natural sounds. Thus, virtually any model that learns to associate naturally occurring peripheral cues with F0 will exhibit poor performance for transposed tones.

To illustrate the effect of optimization for one experiment, Fig. 3.3F displays the average F0 discrimination thresholds for each of the worst, middle, and best 10% of networks (sorted by performance on the validation set). It is visually apparent that top-performing networks exhibit more similar psychophysical behavior to humans than worse-performing networks. See Supplementary Fig. 3.9 for analogous results for the other four experiments from Fig. 3.2. 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 behavior are byproducts of optimization under natural constraints.

Because the space of network architectures is large, it is a challenge to definitively asso-

ciate particular network motifs with good performance and/or human-like behavior. However, we found that very shallow networks both performed poorly on the training task and exhibited less similarity with human behavior (Supplementary Fig. 3.10). This result provides evidence that deep networks (with multiple hierarchical stages of processing) better account for human pitch behavior than relatively shallow networks.

2.2.10 Human-like behavior requires a biologically-constrained cochlea

To test whether a biologically-constrained cochlear model was necessary for human-like pitch behavior, we trained networks to estimate F0 directly from sound waveforms (Fig. 3.4A). We replaced the cochlear model with a bank of 100 one-dimensional convolutional filters operating directly on the audio. The weights of these first-layer filters were optimized for the F0 estimation task along with the rest of the network.

The learned filters deviated from those in the ear, with best frequencies tending to be lower than those of the hardwired peripheral model (Fig. 3.4B). Networks with learned cochlear filters also exhibited less human-like behavior than their counterparts with the fixed cochlear model (Fig, 4C&D). In particular, networks with learned cochlear filters showed little ability to extract pitch information from high-numbered harmonics. Discrimination thresholds for higher harmonics were poor (Fig. 3.4C, Expt. A) and networks did not exhibit phase effects (Fig. 3.4C, Expt. A & B). Accordingly, human-model similarity was substantially lower with learned cochlear filters for two of five psychophysical experiments (Fig. 3.4D; Expt. A: $t(18) = 5.23, p < 0.001, d = 2.47$; Expt. B: $t(18) = 12.69, p < 0.001, d = 5.98$). This result suggests that a human-like cochlear representation is necessary to obtain human-like behavior, but also that the F0 estimation task on its own is insufficient to produce a human-like cochlear representation, likely because the cochlea is shaped by many auditory tasks. Thus, the cochlea may be best considered as a constraint on pitch perception rather than the other way around.

2.2.11 Dependence of pitch behavior on the cochlea

To gain insight into what aspects of the cochlea underlie the characteristics of pitch perception, we investigated how the model behavior depends on its peripheral input. 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. 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

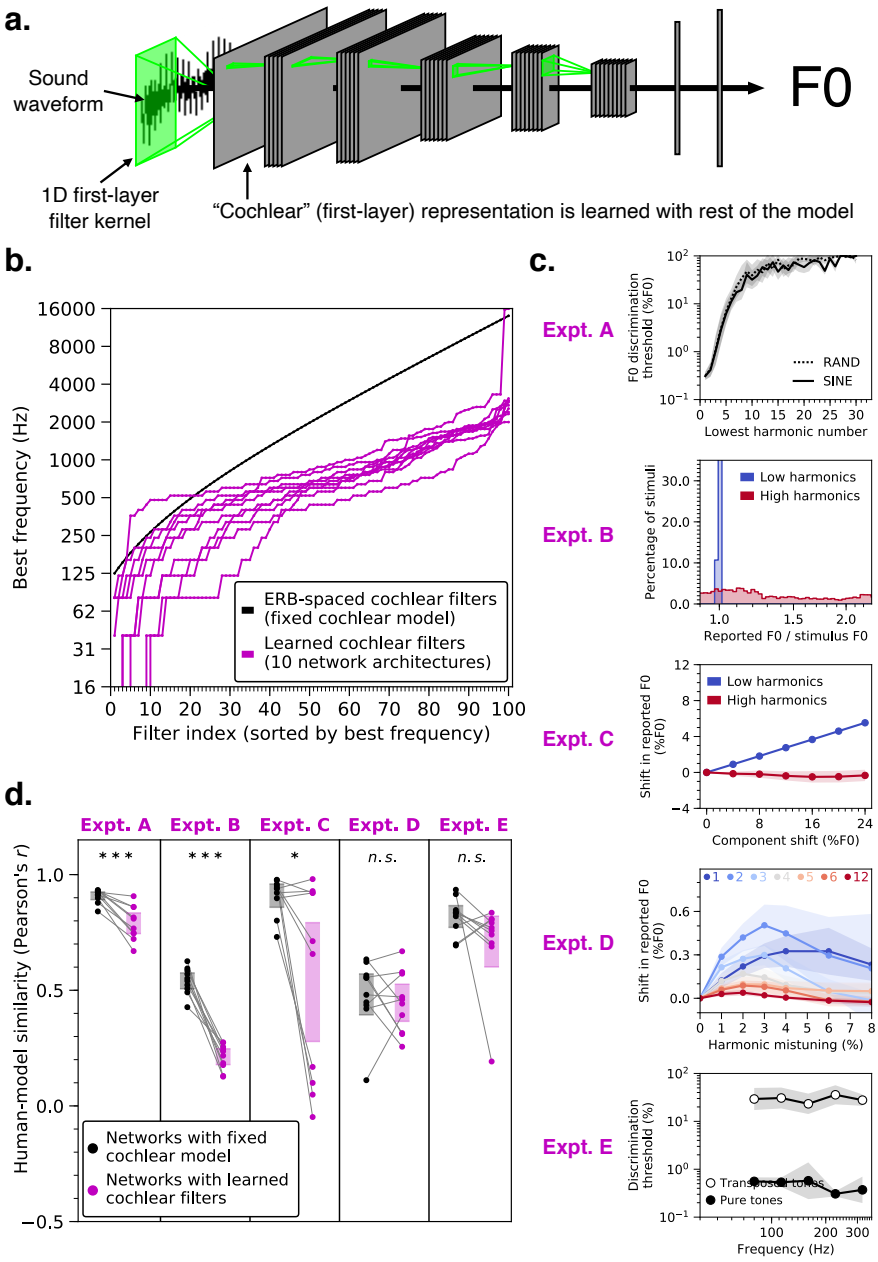

Figure 2.4: Networks trained to estimate F0 directly from sound waveforms exhibit less human-like pitch behavior.

(A) Schematic of model structure. Model architecture was identical to that depicted in Fig. 3.1A, except that the hardwired cochlear input representation was replaced by a layer of 1-dimensional convolutional filters operating directly on sound waveforms. The first-layer filter kernels were optimized for the F0 estimation task along with the rest of the network weights. We trained the 10 best networks from our architecture search with these learnable first-layer filters. (B) The best frequencies (sorted from lowest to highest) of the 100 learned filters for each of the 10 network architectures are plotted in magenta. For comparison, the best frequencies of the 100 cochlear filters in the hardwired peripheral model are plotted in black. (C) Effect of learned cochlear filters on network behavior in all five main psychophysical experiments (see Fig. 3.2A-E): F0 discrimination as a function of harmonic number and phase (Expt. A), pitch estimation of alternating-phase stimuli (Expt. B), pitch estimation of frequency-shifted complexes (Expt. C), pitch estimation of complexes with individually mistuned harmonics (Expt. D), and frequency discrimination with pure and transposed tones (Expt. E). Lines plot means across the 10 networks; error bars plot 95% confidence intervals, obtained by bootstrapping across the 10 networks. (D) Comparison of human-model similarity metrics between networks trained with either the hardwired cochlear model (black) or the learned cochlear filters (magenta) for each psychophysical experiment. Asterisks indicate statistical significance of two-sample t-tests comparing the two cochlear model conditions: $^{***}p < 0.001, ^{*}p = 0.016$. Error bars indicate 95% confidence intervals bootstrapped across the 10 network architectures.

present results for a single psychophysical result that was particularly diagnostic: the effect of lowest harmonic number on F0 discrimination thresholds (Fig. 3.2A, solid line). Results for other experiments are generally congruent with the overall conclusions and are shown in Supplementary Figures. We first present experiments manipulating the fidelity of temporal coding, followed by experiments manipulating frequency selectivity along the cochlea's length.

2.2.12 Human-like behavior depends critically on phase locking

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 [32], which are impractical to alter in vivo but which can be modified in silico via the simulated inner hair cell's lowpass filter [5]. We separately trained networks with lowpass cutoff frequencies of 50 Hz, 320 Hz, 1000 Hz, 3000 Hz (the nerve model's default value, 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 was eliminated, meaning the network only had access to cues from the place of excitation along the cochlea (Fig. 3.5A). As the cutoff frequency was increased, the network gained 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 cochleae.

Reducing the upper limit of phase locking qualitatively changed the model's psychophysical behavior and made it less human-like. As shown in Fig. 3.5B&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, $t(18) = 4.39, p < 0.001, d = 1.96$; 320 Hz, $t(18) = 11.57, p < 0.001, d = 5.17$; 50 Hz, $t(18) = 9.30, p < 0.001, d = 4.16$; 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 [9], [10]. 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, $t(18) = 5.15, p < 0.001, d = 2.30$; 50 Hz, $t(18) = 10.10, p < 0.001, d = 4.52$; two-sample t-tests comparing to the 3000 Hz condition; the transition point was on average

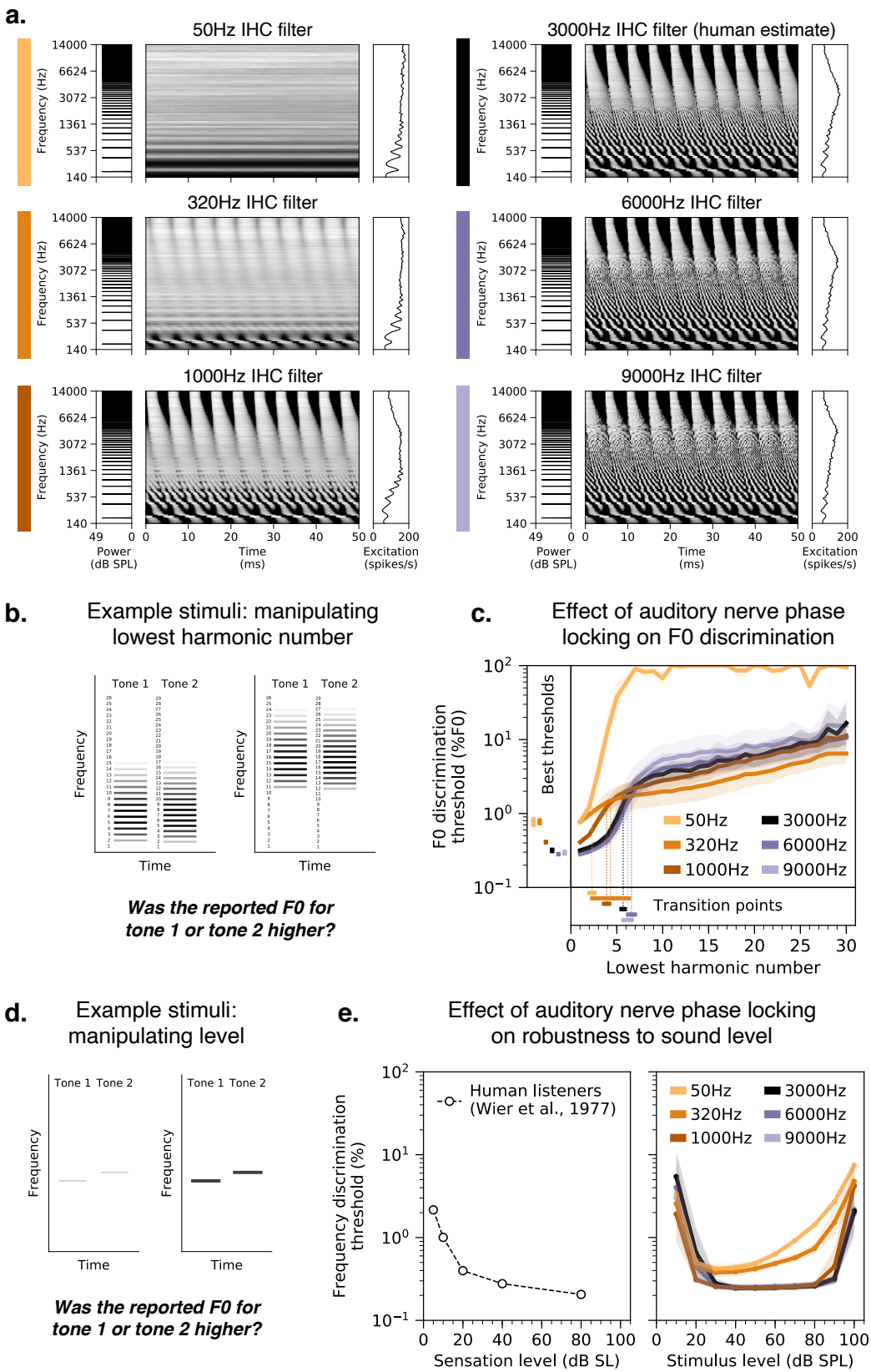

Figure 2.5: Pitch perception is impaired in networks optimized with degraded spike timing in the auditory nerve.

(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. 3.1A, 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. Here and in E, lines plot means across the 10 networks; error bars plot 95% confidence intervals, obtained by bootstrapping across the 10 networks. (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[50].

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. 3.5C), suggesting that superhuman temporal resolution would not continue to improve pitch perception (at least as assessed here). Discrimination thresholds for high-numbered harmonics were in fact slightly worse for increased cutoff frequencies. One explanation is that increasing the model's access to fine timing information biases the learned strategy to rely more on this information, which is less useful for determining the F0 of stimuli containing only high-numbered harmonics. 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 [24]. Consistent with this idea, frequency discrimination thresholds (Fig. 3.5D) measured from networks with lower phase locking cutoffs were less invariant to level than networks trained with normal spike-timing information (Fig. 3.5E, right). Thresholds for models with limited phase locking became progressively worse for louder tones, unlike those for humans (Fig. 3.5E, left) [82]. 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. 3.5C). Similar effects were observed when thresholds were measured with complex tones (data not shown).

To control for the possibility that the poor performance of the networks trained with lower phase locking cutoffs might be specific to the relatively small number of simulated auditory nerve fibers in the model, we generated an alternative representation for the 50 Hz cutoff condition, using 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. 3.5C&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 Supplementary Fig. 3.11.

We simulated the full suite of psychophysical experiments on all networks with altered cochlear temporal resolution (Supplementary Fig. 3.12). 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.

2.2.13 Human-like behavior depends less on cochlear filter bandwidths

The role of cochlear frequency tuning in pitch perception has also been the source of long-standing debates [60], [66], [76], [81], [83], [84]. 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, without retraining the model) had almost no effect on network behavior (Supplementary Fig. 3.13). 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 tuning 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 [85]. The effect of this manipulation is visually apparent in the width of nerve fiber tuning curves as well as in the number of harmonics that produce distinct peaks in the cochlear excitation patterns (Fig. 3.6A).

We also modified the cochlear model to be linearly spaced (Fig. 3.6B), uniformly distributing the characteristic frequencies of the model nerve fibers along the frequency axis and equating their filter bandwidths. Unlike a normal cochlea, which resolves only low-numbered harmonics, the linearly spaced alteration yielded a peripheral representation where all har-

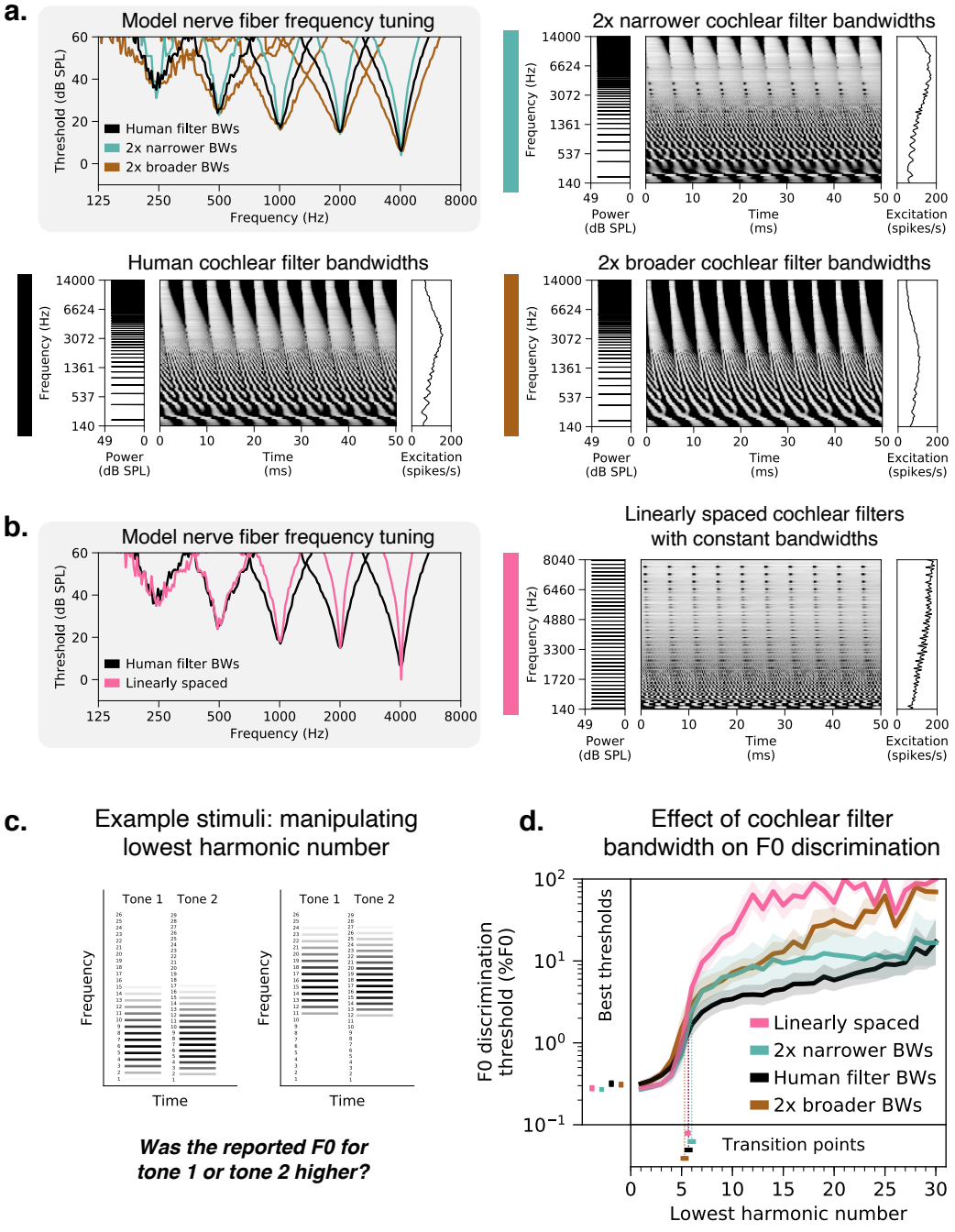

Figure 2.6: Cochlear frequency tuning has relatively little effect on pitch perception.

(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 auditory nerve tuning curves measured from five individual fibers per condition (upper left panel). Tuning curves plot thresholds for each fiber as a function of pure tone frequency. 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) Cochlear filters modeled on the human ear were replaced with a set of linearly spaced filters with constant bandwidths in Hz. Pure tone tuning curves measured with linearly spaced filters are much sharper than those estimated for humans at higher frequencies (left panel; note the log-spaced frequency scale). The right panel shows the simulated auditory nerve representation of the 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. Lines plot means across the 10 networks; error bars indicate 95% confidence intervals bootstrapped across the 10 networks.

monics 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 bandwidths exhibit relatively similar F0 discrimination behavior (Fig. 3.6C&D). Broadening filters by a factor of two had no significant effect on the best thresholds ($t(18) = 0.40, p = 0.69$, 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 ($t(18) = 2.74, p = 0.01, d = 1.23$) but very small (0.27% vs. 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 ($t(18) = 1.88, p = 0.08$). In addition, the dependence of thresholds on harmonic number was fairly similar in all cases (Fig. 3.6D). The transition between good and poor thresholds occurred around the sixth harmonic irrespective of the cochlear bandwidths (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, $t(18) = 0.37, p = 0.71$; t-tests comparing to the normal human tuning condition).

All three models with altered cochlear filter bandwidths produced worse thresholds for stimuli containing only high-numbered harmonics (Fig. 3.6D). This effect is expected for the narrower and linearly-spaced conditions (smaller bandwidths result in reduced envelope cues from beating of adjacent harmonics), but we do not have an explanation for why networks with broader filters also produced poorer thresholds. One possibility that we ruled out is overfitting of the network architectures to the human cochlear filter bandwidths; validation set accuracies were no worse with broader filters ($t(18) = 0.66, p = 0.52$). However, we note that all of the models exhibit what would be considered poor performance for stimuli containing only high harmonics (thresholds are at least an order of magnitude worse than

they are for low harmonics), and are thus all generally consistent with human perception in this regime.

We also simulated the full suite of psychophysical experiments from Fig. 3.2 on networks with altered frequency tuning. Most experimental results were robust to peripheral frequency tuning (Supplementary Fig. 3.14).

2.2.14 Dependence of pitch behavior on training set sound statistics

In contrast to the widely debated roles of peripheral cues, the role of natural sound statistics in pitch has been little discussed throughout the history of hearing research. To investigate how optimization for natural sounds 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.

2.2.15 Altered training set spectra produce altered behavior

One salient property of speech and instrument sounds is that they typically have more energy at low frequencies than high frequencies (Fig. 3.7A, 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. 3.7A, left column, orange line) and then measured their F0 discrimination thresholds (Fig. 3.7B). For comparison, we performed the same experiment with lowpass-filtered sounds.

Thresholds measured from networks optimized for highpass sounds exhibited a much weaker dependence on harmonic number than if optimized for natural sounds (Fig. 3.7C, left column). This difference produced an interaction between the effects of harmonic number and the training condition ($F(2.16, 38.85) = 72.33, p < 0.001, \eta^2_{partial} = 0.80$). 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 (Supplementary Fig. 3.15) and observed several other striking differences in their performance characteristics. In particular, networks optimized for highpass-filtered natural sounds exhibited better discrimination thresholds for transposed tones than pure tones ($t(18) = 9.92, p < 0.001, d = 4.43$, two-sided two-sample t-test comparing pure tone and transposed tone thresholds averaged across frequency), a complete reversal of the human result. These results illustrate that the properties of pitch perception are not strictly a function of the information available in the periphery – performance characteristics can depend strongly on the

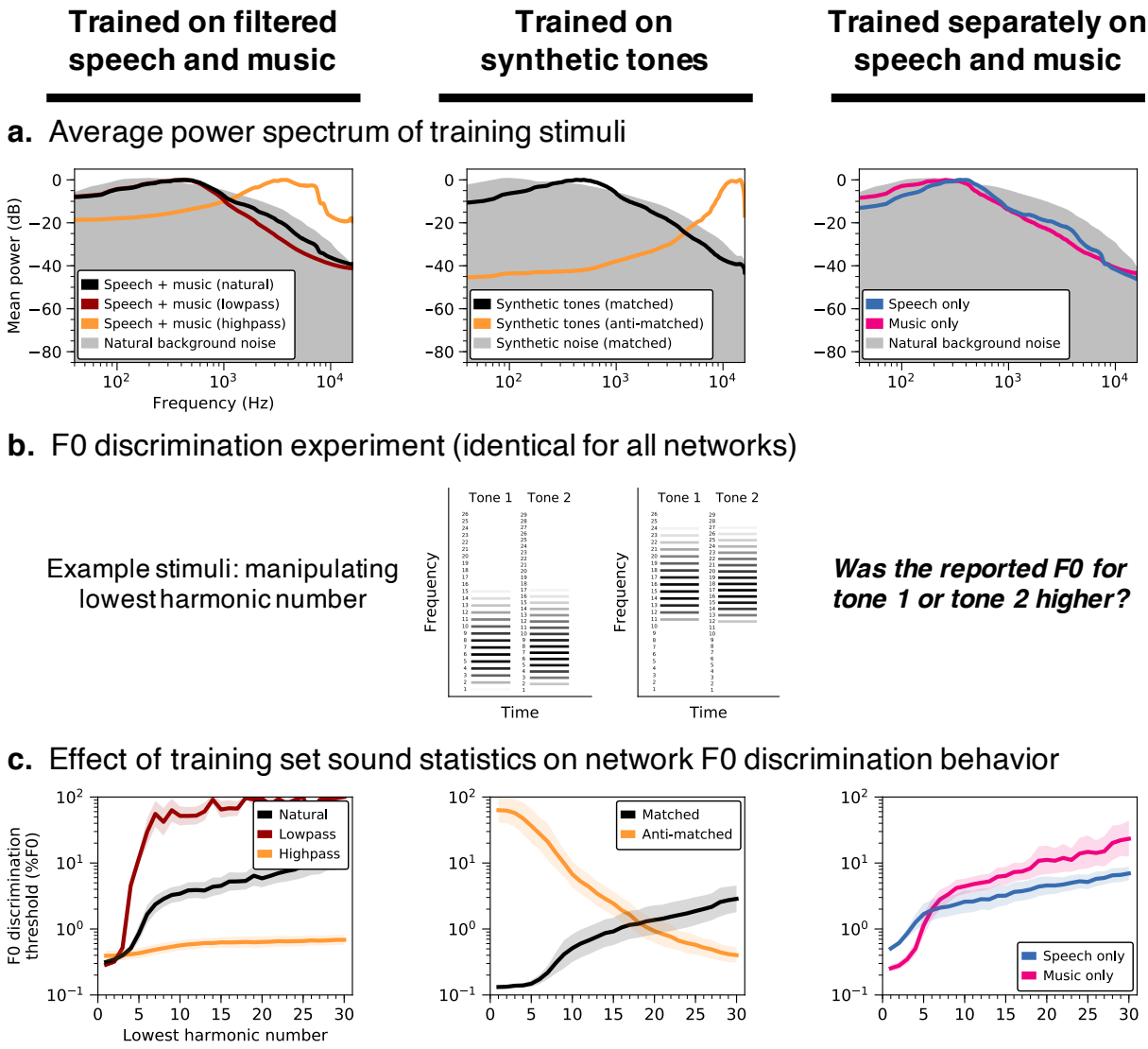

Figure 2.7: Pitch perception depends on training set sound statistics.

(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 sounds were embedded in for training purposes. (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. Lines plot means across the 10 networks; error bars indicate 95% confidence intervals bootstrapped across the 10 networks.

"environment" in which a system is optimized.

2.2.16 Natural spectral statistics account for human-like behavior

To isolate the acoustic properties needed to reproduce human-like pitch behavior, we also trained networks on synthetic tones embedded in masking noise, with spectral statistics matched to those of the natural sound training set (Fig. 3.7A, center 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. 3.7C, center 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 Supplementary Fig. 3.15 for results on the full suite of psychophysical experiments).

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. 3.7A, center 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. 3.7C, center 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.

2.2.17 Music-trained networks exhibit better pitch acuity

We also trained networks separately using only speech or only music stimuli (Fig. 3.7A, right column). Consistent with the more accurate pitch discrimination found in human listeners with musical training [86], networks optimized specifically for music have lower discrimination thresholds for stimuli with low-numbered harmonics (Fig. 3.7C, right column; $t(18) = 9.73, p < 0.001, d = 4.35$, two-sample t-test comparing left-most conditions – which

produce the best thresholds – for speech and music training). As a test of whether this result could be explained by cochlear processing, we repeated this experiment on networks with learnable first-layer filters (as in Fig. 3.4A) and found that networks optimized specifically for music still produced lower absolute thresholds (Supplementary Fig. 2.18). This result likely reflects the greater similarity of the synthetic test tones (standardly used to assess pitch perception) to instrument notes compared to speech excerpts, the latter of which are less perfectly periodic over the stimulus duration.

2.2.18 Training set noise required for "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 in our training set. Networks trained in noisy environments (Fig. 2.8, left) resembled humans in accurately inferring F0 even when the F0 was not physically present in the stimuli (thresholds for stimuli with lowest harmonic number between 2 and 5 were all under 1%). This "missing fundamental illusion" was progressively weakened in networks trained in higher SNRs (Fig. 2.8, center and right), 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. 2.8B, row 2) and frequency-shifted complexes (Fig. 2.8B, row 3), apparently ignoring high-numbered harmonics (correlations with human results were lower in both experiments; $t(18) = 9.08, p < 0.001, d = 4.06$ and $t(18) = 4.41, p < 0.001, d = 1.97$, comparing high vs. no training noise). Conversely, discrimination thresholds for pure tones (Fig. 2.8B, row 5) remained good (below 1%), as though the networks learned 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.

2.2.19 Network neurophysiology

Although our primary focus in this paper was to use DNNs to understand behavior in normative terms, we also examined whether the internal representations of our model might exhibit established neural phenomena.

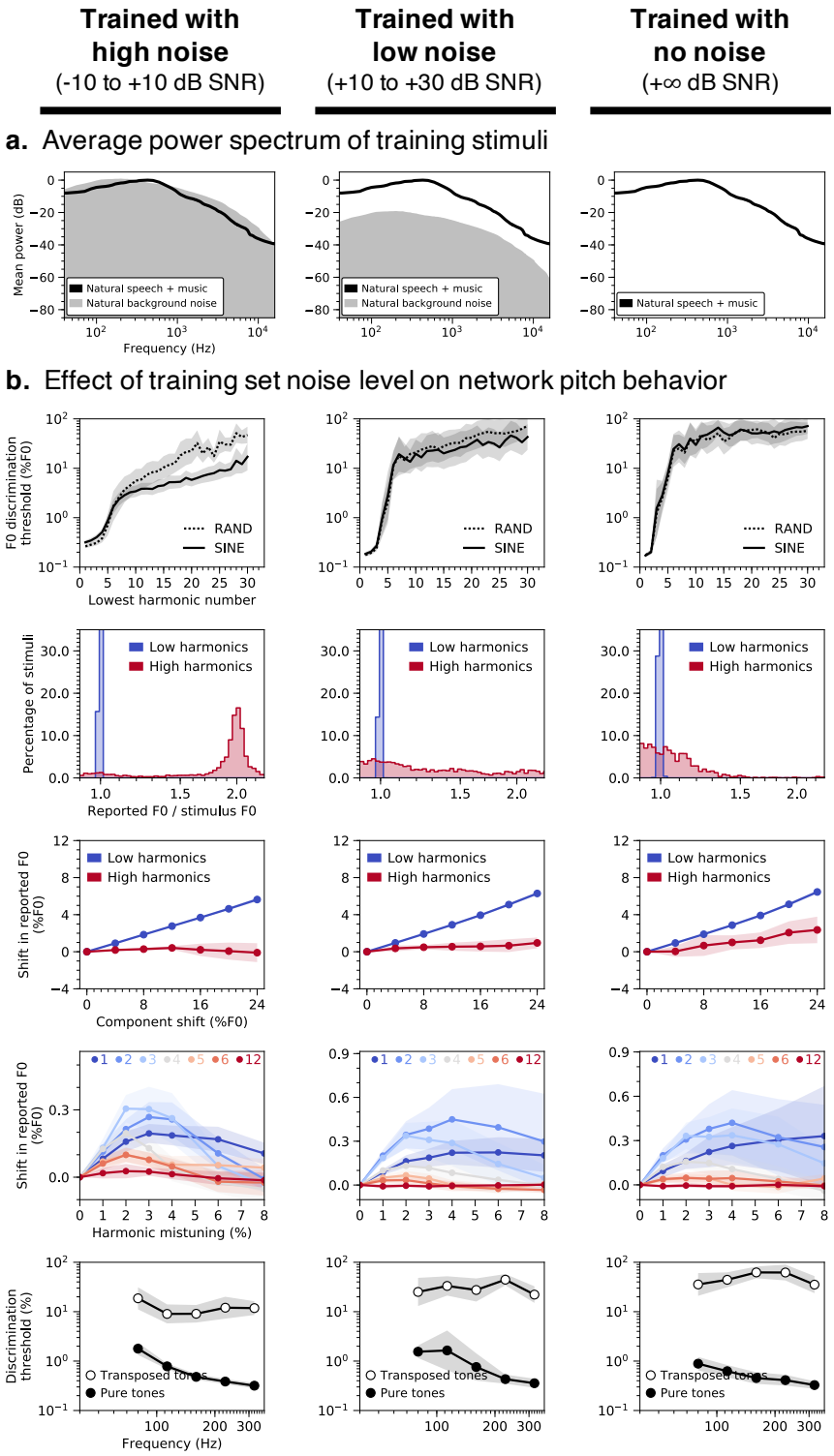

Figure 2.8: Key characteristics of human pitch behavior only emerge in noisy training conditions.

(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. 3.2A-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). Lines plot means across the 10 networks; error bars indicate 95% confidence intervals bootstrapped across the 10 networks.

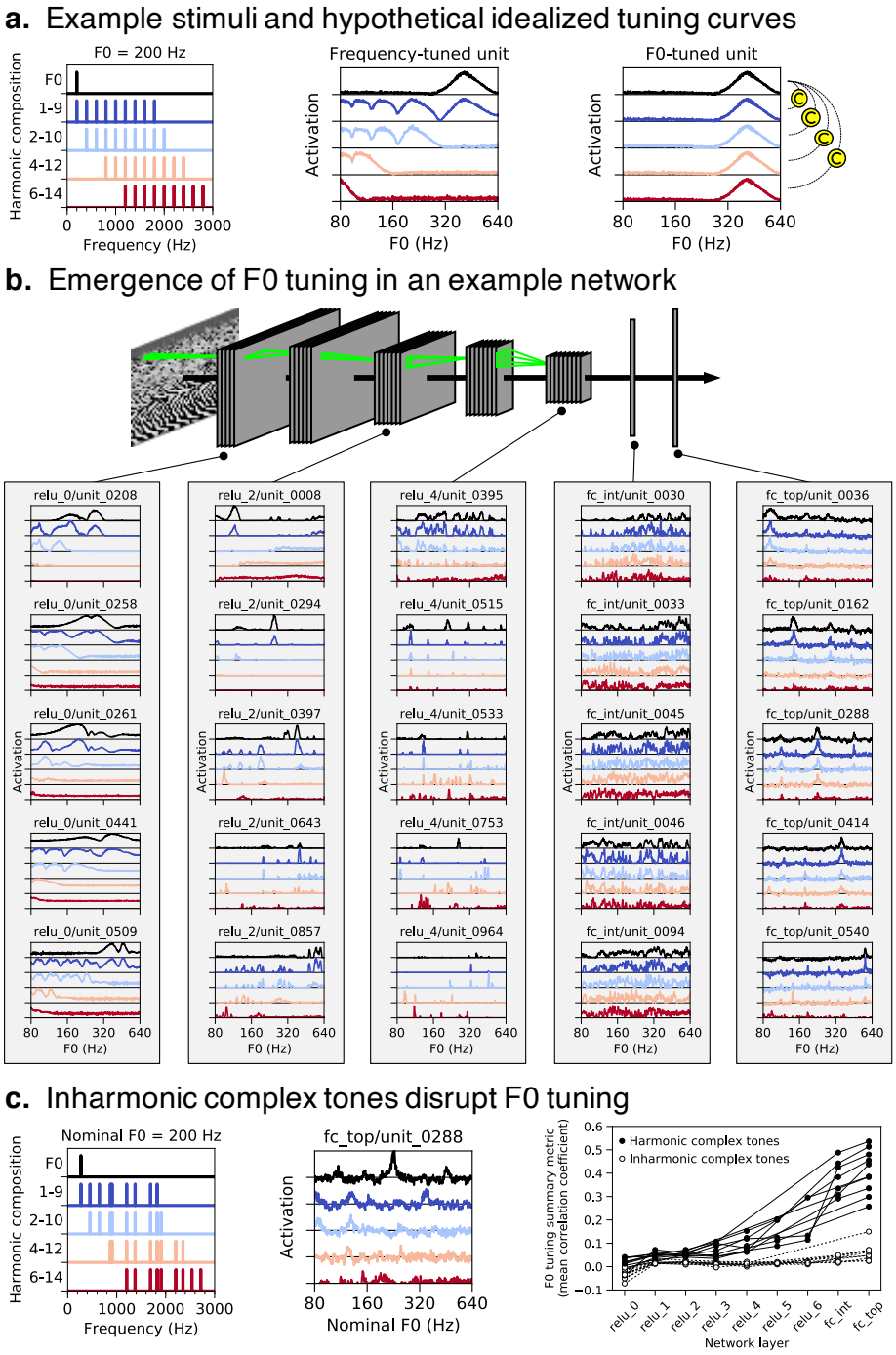

Figure 2.9: Network neurophysiology.

Network activations were measured in response to pure tones and complex tones with four different harmonic compositions. (A) Left: Power spectra for stimuli with 200 Hz F0. Center: expected F0 tuning curves for an idealized frequency-tuned unit. The tuning curves are color-matched to the corresponding stimulus (e.g., black for pure tones and red for harmonics 6-14). A frequency-tuned unit should respond to pure tones near its preferred frequency (414 Hz) or to complex tones containing harmonics near its preferred frequency (e.g., when F0 = 212, 138, 103.5, or 82.8 Hz, i.e. 414/2, 414/3 or 414/4 Hz). Right: expected F0 tuning curves for an idealized F0-tuned unit. An F0-tuned unit should produce tuning curves that are robust to harmonic composition. The strength of a unit's F0 tuning can thus be quantified as the mean correlation between the pure tone (frequency) tuning curve and each of the complex tone tuning curves. (B) F0 tuning curves measured from five representative units in each of five network layers. Units in the first layer ($relu_0$) seem to exhibit frequency tuning. Units in the last layer (fc_{top}) exhibit F0 tuning. (C) Left: Nominal F0 tuning curves were measured for complex tones made inharmonic by jittering component frequencies. Center: Such curves are shown for one example unit in the network's last layer. Unlike for harmonic tones, the tuning curves for tones with different frequency compositions do not align. Right: The overall F0 tuning of a network layer was computed by averaging the F0 tuning strength across all units in the layer. A unit's F0 tuning strength was quantified as the mean correlation between the pure tone (frequency) tuning curve and each of the complex tone tuning curves. For each of our 10 best network architectures, overall F0 tuning (computed separately using either harmonic or inharmonic complex tones) is plotted as a function of network layer. Network units become progressively more F0-tuned deeper into the networks, but only for harmonic tones. (D) Left: Population responses of pitch-selective units in marmoset auditory cortex, human auditory cortex, and our model's output layer, plotted as a function of lowest harmonic number. Marmoset single-unit recordings were made from 3 animals and error bars indicate SEM across 50 neurons (re-plotted from [28]). Center: Human fMRI responses to harmonic tones, as a function of their lowest harmonic number. Data were collected from 13 participants and error bars indicate within-subject SEM (re-plotted from [29]). 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). Right: Network unit activations to harmonic tones as a function of lowest harmonic number. Activations were averaged across all units in the final fully connected layer of our 10 best network architectures (error bars indicate 95% confidence intervals bootstrapped across the 10 best network architectures).

We simulated electrophysiology experiments on our best-performing network architecture by measuring time-averaged model unit activations to pure and complex tones varying in harmonic composition (Fig. 2.9A). F0 tuning curves of units in different network layers (Fig. 2.9B) illustrate a transition from frequency-tuned units in the first layer ($relu_0$, where units responded whenever a harmonic of a complex tone aligned with their pure-tone tuning) to complex tuning in intermediate layers ($relu_2$, $relu_4$, and fc_{int}) to unambiguous F0 tuning in the final layer (fc_{top}), where units responded selectively to specific F0s across different harmonic compositions. These latter units thus resemble pitch-selective neurons identified in primate auditory cortex [64] in which tuning to the F0 of missing-fundamental complexes aligns with pure tone tuning.

We quantified the F0 tuning of individual units by measuring the correlation between pure tone and complex tone tuning curves. High correlations between tuning curves indicate F0 tuning invariant to harmonic composition. In each of the 10 best-performing networks, units became progressively more F0-tuned deeper into the network (Fig. 2.9C, right, solid symbols). Critically, this result depended on the harmonicity of the tones. When we repeated the analysis with complex tones made inharmonic by jittering component frequencies [58] (Fig. 2.9C, left), network units no longer showed F0 tuning (Fig. 2.9B, center) and the dependence on network layer was eliminated (Fig. 2.9C, right, open symbols). In this respect the units exhibit a signature of human F0-based pitch, which is also disrupted by inharmonicity [58], [87], and of pitch-tuned neurons in non-human primates [88].

To compare the population tuning to that observed in the auditory system, we also

measured unit activations to harmonic complexes as a function of the lowest harmonic in the stimulus. The F0-tuned units in our model's final layer responded more strongly when stimuli contained low-numbered harmonics (Fig. 2.9D, right; main effect of lowest harmonic number on mean activation, $F(1.99, 17.91) = 134.69, p < 0.001, \eta^2_{partial} = 0.94$). This result mirrors the response characteristics of pitch-selective neurons (measured with single-unit electrophysiology) in marmoset auditory cortex (Fig. 2.9D, left) [64] and pitch-selective voxels (measured with fMRI) in human auditory cortex (Fig. 2.9D, center) [65].

2.3 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 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 the fundamental frequency of natural sounds, in natural conditions, given a human cochlea.

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 [63], [67]–[74]. Our model's initial stages incorporated detailed simulations of the auditory nerve, 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 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 of frequency from the auditory nerve [9], [10], 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, DNNs instantiate general-purpose functions that can be optimized to perform a training task. They learn to use 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 (Supplementary Fig. 3.10).

Although mechanistic explanations of pitch perception are widely discussed [63], [67]–[72], [74], 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 [60], [76], [81], [84]. However, we found that the dependence on harmonic number could be fully reversed for sufficiently unnatural sound training sets (Fig. 3.7C). Moreover, the dependence was stable across changes in cochlear filter bandwidths (Fig. 3.6C). 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 influence of the natural environment was evident when we eliminated background noise from the training set (Fig. 2.8). Networks trained without background noise did not extract F0 information from high-numbered harmonics, relying entirely on the lowest-numbered harmonics. Such a strategy evidently works well for idealized environments (where the lowest harmonics are never masked by noise), but not for realistic environments containing noise, and diverges from the strategy employed by human listeners. This result suggests that pitch is also in part a consequence of needing to hear in noise, and is consistent with evidence that human pitch perception is highly noise-robust [89]. Together, these two 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 [9], [10], [90], [91]. Although human listeners

often improve on such tasks with practice, there is not much reason to expect humans 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.

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 [68], [69], whereas temporal models quantify signatures of periodicity in temporal patterns of spikes [67], [71]. 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 peripheral resolution in place and time by altering the simulated cochlea. These manipulations provided evidence that fine-grained peripheral timing is critical for normal pitch perception (Fig. 3.5C&E), and that fine-grained place-based frequency tuning is less so (Fig. 3.6). 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. These conclusions were enabled by combining a realistic model of the auditory periphery with task-optimized neural networks.

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. 3.2A 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 [28]. The uniform distributions over sound level and SNR in our training dataset were also not matched in a principled way to the natural world. Discrepancies may also reflect shortcomings of our F0 estimation task (which used only 50ms clips) or peripheral model, which although state-of-the-art and relatively well validated, is imperfect (e.g., peripheral representations consisted of firing rates rather than spikes).

We note that the ear itself is the product of evolution and thus likely itself reflects properties of the natural environment [92]. 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. Consistent with this view, when we replaced the fixed peripheral model with a set of learnable filters operating directly on sound waveforms, networks exhibited less human-like pitch behavior (Fig. 3.4). This result suggests it could be fruitful to incorporate additional stages of peripheral physiology, which might similarly provide constraints on pitch perception.

Our model shares many of the commonly-noted limitations of DNNs as models of the brain [93], [94]. 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 similarities have been documented between trained neural network representations and brain representations [13], [19], and although we saw some such similarities ourselves in the network's activations (Fig. 2.9C), the inconsistencies with biology could 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.

The model 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 [95] or melody [96]. In some cases relative pitch involves comparisons of the spectrum rather than the F0 [58], [87] and/or can be biased by changes in the timbre of a sound [97], for reasons that are not well understood. The framework used here could help to develop normative understanding of such effects, by incorporating more complicated tasks (e.g., involving speech or music) and then characterizing the pitch-related behavior that results. DNNs 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 [13], [26], [98].

The approach we used here has natural extensions to understanding other aspects of hearing [34], 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 [4] and which often entails particular types of deficits in pitch perception [83]. 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 [99]. Models optimized with different types of simulated electrical stimulation could clarify the patterns of behavior to expect. Models trained with either acoustically- or electrically-stimulated peripheral auditory representations (or combinations thereof) and then tested with electrically-stimulated input could yield insights into the variable outcomes of pediatric cochlear implantation. Similar approaches could be applied to study acclimatization to hearing aids in adults.

There is also growing evidence for species differences in pitch perception [100], [101]. Our approach could be used to relate species differences in perception to species differences in the cochlea [102] or to differences in the acoustic environment and/or tasks a species may be optimized for. While our results suggest that differences in cochlear filters alone are unlikely to explain differences in pitch perception abilities across species, they leave open the possibility that human pitch abilities reflect the demands of speech and music, which plausibly require humans to be more sensitive to small F0 differences than other species. This issue could be clarified by optimizing network representations for different auditory tasks.

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.

2.4 Methods

2.4.1 Natural sounds training dataset - overview

The main training set consisted of 50ms excerpts of speech and musical instruments. This duration was chosen to enable accurate pitch perception in human listeners [103], but to be 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. Overall stimulus presentation levels were drawn uniformly between 30 dB SPL and 90 dB SPL. All training stimuli were sampled at 32 kHz.

2.4.2 Speech and music training excerpts

We used STRAIGHT [104] to compute time-varying F0 and periodicity traces for sounds in several large corpora of recorded speech and instrumental music: Spoken Wikipedia Corpora (SWC) [105], Wall Street Journal (WSJ), CMU Kids Corpus, CSLU Kids Speech, NSynth [106], and RWC Music Database. 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 $\pm 3\%$ 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 training 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). Segments were assigned to bins according to their time-averaged F0. The resulting training dataset consisted of 3000 exemplars 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 the goal of using 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)

2.4.3 Background noise for training data

To make the F0 estimation task more difficult and to simulate naturalistic listening conditions, each speech or instrument excerpt in the training dataset was embedded in natural background noise. The signal-to-noise ratio for each training example was drawn uniformly between -10 dB and +10 dB. Noise source clips were taken from a subset of the AudioSet corpus [107], 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) [108] from all recordings, and discarding segments over which these statistics were not stable in time, as in previous studies [109]. 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.

2.4.4 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 [4], [5]. 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 synapse between inner hair cells and auditory nerve fibers. Although the model's responses have only been directly compared to recordings made in nonhuman animals, some model parameters have been inferred for humans (such as the bandwidths of cochlear filters) on the basis of behavioral and otoacoustic measurements [85].

Because the majority of auditory nerve fibers, especially those linked to feedforward projections to higher auditory centers, have high spontaneous firing rates [110], [111], we used exclusively high spontaneous rate fibers (70 spikes/s) as the input to our model. 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 additionally 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 Supplementary Fig. 2.19. 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 exclusively used high spontaneous rate fibers in all other experiments.

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 [112] 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 [113]. We confirmed that increasing the number of channels by a factor of 10 had little effect on the behavioral results from our main natural sound training condition (Supplementary 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. The middle 50ms was then excerpted, leaving a 100-fiber by 1000-timestep array of instantaneous firing rates that constituted the input to the neural networks.

2.4.5 Deep neural network models - overview

The 100-by-1000 simulated auditory nerve representations were passed into deep convolutional neural networks, each consisting of a series of feedforward layers. These layers were hierarchically organized and instantiated one of a number of 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 for a stimulus should be unimodal. 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 (Supplementary Fig. 2.20A). This performance asymptote could reflect the limits of the F0 labels the network was trained on. As it happened, with this bin width of 1/16 of a semitone it was possible to attain discrimination thresholds for synthetic tones that were on par with the best thresholds typically measured in human listeners ($\sim 0.1 - 0.4\%$) [76], [81] for some model architectures and auditory nerve settings. Discrimination thresholds were worse for wider classification bins (Supplementary Fig. 2.20B). 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.

2.4.6 Definitions of constituent neural network operations

Convolutional layer: A convolutional layer implements the convolution of a bank of N_k two-dimensional linear 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 a natural choice for models of sensory systems as their input has translation-invariant temporal statistics. Because translation invariance does not hold for the frequency dimension, convolution in frequency is less obviously a natural model constraint. However, many types of sound signals are well described by approximate translation invariance in local neighborhoods of the frequency domain, and classical auditory models can often be described as imposing convolution in

frequency [114], [115]. Moreover, imposing convolution greatly reduces the number of parameters to be learned. We have empirically found that auditory neural network models often train more readily when convolution in frequency is imposed, 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 timesteps, and a placeholder 1 in the filter kernel dimension.

A convolutional layer is defined by five parameters:

- h : height of the convolutional filter kernels (number of filter taps in the frequency dimension)
- w : width of the convolutional filter kernels (number of filter taps in the time dimension)
- N_k : number of different convolutional filter kernels
- W : Trainable weights for each of the
- N_k filter kernels; W has shape $[h, w, M_k, N_k]$
- 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:

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

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 overlapped 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 [93]. 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

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

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. 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 $[N_f \times N_t \times N_k]$. Then, X_{flat} is linearly transformed to give an output Y with shape $[N_{out}]$:

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

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 that the expected sum over all outputs is equal to the expected sum over all inputs. The $r \times N_{in}$ 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. 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:

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

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

2.4.7 Model optimization - architecture search

All of our DNN 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; Supplementary Fig. 3.10) regardless of the number of fully connected layers. When we refer to network "activations" in a given convolutional layer (Fig. 2.9), 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 DNN 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 16, 32, or 64 (each sampled with probability=1/3). The number of kernels in each successive layer could 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 DNNs, we imposed 16 and 1024 as lower and upper bounds on the number of kernels in a single layer and capped the frequency $\times$ 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.

2.4.8 Model optimization - 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. 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 single DNN architecture, we selected the 10 architectures that produced the highest validation set accuracies

to use as our model experimental "participants" (collectively referred to as "the model"). We re-trained all 10 architectures for each manipulation of the peripheral auditory model (Figs. 4-6) and the training set sound statistics (Figs. 7 and 8). The 10 different network architectures are described in Supplementary Table 3.1.

2.4.9 Network psychophysics - overview

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. These experiments are denoted A through E in the following sections to align with Fig. 3.2 (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 50ms to accommodate the input length for the 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. In most cases, individual subject data was not available (the original studies were performed 17 to 36 years prior to this work), so we report only across-subject means and do not include error bars for human data.

2.4.10 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 [81], 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 were audible.

Human experiment

The human F0 discrimination thresholds (previously published by Bernstein and Oxenham) 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 [63]. Each participant completed 4 adaptive tracks per condition (where a condition had a particular lowest harmonic number and either random or sine phase). 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.

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 ($n_{low} = 1, 2, 3, \ldots 30$), and 10 reference F0s ($F_{(0,ref)}$) spaced uniformly on a logarithmic scale between 100 and 300 Hz. Within each condition, each network was evaluated on 121 stimuli with slightly different F0s (within $\pm 6\%$ 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 the frequency of the lowest harmonic for that condition and F0 ($n_{low} \times F_{(0,ref)}$). 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 7,260 possible pairwise comparisons between the 121 stimuli for a condition. 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. We next constructed a psychometric function by plotting the percentage of correct trials as a function of %F0 difference between two stimuli. We then averaged psychometric functions across the 10 reference F0s with the same harmonic phase and lowest harmonic number. Network thresholds were thus based on 1210 stimuli (72,600 pairwise F0 discriminations) per condition. Normal cumulative distribution functions were fit to the 60 resulting psychometric functions (2 phase conditions x 30 lowest harmonic numbers). To match human F0 discrimination thresholds, which were measured with a 2-down-1-up adaptive algorithm, 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

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) assumed that F0 discrimination thresholds were constant for lowest harmonic numbers between 1 and 5 (supported by other published data [65], [81]). We then computed the Pearson correlation coefficient between log-transformed vectors of human and network thresholds.

2.4.11 Experiment B: pitch of alternating-phase harmonic complexes

Experiment B reproduced the stimulus manipulation of Shackleton and Carlyon (1994) to test if our networks exhibited 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 the lowest F0 our networks could report was 80 Hz. 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 provides a proxy for 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) as 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.

2.4.12 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 $x = 1 - |(F - F_e)/(1.5 \times F0)|$ 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 frequency component shifts of +0, +4, +8, +12, +16, +20, and +24 %F0. For each combination of spectral envelope and frequency component shift, we generated stimuli with 3917 nominal F0s spaced log-uniformly between 80 and 480 Hz (83,391 stimuli in total). These stimuli are a superset of those used in the human experiment, which measured shifts for three F0s (100, 200, and 400 Hz) and four component shifts (+0, +8, +16, +24 %F0). As in the original study, stimuli were presented at overall levels of 70 dB SPL.

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. 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.

Model experiment

For the model experiment, we again took network F0 predictions for the 83,391 frequency-shifted complexes as the "perceived" F0s. F0 predictions were restricted 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

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.

2.4.13 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 exhibited 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 147 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. Humans were tested on 126 of the 147 conditions in the model experiment (3 nominal F0s x 7 component shifts x 6 harmonic numbers) – the conditions with a harmonic number of 12 were not included.

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 26,166 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 147 mean shift values corresponding to the 147 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 147 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.

2.4.14 Experiment E: frequency discrimination with pure and transposed tones

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 10,080 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 233,586 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 $\times$ 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.

2.4.15 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. 3.5D).

Stimuli

We generated pure tones at 6,144 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. 3.5E.

Model experiment

We used the same procedure used in Experiments A and E to measure frequency discrimination thresholds. 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.

2.4.16 Auditory nerve manipulations - overview

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 the trained networks using modified auditory nerve representations of the same test stimuli. We used this approach to investigate whether a biologically-constrained cochlea is necessary to obtain human-like pitch behavior and to evaluate the dependence of network pitch behavior on both temporal and "place" information in the auditory nerve representation. The only exception to this general procedure was in the experiment that tested the effect of flattening the excitation pattern, which was performed on networks that were trained on normal auditory nerve representations (see below).

2.4.17 Replacing the hardwired cochlear model with learnable filters

We replaced the hardwired auditory nerve model with a convolutional layer whose weights could be optimized alongside the rest of the DNN (Fig. 3.4). The convolutional layer consisted of 100 one-dimensional filter kernels (each with 801 taps) that operated directly on 32 kHz audio, applied using "valid" convolution. The audio input to the network was 75ms in duration such that the valid output of convolution was 50ms, as in the hardwired cochlear representation. Outputs from the 100 filters were stacked, half-wave rectified, and then resampled to 20 kHz, resulting in a first-layer representation with 100 "nerve fibers" and 1000 timesteps to match the size and temporal resolution of the hardwired cochlear

representations. We separately trained the 10 best network architectures from the original architecture search with this learnable "cochlear" layer. The best frequency of a learned filter was determined after training from the maximum value of its transfer function.

2.4.18 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 [5]. 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. 3.5A). 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 for the 50 Hz hair cell cutoff preserved all stimulus-related information), 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, allowing us to use the same network architectures as in all other training conditions. 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 upsampled 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 original 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 (Supplementary Fig. 3.11).

2.4.19 Eliminating place cues by flattening the excitation pattern

To test if our trained networks made use of peaks and valleys in the time-averaged excitation pattern (which provide "place" cues to F0), we tested networks on nerve representations with artificially flattened excitation patterns. Nerve representations were flattened by separately scaling each frequency channel of the nerve representation to have the same time-averaged response. Each row (nerve fiber) of the nerve representation was divided by its time-averaged firing rate and multiplied by the mean firing rate across all rows, yielding an excitation-flattened nerve representation with the same mean firing rate as the original. This manipulation was separately applied to each psychophysical stimulus.

2.4.20 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 [85]. 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 the frequency scaling found 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 tuning curves (detection thresholds as a function of frequency) for simulated nerve fibers with characteristic frequencies of 250, 500, 1000, 2000, 4000 Hz (Fig. 3.6A&B). Mean firing rate responses were computed for each fiber to 50ms pure tones with frequencies between 125 Hz and 8000 Hz.

Thresholds were defined as the minimum sound level required to increase the fiber's mean firing rate response 10% above its spontaneous rate (i.e., dB SPL required for 77 spike/s).

2.4.21 Sound statistics manipulations - overview

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.

2.4.22 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).

2.4.23 Training on spectrally matched and anti-matched synthetic tones

To investigate the extent to which network pitch behavior could be explained by low-order 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 [104] 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 spectral envelope using 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.

2.4.24 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 background noise clips sampled from the same sources used for the combined dataset (SNRs drawn uniformly from -10 to +10 dB).

2.4.25 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 that were masked by noise in the original 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.

2.4.26 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 network layers as model "neurons" and looked at their tuning using their average activations across time to different stimuli. We measured tuning properties using equal-amplitude sine-phase harmonic complex tones (45 dB SPL per frequency component) in threshold equalizing noise (10 dB SPL per ERB). We used a total of 11,520 stimuli: 5 unique harmonic compositions (pure tones or successive harmonics 1-9, 2-10, 4-12, or 6-14) $\times$ 2304 unique F0s (logarithmically-spaced between 80 and 640 Hz). For each unit, we constructed an F0 tuning curve for each harmonic composition by averaging activations to stimuli within the same F0 classification bin (i.e., within 1/16 semitone bins). Tuning curves were normalized separately for each unit by dividing by the unit's maximum response across the full stimulus set. Units that produced a response of zero to all of the test stimuli were excluded from analysis ($<$1% of units).

As a measure of the strength of F0 tuning in a single model unit, we computed the mean Pearson correlation coefficient between the pure tone (frequency) tuning curve and each of the complex tone tuning curves. A perfectly F0-tuned unit should selectively respond to pure and complex tones of a preferred F0 independent of harmonic composition (Fig. 2.9A),

yielding a high mean correlation coefficient. We measured the F0 tuning correlation of each unit in each layer (all ReLU activations following convolutional layers and fully connected layers) of the 10 best-performing network architectures. The F0 tuning strength of a network layer was computed by averaging this metric across all units in the layer.

To test if the observed F0 tuning depended on the harmonicity of the stimuli, we repeated this analysis with complex tones made inharmonic by randomly shifting component frequencies with a fixed jitter pattern [58]. The jitter pattern allowed for each individual component (after the F0 component) to be shifted by up to $\pm 50\%$ of the F0. Jitter values for each component were drawn uniformly from -50% to +50% with rejection sampling to ensure adjacent components were separated by at least 30 Hz to minimize salient differences in beating. Each component's frequency was the original harmonic's frequency plus the jitter value multiplied by the nominal F0. Like harmonic tones, inharmonic tones generated in this way have frequency components that increase in spacing as the nominal F0 is increased, but unlike harmonic tones they lack a fundamental frequency in the range of audible pitch (i.e., above 30Hz [115]). Empirically, human perceptual signatures of F0-based pitch are disrupted by inharmonicity [58], [87], [89], making it a way to distinguish human-like representations of F0 from coarser representations of frequency spacing. The same jitter pattern (i.e., the same mapping of nominal harmonic numbers to jitter value) was applied to all stimuli regardless of F0 and harmonic composition. As in the F0-tuning analysis with harmonic tones, we measured the average strength of nominal F0 tuning for each layer in the 10 networks. To ensure results were not unduly biased by a single random jitter pattern, the analysis was repeated five times with different random seeds. F0 tuning summary metrics reported in Fig. 2.9C are averaged across these five random seeds.

We measured population responses as a function of lowest harmonic number (Fig. 2.9D) using a superset of the harmonic complex tones used to measure F0 tuning. The dataset was expanded to include all complexes containing 9 successive harmonics, with lowest harmonic numbers 1 to 15 (e.g., $1-9, 2-10, 3-11, \ldots 15-26$). 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 700 units in each network's final fully connected layer. We then constructed lowest-harmonic-number tuning curves by taking responses to stimuli at the best F0 with lowest harmonic numbers 1 to 15. 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 [64] and pitch-selective voxels in human auditory cortex [65]. 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 fMRI responses to bandpass-filtered sine-phase harmonic complex tones and frequency-matched noise. The 12 participants (4 male, 8 female, ages 21-28) were non-musicians with normal hearing. 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.

2.4.27 Statistics - analysis of human data

Human data was scanned from original figures or provided by the authors of the original papers. We did not have access to data from individual human participants, and so did not plot error bars on the graphs of human results.

2.4.28 Statistics - analysis of human-model comparison metrics

Human-model comparison metrics were computed separately for each psychophysical experiment (as described in the above sections on each experiment and its analysis) 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. 3.3, 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 (Supplementary Fig. 3.10), 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 comparing the human-model similarity metrics of the 54 single-convolutional-layer networks to those of the other 346 networks. 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.

These human-model similarity metrics were used to analyze two other experiments (in all others we used more fine-grained analysis of best thresholds and transition points, described below). To assess the statistical significance of changes in human-model similarity when networks were optimized with a learned “cochlea” (Fig. 3.4) or in the absence of background

noise, 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 (Supplementary Table 1). We performed two-sample t-tests (each sample containing results from the 10 independently trained networks) to compare human-model comparison metrics between training conditions. Effect sizes were quantified as Cohen's d and reported for all such tests that indicated statistically significant differences. Because human-model similarity metrics were bounded between -1 and 1, we passed the metrics through an inverse normal cumulative distribution function before performing t-tests. All t-tests and rank-sum tests were two-sided.

2.4.29 Statistics - analysis of best thresholds and transition points

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 [76], [81]. To assess the effect of altered cochlear input or training sound statistics (Figs. 5, 6, 7), we thus focused on two measures: first, the absolute F0 discrimination acuity of our model when all low-numbered harmonics were present ("best threshold"), and second, the harmonic number at which discrimination ability transitioned from good to poor ("transition point"). 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. To quantify effect sizes, Cohen's d is reported for all two-sample t-tests that indicated statistically significant differences.

2.4.30 Statistics - ANOVAs on discrimination thresholds

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 and stimulus presentation level) as within-subject factors. The specific pairings of these different factors were: stimulus presentation level vs. auditory nerve phase locking cutoff (Fig. 3.5E), lowest harmonic number vs. training set spectral statistics (Fig. 3.7C), and lowest harmonic number vs. training set noise level (Fig. 2.8). We also performed a repeated-measures ANOVA to test for a main effect of lowest har-

monic number on population responses in the network's last layer (Fig. 2.9C). 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.

2.5 Supplementary information

2.5.1 Data availability

The Wall Street Journal (LDC93S6A), CMU Kids Corpus (LDC97S63), and CSLU Kids Speech (LDC2007S18) audio datasets used in this study are available from the Linguistic Data Consortium (https://www.ldc.upenn.edu). The Spoken Wikipedia Corpora audio dataset is available at https://nats.gitlab.io/swc. The RWC Music Database is available at https://staff.aist.go.jp/m.goto/RWC-MDB. The pitch datasets we compiled from these publicly available corpora, along with the psychophysical test stimulus sets, are available at: https://github.com/msaddler/pitchnet.

2.5.2 Code availability

Source code for the Bruce et al. (2018) auditory nerve model is available from the authors. We developed a Python wrapper around the model, which supports flexible manipulation of cochlear filter bandwidths and the upper limit of phase locking. This wrapper is available at https://github.com/msaddler/bez2018model. Code to implement and analyze our deep neural network pitch models (including trained network weights) is available at https://github.com/msaddler/pitchnet.

2.5.3 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, Brian Moore, and Hideki Kawahara for helpful discussions. Work supported by NSF grant BCS-1634050 and NIH grant R01DC017970.

2.5.4 Author contributions

All authors conceived the project and designed the experiments. R.G. ran and analyzed pilot versions of the experiments. M.R.S. ran the experiments described in the paper, analyzed

the data, and made the figures. M.R.S. and J.H.M. drafted the manuscript. All authors edited the manuscript.

2.5.5 Supplementary figures

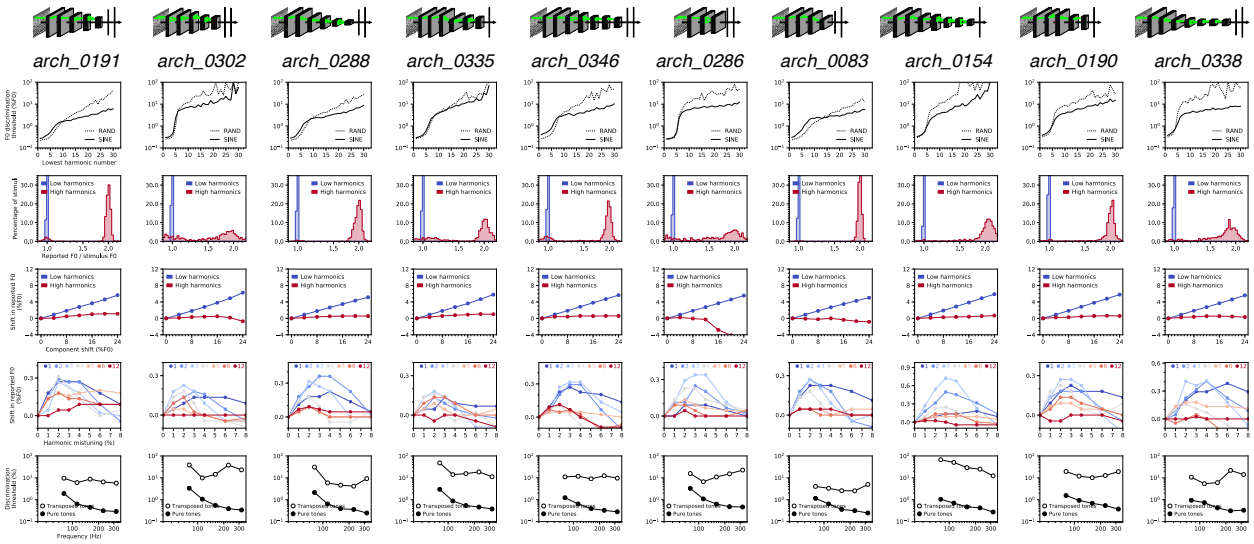

Figure 2.10: Pitch behavior of the 10 best network architectures ranked by F0 estimation performance on natural sounds.

Each column shows results from a single neural network architecture (depicted at the top). Detailed descriptions of each architecture are provided in Supplementary Table 1. The five rows in this grid correspond to the five main psychophysical experiments (see Fig. 3.2a-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).

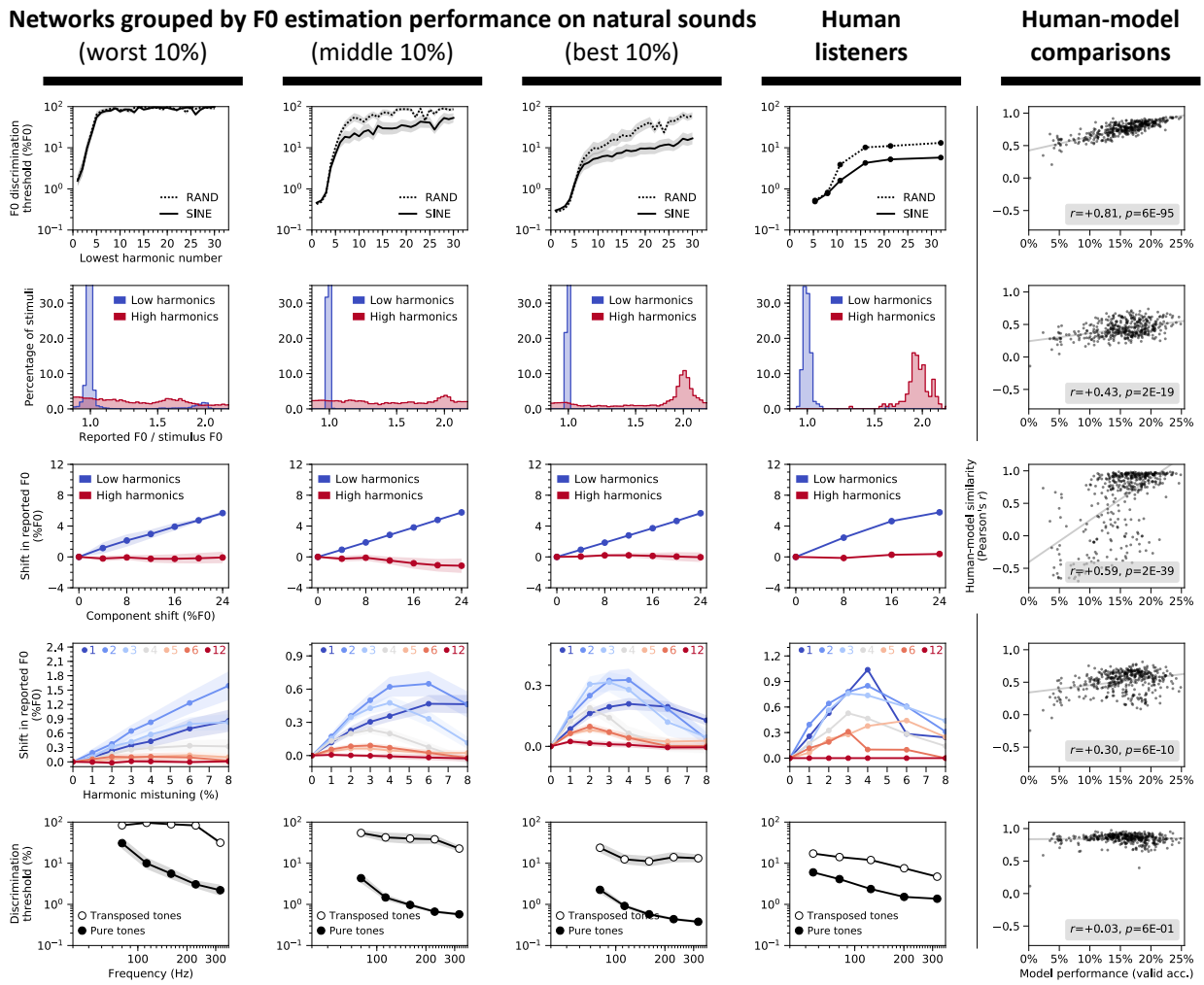

Figure 2.11: Network architectures producing better F0 estimation for natural sounds exhibit more human-like pitch behavior.

The five rows in this grid correspond to the five main psychophysical experiments (see Fig. 3.2a-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). First three columns show results measured from the 40 worst, middle, and best-performing network architectures (out of the 400 architectures trained in our architecture search, ranked by F0 estimation performance on natural sounds), respectively. Error bars plot bootstrapped 95% confidence intervals around the mean across the 40 networks. Human results (reproduced from Fig. 3.1) are shown in column 4. Column 5 contains scatter plots of human-model behavioral similarity (quantified as a correlation between the results of each model and that of humans) vs. validation set accuracy for all trained networks (reproduced from Fig. 3.3). Pearson correlations between validation set accuracy and human-model similarity for each experiment are noted in the legends.

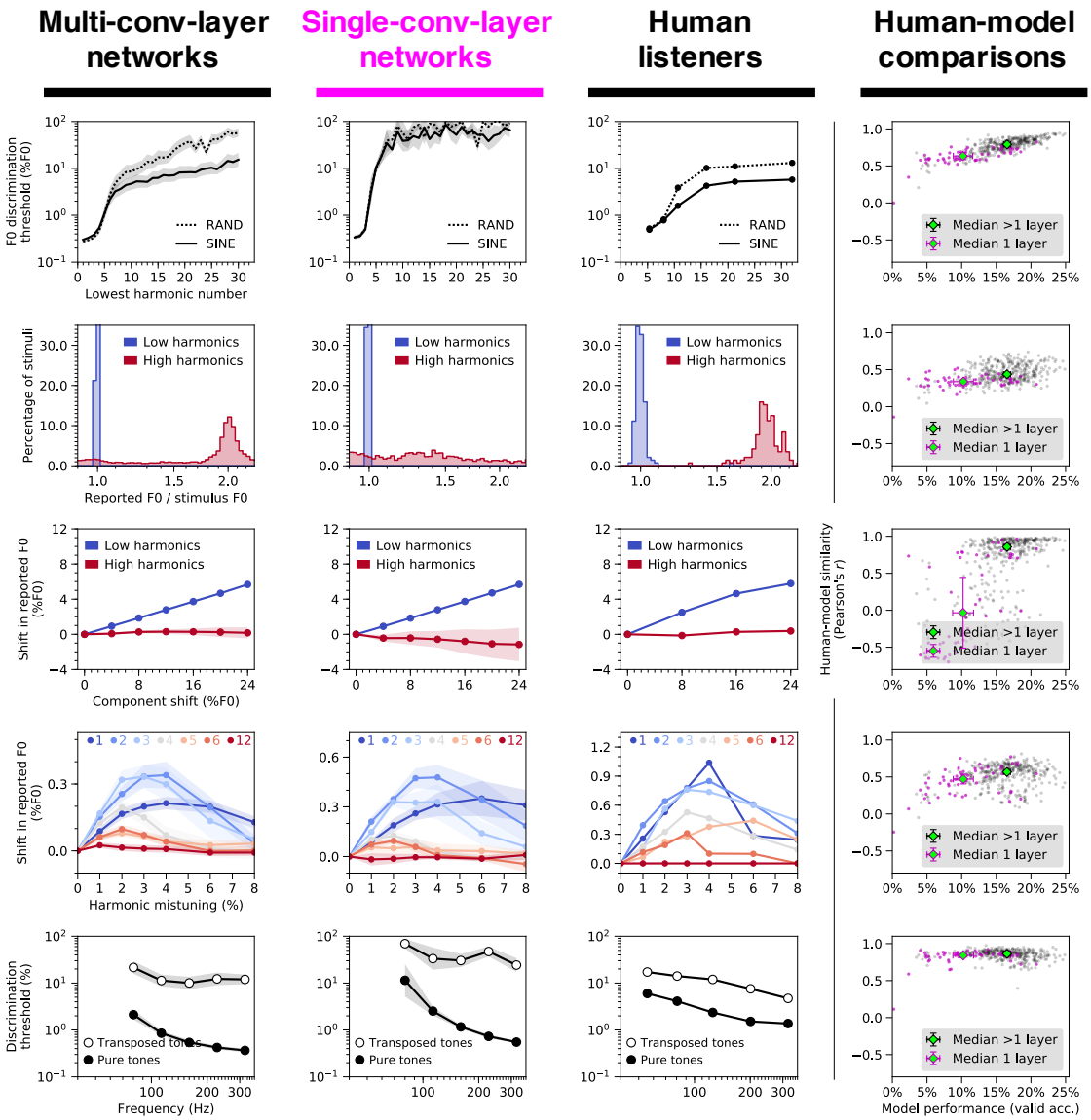

Figure 2.12: Deep networks better account for human psychophysical behavior than networks with just one convolutional layer.

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), shown on the x-axis of the graphs in the right column. The single-convolutional-layer networks also produce less human-like psychophysical results. The five rows in this grid correspond to the five main psychophysical experiments (see Fig. 3.2a-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). In column 1, network psychophysical results are shown averaged across the top 10% (34 of 346) of networks (ranked by validation set performance) containing more than one convolutional layer. In column 2, results are shown averaged across the top 10% (5 of 54) of networks containing just one convolutional layer. Human results are shown in column 3. Column 4 contains scatter plots of human-model behavioral similarity (quantified as a correlation between the results of each model and that of humans) vs. validation set accuracy for all trained networks. Gray and magenta data points correspond to multi-convolutional-layer and single-convolutional-layer networks, respectively. Median data points are included for each group to illustrate how single-layer networks generally both performed worse on the F0 estimation task and produced poorer matches to human behavior. Error bars plot bootstrapped 95% confidence intervals around the mean across architectures. When pooled across all five experiments, human-model similarity was lower for single-convolutional-layer networks than the remaining 346 multi-convolutional-layer networks ($z = 9.24, p < 0.001$, two-sided Wilcoxon rank-sum test). Moreover, the top 40% of networks ranked according to overall human-model similarity consisted entirely of multi-convolutional-layer networks.

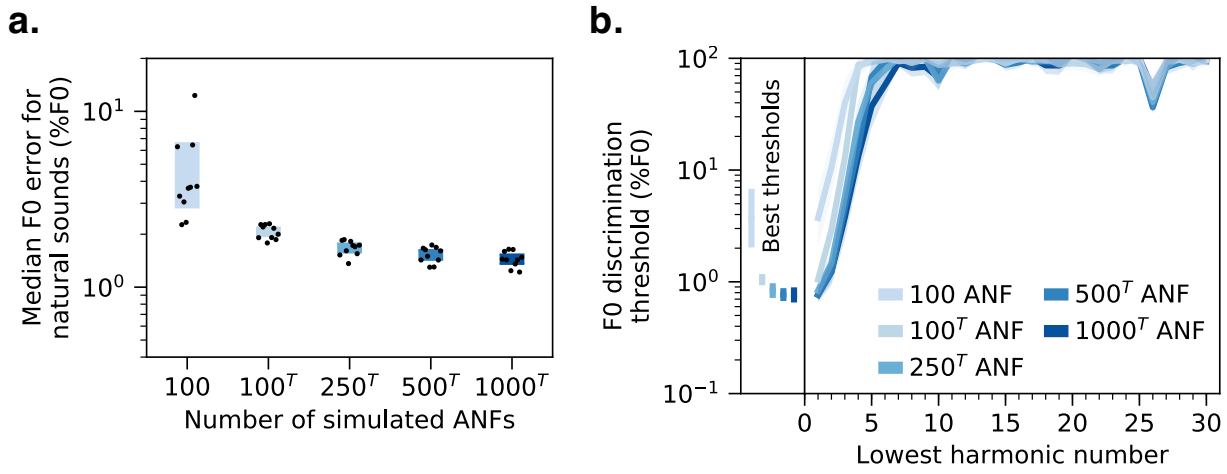

Figure 2.13: Effect of number of auditory nerve fibers on F0 estimation error and discrimination thresholds measured from networks without spike-timing information.

(a) Median F0 estimation error on the natural sounds validation set for 10 networks trained and tested with five different peripheral model configurations varying in the number of input auditory nerve fibers (ANFs) and in whether the frequency and time dimensions were "transposed" (indicated by superscript T). The network architectures were constrained to take a 100-by-1000 array as input. The default input representation was 100 frequency channels (auditory nerve fibers) by 1000 timesteps. A transposed input representation consists of 1000 frequency channels by 100 timesteps. To manipulate the number of nerve fibers while keeping the size of the input representation fixed, transposed peripheral representations with fewer than 1000 nerve fibers were upsampled to 1000 (via linear interpolation) along the frequency (auditory nerve fiber) dimension. Time was sampled at 2 kHz to yield 100 timesteps. For all networks included in the plot, the IHC filter cutoff frequency was set to 50 Hz to eliminate all phase-locked temporal information (see Fig. 3.5). (b) F0 discrimination thresholds as a function of lowest harmonic number of synthetic tones, measured from the same networks as (a). The best thresholds are re-plotted to the left of the main axes. Error bars plot bootstrapped 95% confidence intervals around the mean across the 10 best network architectures.

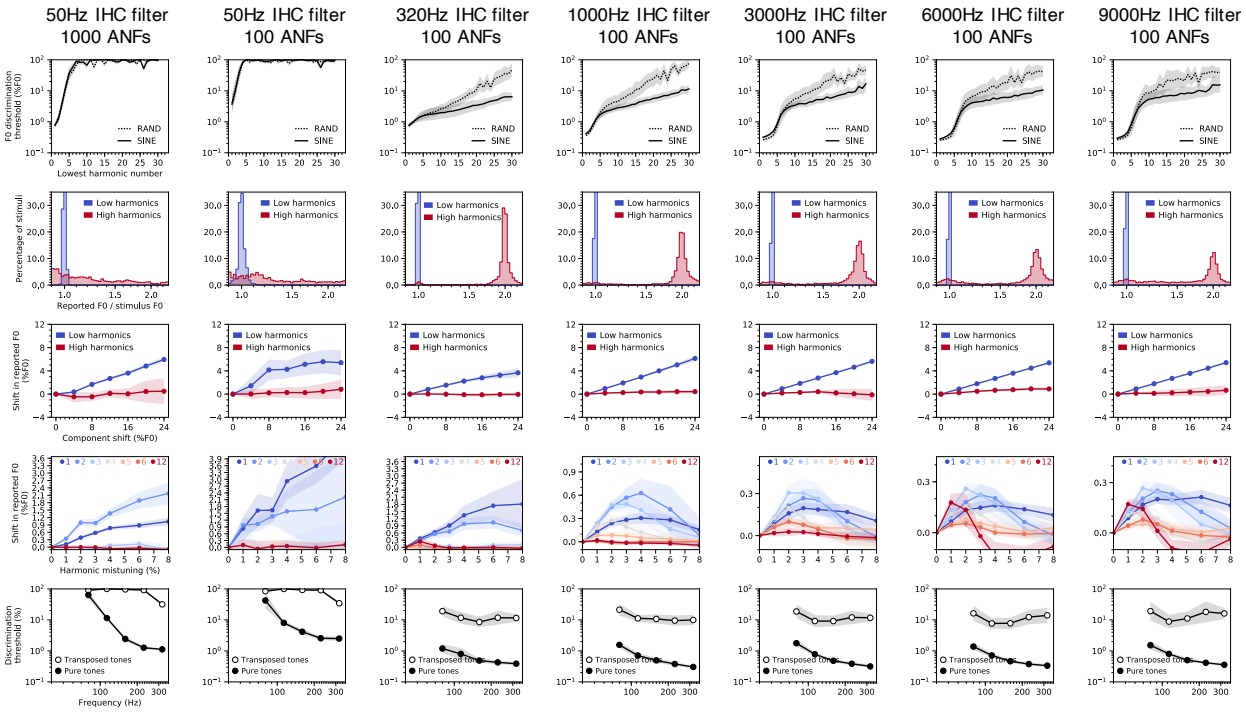

Figure 2.14: Effects of phase locking cutoff on network pitch behavior.

Columns correspond to networks trained and tested with seven different configurations of the peripheral auditory model. Configurations differed in the upper frequency limit of auditory nerve phase locking (inner hair cell lowpass filter cutoff) and the number of auditory nerve fibers (ANFs) (see Fig. 3.5). Rows correspond to the five main psychophysical experiments (see Fig. 3.2a-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 plot bootstrapped 95% confidence intervals around the mean across the 10 best network architectures.

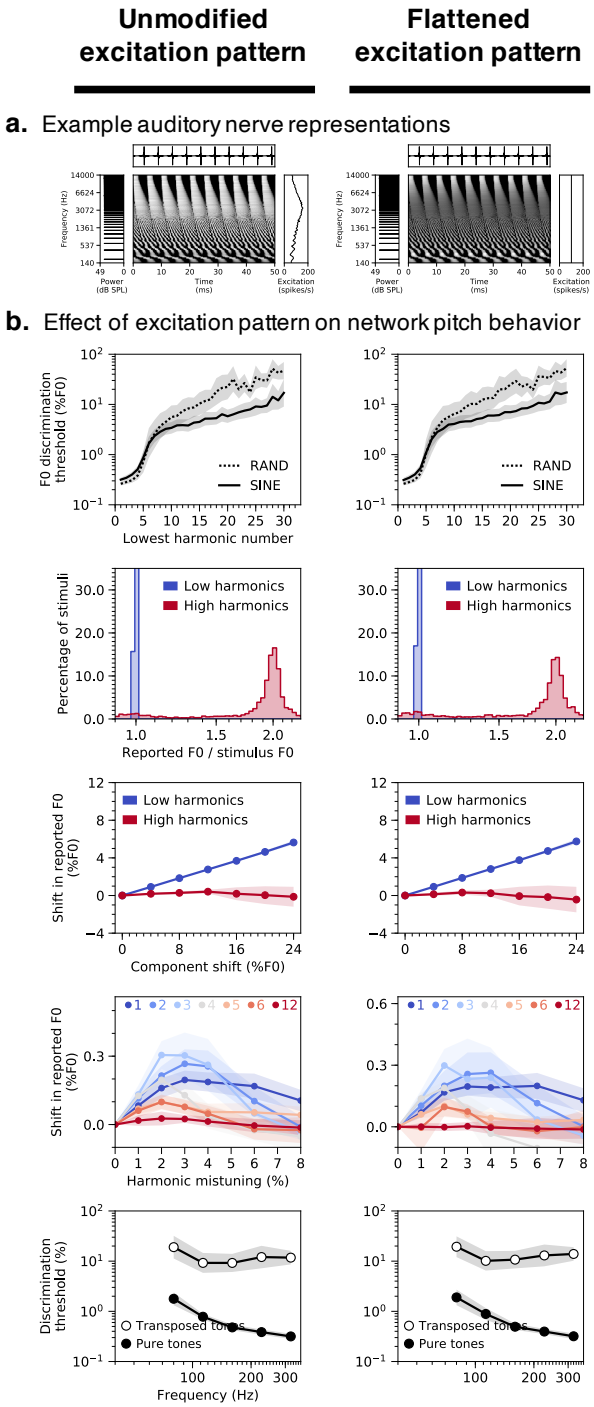

Figure 2.15: Networks do not rely on place cues to F0 in the excitation pattern to produce human-like pitch behavior.

(a) Simulated peripheral representations of the same stimulus (harmonic tone with 200 Hz F0) with unmodified (left column) or flattened (right column) time-averaged excitation patterns. The peaks and valleys in the unmodified excitation pattern, which provide place cues to F0, were eliminated by separately scaling each frequency channel of the nerve representation to have the same time-averaged response. (b) Psychophysical results from the best-performing network architecture tested on auditory nerve representations with either unmodified or flattened excitation patterns. In both cases, the model was trained on auditory nerve representations with unmodified excitation patterns. Rows correspond to the five main psychophysical experiments (see Fig. 3.2a-e): F0 discrimination as a function of harmonic number and phase (row 1), pitch estimation of alternating-phase stimuli (row 2), frequency-shifted complexes (row 3), and complexes with individually mistuned harmonics (row 4), and frequency discrimination with pure and transposed tones (row 5). Error bars plot bootstrapped 95% confidence intervals around the mean across the 10 best network architectures.

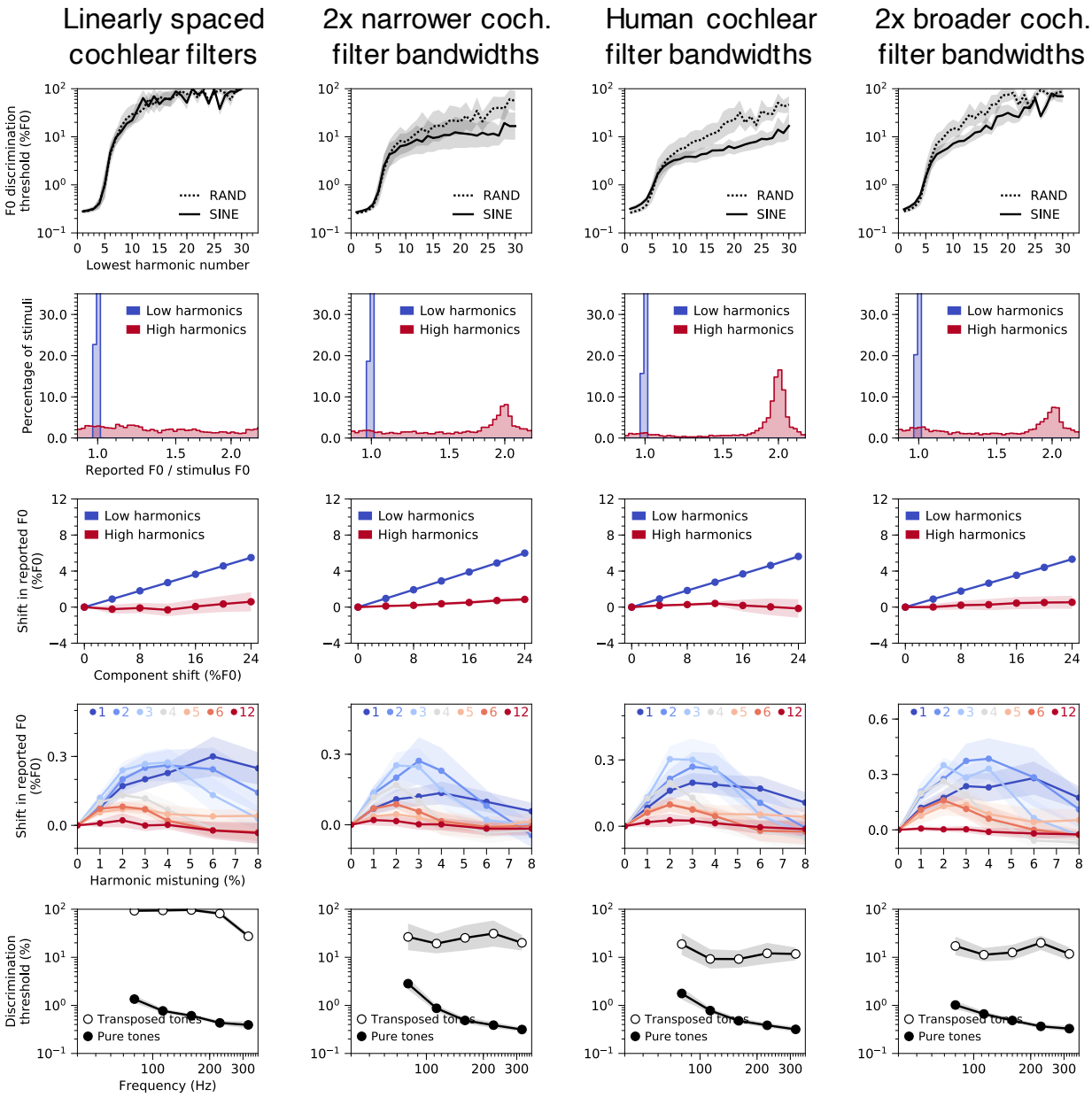

Figure 2.16: Effects of altered cochlear frequency selectivity on network pitch behavior.

The four columns correspond to networks trained and tested with four different settings of cochlear frequency selectivity: linearly spaced cochlear filters with constant bandwidths (column 1), cochlear filters with bandwidths two times narrower than those estimated for normal hearing-humans but normally spaced (i.e., evenly spaced on an ERB scale, to best approximate the spacing believed to characterize the ear) (column 2), normally spaced cochlear filters with bandwidths matched to those of normal-hearing humans (column 3), and cochlear filters with two times broader bandwidths but normally spaced (column 4) (see Fig. 3.6). Rows correspond to the five main psychophysical experiments (see Fig. 3.2a-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 plot bootstrapped 95% confidence intervals around the mean across the 10 best network architectures. Psychophysical results were qualitatively robust to changes in peripheral frequency tuning. The main exceptions 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; phase randomization: $t(18) = 3.13, p < 0.01, d = 1.40$; alternating phase: $t(18) = 6.50, p < 0.001, d = 2.91$; two-sided two-sample t-tests). These results are to be expected because the sharp tuning of the linearly spaced filters (Fig. 3.6B) results in less interaction between adjacent harmonics, which is believed to drive phase effects.

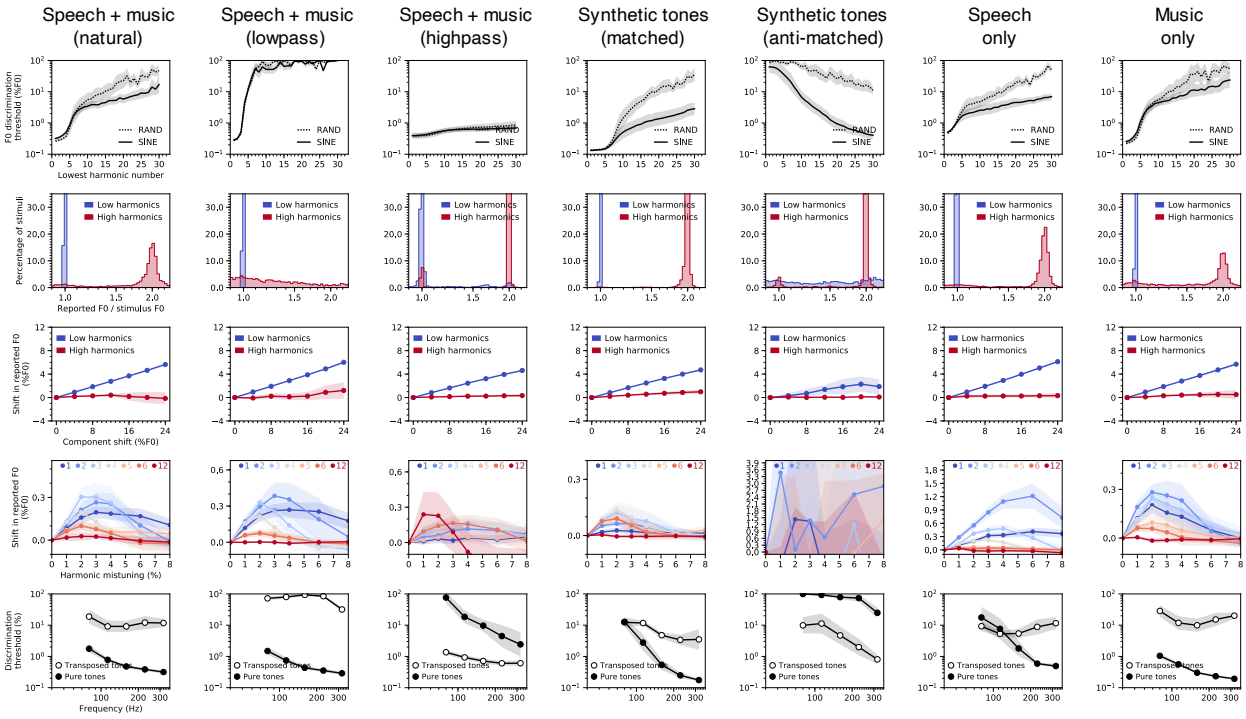

Figure 2.17: Effects of training set sound statistics on network pitch behavior.

Columns correspond to different training datasets (described in column titles). Rows correspond to the five main psychophysical experiments (see Fig. 3.2a-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). Results are shown for the best-performing network architecture, averaged across 10 instances of the architecture trained from different random initializations. Error bars plot bootstrapped 95% confidence intervals around the mean across the 10 best network architectures.

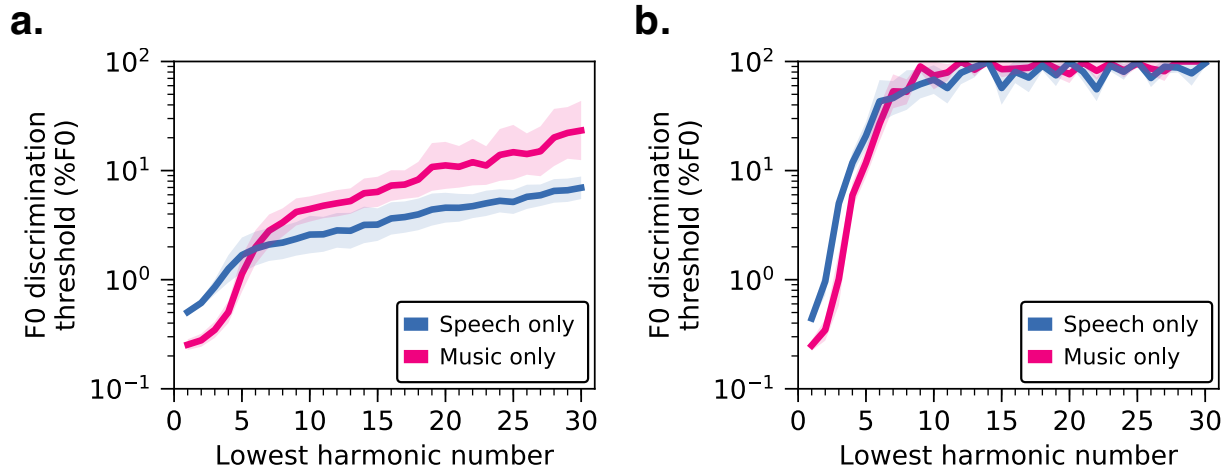

Figure 2.18: F0 discrimination thresholds as a function of lowest harmonic number, measured from networks trained separately on speech-only and music-only datasets.

Results from networks trained on simulated auditory nerve representations produced by a fixed peripheral auditory model (reproduced from Fig. 3.7c). (b) Results from networks trained directly on sound waveforms (first-layer "cochlear" filters were learned alongside the rest of the network weights; see Fig. 3.4). Error bars plot bootstrapped 95% confidence intervals around the mean across the 10 best network architectures.

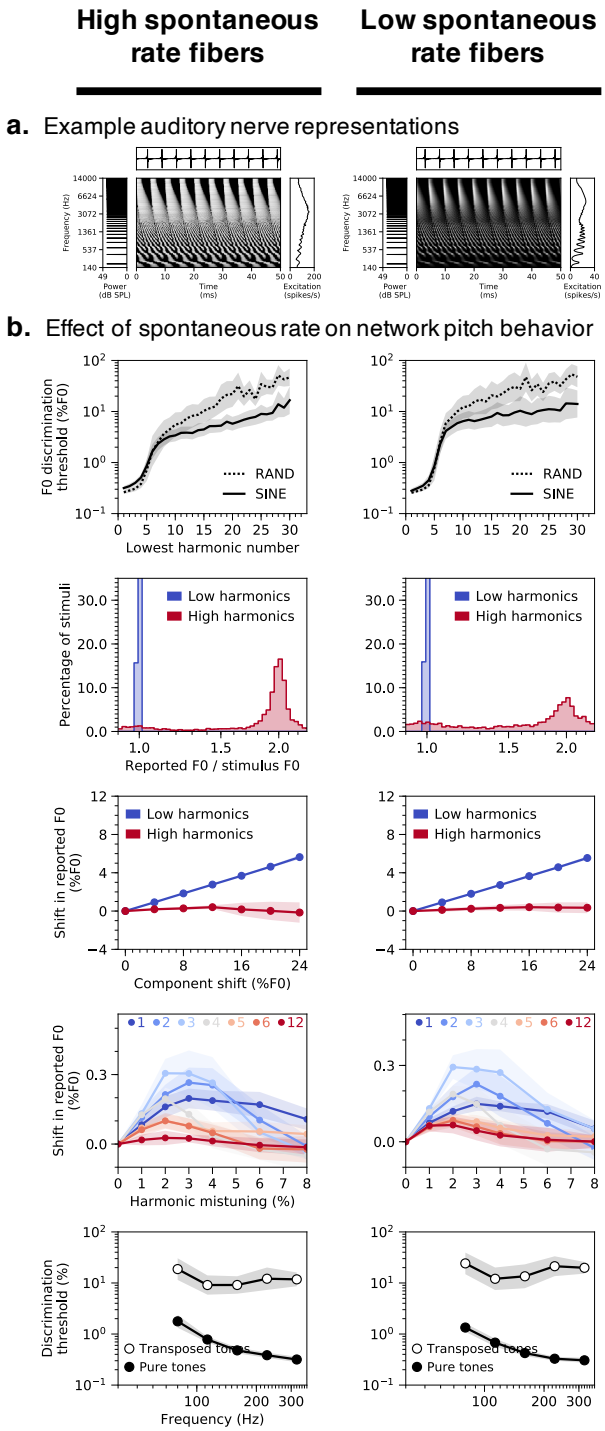

Figure 2.19: Effects of auditory nerve fiber type (high vs. low spontaneous rate) on network pitch behavior.

(a) Simulated peripheral representations of the same stimulus (harmonic tone with 200 Hz F0) with high (70 spikes/s; left column) and low (0.1 spikes/s; right column) spontaneous rate auditory nerve fibers. (b) Psychophysical results from the best-performing network architecture trained and tested with each of the two different nerve fiber types. Rows correspond to the five main psychophysical experiments (see Fig. 3.2a-e): F0 discrimination as a function of harmonic number and phase (row 1), pitch estimation of alternating-phase stimuli (row 2), frequency-shifted complexes (row 3), and complexes with individually mistuned harmonics (row 4), and frequency discrimination with pure and transposed tones (row 5). Error bars plot bootstrapped 95% confidence intervals around the mean across the 10 best network architectures.

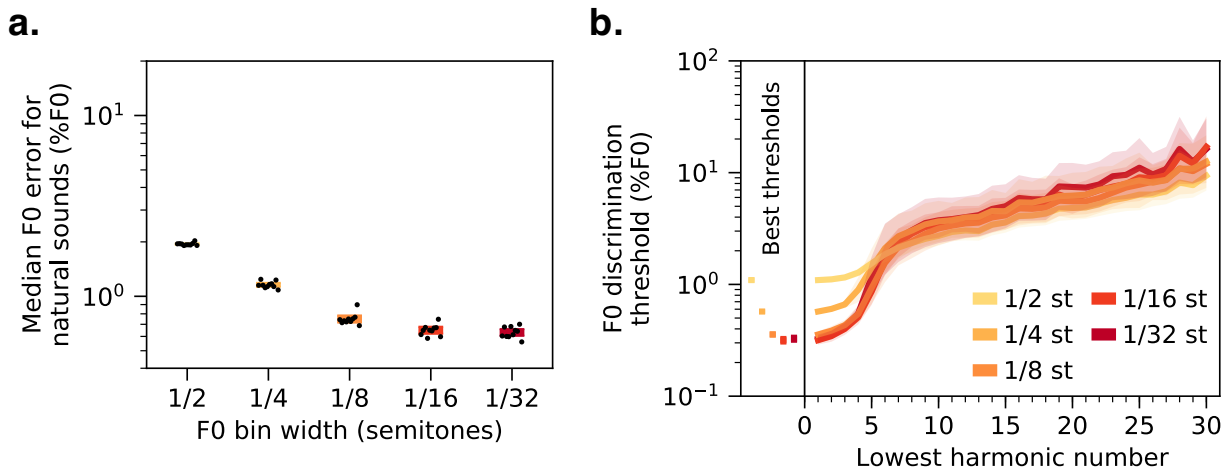

Figure 2.20: Effects of F0 classification bin width on F0 estimation error and discrimination thresholds.

(a) Median F0 estimation error on the natural sounds validation set for 10 networks trained and tested with five different F0 classification bin widths. (b) F0 discrimination thresholds as a function of lowest harmonic number of synthetic tones, measured from the same networks as a. The best thresholds are re-plotted to the left of the main axes. Error bars plot bootstrapped 95% confidence intervals around the mean across the 10 best network architectures.

Architecture	*arch_0191*	*arch_0302*	*arch_0288*	*arch_0335*	*arch_0346*	*arch_0286*	*arch_0083*	*arch_0154*	*arch_0190*	*arch_0338*
Validation set rank (of 400)	1	2	3	4	5	6	7	8	9	10
Validation set accuracy (%)	24.9	24.6	24.1	23.6	23.6	23.5	23.4	23.2	22.7	22.7
Operation	input [100, 1000, 1]	input [100, 1000, 1]	input [100, 1000, 1]	input [100, 1000, 1]	input [100, 1000, 1]	input [100, 1000, 1]	input [100, 1000, 1]	input [100, 1000, 1]	input [100, 1000, 1]	input [100, 1000, 1]
1	conv_1 [2, 83, 32]	conv_1 [1, 250, 32]	conv_1 [3, 77, 64]	conv_1 [1, 71, 32]	conv_1 [3, 53, 32]	conv_1 [1, 180, 64]	conv_1 [2, 82, 32]	conv_1 [1, 70, 64]	conv_1 [2, 97, 32]	conv_1 [2, 110, 64]
2	relu_1 [99, 918, 32]	relu_1 [100, 751, 32]	relu_1 [98, 924, 64]	relu_1 [100, 930, 32]	relu_1 [98, 948, 32]	relu_1 [100, 821, 64]	relu_1 [99, 919, 32]	relu_1 [100, 931, 64]	relu_1 [99, 904, 32]	relu_1 [99, 891, 64]
3	pool_1 [1, 2]	pool_1 [1, 5]	pool_1 [1, 2]	pool_1 [1, 1]	pool_1 [1, 2]	pool_1 [2, 6]	pool_1 [1, 2]	pool_1 [1, 5]	pool_1 [1, 6]	pool_1 [3, 2]
4	norm_1 [99, 459, 32]	norm_1 [100, 151, 32]	norm_1 [98, 462, 64]	norm_1 [100, 930, 32]	norm_1 [98, 474, 32]	norm_1 [50, 137, 64]	norm_1 [99, 460, 32]	norm_1 [100, 187, 64]	norm_1 [99, 151, 32]	norm_1 [33, 446, 64]
5	conv_2 [1, 164, 64]	conv_2 [19, 11, 64]	conv_2 [1, 193, 128]	conv_2 [1, 114, 32]	conv_2 [1, 60, 64]	conv_2 [2, 37, 128]	conv_2 [1, 162, 64]	conv_2 [7, 21, 128]	conv_2 [5, 11, 64]	conv_2 [1, 126, 128]
6	relu_2 [99, 296, 64]	relu_2 [82, 141, 64]	relu_2 [98, 270, 128]	relu_2 [100, 817, 32]	relu_2 [98, 415, 64]	relu_2 [49, 101, 128]	relu_2 [99, 299, 64]	relu_2 [94, 167, 128]	relu_2 [95, 141, 64]	relu_2 [33, 321, 128]
7	pool_2 [3, 7]	pool_2 [1, 7]	pool_2 [4, 3]	pool_2 [1, 3]	pool_2 [2, 4]	pool_2 [1, 1]	pool_2 [2, 2]	pool_2 [4, 3]	pool_2 [1, 1]	pool_2 [3, 3]
8	norm_2 [33, 43, 64]	norm_2 [82, 21, 64]	norm_2 [25, 90, 128]	norm_2 [100, 273, 32]	norm_2 [49, 104, 64]	norm_2 [49, 101, 128]	norm_2 [50, 150, 64]	norm_2 [24, 56, 128]	norm_2 [95, 141, 64]	norm_2 [11, 107, 128]
9	conv_3 [5, 9, 128]	conv_3 [12, 9, 128]	conv_3 [8, 10, 128]	conv_3 [1, 86, 64]	conv_3 [3, 46, 128]	conv_3 [15, 10, 128]	conv_3 [1, 72, 128]	conv_3 [4, 26, 256]	conv_3 [1, 56, 128]	conv_3 [4, 30, 256]
10	relu_3 [29, 35, 128]	relu_3 [71, 13, 128]	relu_3 [18, 81, 128]	relu_3 [100, 188, 64]	relu_3 [47, 59, 128]	relu_3 [35, 92, 128]	relu_3 [50, 79, 128]	relu_3 [21, 31, 256]	relu_3 [95, 86, 128]	relu_3 [8, 78, 256]
11	pool_3 [1, 7]	pool_3 [3, 1]	pool_3 [2, 6]	pool_3 [4, 1]	pool_3 [1, 6]	pool_3 [1, 1]	pool_3 [4, 2]	pool_3 [1, 6]	pool_3 [4, 7]	pool_3 [2, 5]
12	norm_3 [29, 5, 128]	norm_3 [24, 13, 128]	norm_3 [9, 14, 128]	norm_3 [25, 188, 64]	norm_3 [47, 10, 128]	norm_3 [35, 92, 128]	norm_3 [13, 40, 128]	norm_3 [21, 6, 256]	norm_3 [24, 13, 128]	norm_3 [4, 16, 256]
13	conv_4 [4, 3, 256]	conv_4 [7, 7, 256]	conv_4 [2, 2, 256]	conv_4 [13, 13, 128]	conv_4 [8, 1, 256]	fc_1 [512]	conv_4 [6, 3, 128]	conv_4 [2, 1, 512]	conv_4 [8, 5, 256]	conv_4 [1, 5, 256]
14	relu_4 [26, 3, 256]	relu_4 [18, 7, 256]	relu_4 [8, 13, 256]	relu_4 [13, 176, 128]	relu_4 [40, 10, 256]	relu_fc_1 [512]	relu_4 [8, 38, 128]	relu_4 [20, 6, 512]	relu_4 [17, 9, 256]	relu_4 [4, 12, 256]
15	pool_4 [2, 1]	pool_4 [1, 1]	pool_4 [2, 2]	pool_4 [1, 8]	pool_4 [2, 2]	norm_fc_1 [512]	pool_4 [2, 5]	pool_4 [2, 1]	pool_4 [1, 2]	pool_4 [1, 3]
16	norm_4 [13, 3, 256]	norm_4 [18, 7, 256]	norm_4 [4, 7, 256]	norm_4 [13, 22, 128]	norm_4 [20, 5, 256]	dropout [512]	norm_4 [4, 8, 128]	norm_4 [10, 6, 512]	norm_4 [17, 5, 256]	norm_4 [4, 4, 256]
17	conv_5 [5, 2, 512]	conv_5 [5, 3, 512]	conv_5 [2, 1, 512]	conv_5 [2, 10, 256]	conv_5 [7, 2, 256]	fc_out [700]	fc_1 [128]	conv_5 [5, 3, 256]	conv_5 [1, 3, 256]	conv_5 [2, 2, 512]
18	relu_5 [9, 2, 512]	relu_5 [14, 5, 512]	relu_5 [3, 7, 512]	relu_5 [12, 13, 256]	relu_5 [14, 4, 256]		relu_fc_1 [128]	relu_5 [6, 4, 256]	relu_5 [17, 3, 256]	relu_5 [3, 3, 512]
19	pool_5 [1, 1]	pool_5 [3, 1]	pool_5 [1, 1]	pool_5 [1, 3]	pool_5 [1, 1]		norm_fc_1 [128]	pool_5 [1, 1]	pool_5 [2, 1]	pool_5 [1, 1]
20	norm_5 [9, 2, 512]	norm_5 [5, 5, 512]	norm_5 [3, 7, 512]	norm_5 [12, 5, 256]	norm_5 [14, 4, 256]		dropout [128]	norm_5 [6, 4, 256]	norm_5 [9, 3, 256]	norm_5 [3, 3, 512]
21	fc_1 [256]	fc_1 [1024]	conv_6 [2, 4, 1024]	conv_6 [3, 2, 512]	conv_6 [2, 2, 512]		fc_out [700]	conv_6 [2, 2, 256]	dropout [6912]	conv_6 [1, 1, 1024]
22	relu_fc_1 [256]	relu_fc_1 [1024]	relu_6 [2, 4, 1024]	relu_6 [10, 4, 512]	relu_6 [13, 3, 512]			relu_6 [5, 3, 256]	fc_out [700]	relu_6 [3, 3, 1024]
23	norm_fc_1 [256]	norm_fc_1 [1024]	pool_6 [1, 1]	pool_6 [2, 1]	pool_6 [2, 1]			pool_6 [1, 1]		pool_6 [1, 1]
24	dropout [256]	dropout [1024]	norm_6 [2, 4, 1024]	norm_6 [5, 4, 512]	norm_6 [7, 3, 512]			norm_6 [5, 3, 256]		norm_6 [3, 3, 1024]
25	fc_out [700]	fc_out [700]	fc_1 [256]	dropout [10240]	conv_7 [1, 1, 512]			conv_7 [3, 1, 256]		conv_7 [1, 1, 1024]
26			relu_fc_1 [256]	fc_out [700]	relu_7 [7, 3, 512]			relu_7 [3, 3, 256]		relu_7 [3, 3, 1024]
27			norm_fc_1 [256]		pool_7 [1, 1]			pool_7 [1, 1]		pool_7 [1, 1]
28			dropout [256]		norm_7 [7, 3, 512]			norm_7 [3, 3, 256]		norm_7 [3, 3, 1024]
29			fc_out [700]		fc_1 [512]			dropout [2304]		fc_1 [256]
30					relu_fc_1 [512]			fc_out [700]		relu_fc_1 [256]
31					norm_fc_1 [512]					norm_fc_1 [256]
32					dropout [512]					dropout [256]
33					fc_out [700]					fc_out [700]
34										

Table 2.1: Details of 10 best network architectures.

Columns correspond to 10 distinct convolutional neural network architectures (the 10 best-performing networks identified in our random architecture search). Rows include descriptors of constituent network operations. Grey horizontal bands group operations by convolutional layer. With two exceptions, all results figures present results averaged across all 10 architectures. The two exceptions are Fig. 2.9a, which features only the best-performing architecture (arch_0191), and Supplementary Fig. 3.8, which displays results separately for each of these 10 architectures. Legend:

- $conv[h, w, k]$: convolutional layer with h = kernel height (frequency dimension), w = kernel width (time dimension), and k = number of kernels
- $relu[N_f, N_t, N_k]$: rectified linear unit activation function operating on inputs with the specified shape (N_f = frequency dimension, N_t = time dimension, and N_k= kernel dimension)
- $pool[s_f, s_t]$: weighted averaged pooling operation with stride s_f in the frequency dimension and stride s_t in the time dimension
- $norm[N_f, N_t, N_k]$: batch normalization operating on inputs with the specified shape (N_f = frequency dimension, N_t = time dimension, and N_k = kernel dimension)
- $fc[N]$: fully-connected layer with N units
- $dropout$: dropout regularization with 50% dropout rate

Chapter 3

The role of temporal coding in hearing: evidence from task-optimization

Mark R. Saddler, Josh H. McDermott

Abstract

Neurons encode information in the timing of their spikes in addition to their firing rates. The fidelity of spike timing is particularly high in the auditory nerve, whose action potentials are phase-locked to the fine-grained temporal structure of sound with sub-millisecond precision. However, the perceptual role of this temporal coding in hearing remains controversial. We investigated the issue with machine learning models optimized for real-world hearing tasks, asking whether phase-locked spike timing in a model's cochlear input was necessary to obtain human-like behavior. Models with high-frequency phase locking replicated human behavior across all tested regimes. Degraded phase-locking produced inhuman performance on some tasks, impairing sound localization, pitch perception, and voice recognition while leaving word recognition largely intact. The results link neural coding to real-world perception and clarify conditions in which prostheses that fail to restore high-fidelity temporal coding (e.g., contemporary cochlear implants) could in principle restore near-normal hearing.

3.1 Introduction

Sensory systems encode information about an organism's environment in the spiking activity of neurons. From electrophysiology, we know a great deal about how stimulus properties are represented at different stages of neural processing. Comparatively less is known about how this information gives rise to complex human behavior. In perceptual science, ideal observer models have long been used to analyze which features of a neural code contribute to behavior [9], [116], [117]. An ideal observer derives the statistically optimal solution to a perceptual task given the information available at some stage of neural processing [11]. Since evolutionary pressures presumably drive biological perceptual systems in the direction of optimal performance, models that instantiate optimal task solutions under different biological constraints might elucidate the underpinnings of human behavior. This approach has produced elegant computational accounts for many aspects of vision [53]–[55], [118], [119] and hearing [9], [10], [68], [120]–[122]; however, it is limited to tasks and stimuli for which provably optimal solutions can be derived (i.e., for which probability distributions of the generative parameters can be specified). This limitation precludes much of real-world behavior as in most cases it is not clear what the generative parameters of natural stimuli even are, much less how to analytically describe their distributions or relate them to neural coding. Because natural stimuli and tasks are precisely those that biological systems are likely to have been optimized for, ideal observers have had limited applicability in domains where they might otherwise be most useful.

Here, we propose machine learning as an alternative approach to link neural coding to behavior. Contemporary machine learning models are highly expressive mathematical functions that can be optimized to perform natural tasks with natural stimuli[1]. These new tools – used in place of provably ideal observers – may offer a way to resurrect this classic approach for real-world perception problems. Previous work has documented how human-like behavior emerges in deep artificial neural networks optimized for natural tasks [12]–[15], [17], [18], [123]. In this work, we show how comparing the behavior of networks optimized to perform tasks using different neural representations can reveal which aspects of neural coding underlie human behavior.

Neurons transmit information in the precise timing of their spikes [124] in addition to

[1] *Natural tasks with natural stimuli* refer to tasks that operationalize perceptual judgments humans make in their daily lives. Examples in hearing include recognizing and localizing everyday sounds in noisy or reverberant environments. We use this terminology to distinguish from less ecological tasks with artificial stimuli that humans only encounter in perceptual experiments, such as discriminating simple synthetic tones and noise bursts. Note, it is possible to have artificial tasks with natural stimuli (e.g., counting phonemes in natural speech) and natural tasks with artificial stimuli (e.g., recognizing words in artificially distorted speech).

their time-averaged firing rates. Temporal coding has been identified across multiple sensory modalities [125]–[128], but spike timing is arguably most precise in the auditory nerve, where action potentials align to the temporal structure of sound. While the temporal envelopes of sound waveforms may be represented with changes in time-averaged firing rates, information in the temporal fine structure (i.e., the individual pressure oscillations) can only be encoded with precise spike timing. As mammalian auditory nerve fibers phase-lock to sound frequencies as high as 3 to 5 kHz [30]–[32], it is clear the auditory system has access to this information from the outset. However, the perceptual role of this information remains debated [33]. The precision of temporal coding decays with each synapse along the ascending auditory pathway, such that physiological mechanisms for extracting information from high-frequency phase locking are likely to be situated early. Despite longstanding interest, no such mechanisms have been discovered [27], [33]. Without a mechanism for reformatting information encoded in peripheral spike timing, it is unclear whether later stages of the auditory system can use this information to mediate behavior.

Information encoded in high-frequency phase locking is widely suspected to be important for sound localization, which relies on microsecond-level timing differences between the two ears. Behavioral evidence is argued to suggest that human pitch and speech perception also rely on cues in the temporal fine structure of sound [35], but, because the auditory nerve cannot be directly observed and manipulated in humans, the perceptual role of high-frequency phase locking remains controversial [28]. Nonetheless, impaired temporal coding is often proposed to underlie speech perception difficulties in people with hearing loss and cochlear implants, particularly in noise [34], [36], [129].

Previous computational evidence in support of a perceptual role comes from ideal observers for simple tasks with artificial stimuli (e.g., discrimination of single frequencies), where incorporating spike timing information sometimes leads to better qualitative fits to human behavior [10]. However, these models generally overestimate human performance [9], plausibly because human perceptual systems were never optimized for psychophysical tasks with artificial stimuli. We investigated the issue in models optimized for ecological tasks by training deep artificial neural networks to perform real-world hearing tasks using simulated auditory nerve representations as input. To ask whether the information encoded in phase-locked auditory nerve spike timing is necessary to account for human behavior, we separately optimized networks with altered phase locking and compared the behavior of trained networks to that of human listeners (Fig. 3.1a). The results provide new evidence for the importance of high-fidelity temporal coding in perception, but indicate that it is implicated in some aspects of perception (sound localization, pitch perception, and voice recognition) more than others (speech recognition), even in noise.

3.2 Results

3.2.1 Auditory nerve model stage

We hard-coded the model input representation to approximate the information the ear sends to the brain. We used a phenomenological model [5] of the auditory periphery to simulate instantaneous firing rate responses of a population of auditory nerve fibers whose frequency tuning and sensitivity was intended to match that of the human ear. Each stimulus was represented as a three-dimensional array of firing rates with shape [N frequency channels, T timesteps, S fiber types]. Due to computational constraints, we limited N to 50 or 100 frequency channels. The number of timesteps T depended on the length of the stimulus (always sampled at 10 or 20 kHz). We used S = 3 to simulate the 3 canonical auditory nerve fiber types which have different spiking thresholds and spontaneous activity [130]. High spontaneous rate fibers have very low thresholds but narrow dynamic ranges such that their firing rates quickly saturate at conversational sound levels. Medium and low spontaneous rate fibers have higher thresholds and broader dynamic ranges but are less numerous in the ear. This frequency-by-time-by-fiber-type array of instantaneous firing rates was then converted to an array of sampled spike counts, representing the population response of 32000 individual auditory nerve fibers per ear (60% high spontaneous rate, 25% medium spontaneous rate, and 15% low spontaneous rate), which served as the input representation to the networks. To our knowledge, our models are the first to perform naturalistic tasks using a near-realistic representation of the information from a sensory receptor organ.

3.2.2 Temporal coding manipulation

The fidelity of temporal coding in the mammalian ear is limited by ion channels in the hair cell membrane, which act as a low-pass filter [32]. While not feasible to manipulate in vivo (particularly in an animal with complex auditory behavior), the upper limit of phase locking can be altered in silico by changing the cutoff frequency of the low-pass filter governing the inner hair cell potential in the peripheral model. We optimized machine models with different cutoffs to ask whether phase locking was necessary to obtain humanlike behavior.

In one training condition, this low-pass cutoff was set to a default value of 3000 Hz, which causes a roll-off in auditory nerve phase locking on par with electrophysiological recordings made in non-human animals. This upper limit is presumed to be shared by humans but is not directly measurable [51], [52], [131]. To investigate the behavioral relevance of temporal coding, we lowered this cutoff to 1000 Hz (eliminating phase locking to high-frequency temporal fine structure), 320 Hz (eliminating phase locking above the fundamental frequency of most

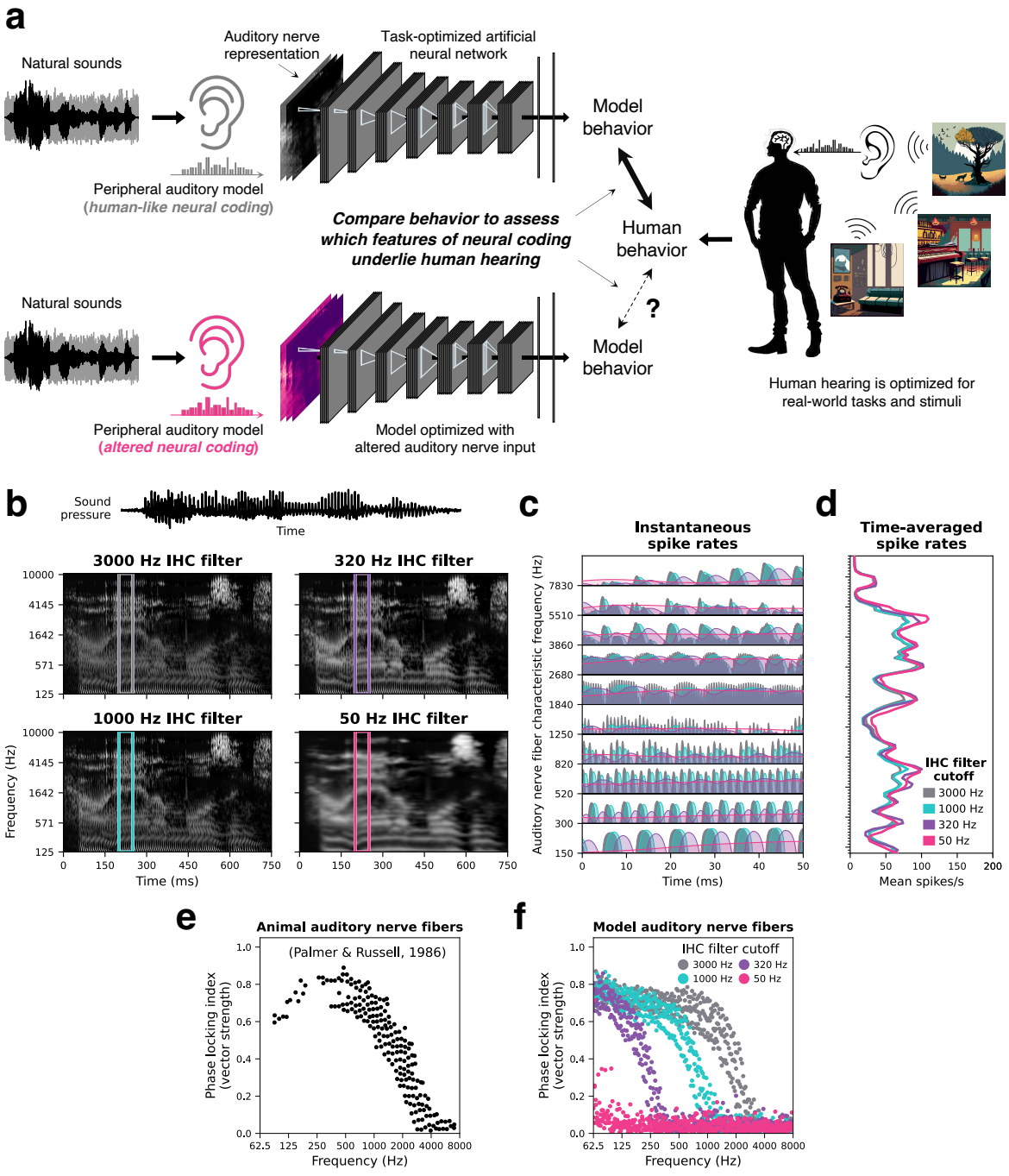

Figure 3.1: Overview.

a. Schematic of the approach. Human hearing behavior is shaped by the ears and the acoustic environment. Models optimized to perform naturalistic tasks with human-like auditory nerve input exhibit human-like behavior. Models optimized with altered auditory nerve input may reveal which features of neural coding underlie human behavior. b. Simulated auditory nerve representations of the same speech waveform with four different configurations of the auditory nerve model. Configurations differed in the inner hair cell low-pass filter cutoff which determines the upper frequency limit of phase locking. The 3000 Hz cutoff is commonly used to model the human auditory system. c. Instantaneous firing rates from example auditory nerve fibers illustrate the degradation of precise spike timing as the phase locking limit is lowered. d. Time-averaged firing rates (across the 50ms window depicted in c) illustrate that lowering the phase locking limit does not disrupt place cues in the overall pattern of excitation across the cochlear frequency axis. e. The roll-off in phase locking strength as a function of frequency has been measured in the auditory nerve fibers of many nonhuman mammals. Classic data in guinea pigs is re-plotted from Palmer and Russel (1986). f. The roll-off in phase locking strength as a function of frequency is plotted for model auditory nerve fibers with four different phase locking limits. The vector strength of phase locking was calculated for a simulated high, medium, and low-spontaneous rate auditory nerve fiber at each frequency and inner hair cell low-pass filter cutoff.

human speech pitch), or 50 Hz (eliminating all phase locking to temporal fine structure). Lowering this cutoff degrades the fidelity of temporal coding, progressively blurring the auditory nerve representation along the time-axis (Fig. 3.1b and c). However, the manipulation had very little effect on the pattern of firing across the cochlear frequency axis (Fig. 3.1d): nerve fibers with very low phase locking limits still encode high frequency sounds in their firing rates, just not with precise spike timing. We separately optimized neural networks operating on auditory nerve representations with these four different cutoff frequencies.

3.2.3 Simplified cochlear model stage

State-of-the-art cochlear models that best capture the nonlinear response properties of the auditory nerve are computationally expensive, which can limit their integration into larger-scale models of the auditory system. To investigate whether the fine-grained details of these models are critical to account for naturalistic behavior, we also optimized networks with a simplified cochlear front-end. The simplified front-end consisted of a linear cochlear filter bank followed by half-wave rectification and low-pass filtering (to impose an upper limit on phase locking), the output of which was passed through sigmoid functions approximating the rate-level functions of high-, medium-, and low-spontaneous-rate fibers [5]. These stages yielded a three-dimensional array of instantaneous auditory nerve firing rates with the same dimensions as the detailed auditory nerve model. The spike sampling procedure was identical for the simplified and detailed cochlear models. We repeated the temporal coding manipulation in the simplified cochlear model (by setting the low-pass filter cutoff to 3000, 1000, 320, and 50 Hz). In addition to testing the importance of a detailed cochlear model, the simplified model also served to rule out the possibility that effects observed with the detailed cochlear model were driven by unintended nonlinear consequences of adjusting filter parameters rather than degraded temporal coding per se.

3.2.4 Artificial neural network model stages and training

The neural network portion of each model consisted of a feedforward series of stages instantiating linear convolution, nonlinear rectification, normalization, and pooling. The parameters of these model stages were optimized to perform auditory tasks via supervised machine learning. Each task was operationalized as a classification problem with a single ground-truth label per stimulus. Training stimuli were compiled from large-scale corpora of natural sounds and were meant to approximate the "auditory diet" that likely shaped biological hearing systems over the course of evolution and development.

The performance of a neural network depends on both the weights directly optimized

for training task performance and the hyperparameters that define the network architecture (e.g., the number of layers and the size and shape of convolutional filter kernels). To ensure these hyperparameters were also optimized for the tasks, we used the top 10 best-performing network architectures previously identified in large-scale random architecture searches conducted for each task (Supplementary Tables 3.1, 3.2, and 3.3) [17], [18], [37]. Results for each task and cochlear model configuration are presented as the average of 10 network architectures, allowing us to provide uncertainty estimates and marginalize across the idiosyncrasies of any single network architecture.

3.2.5 Model tasks and roadmap

Models were optimized to perform four different auditory tasks: sound localization, pitch perception, voice recognition, and word recognition. For each task, we separately trained models with four different auditory nerve phase locking limits and then compared their behavior to that of humans. In the following sections, we consider each task in turn. In each case, we first compare humans and models tested on naturalistic stimuli in noise, asking whether phase-locked spike timing is needed to obtain human-level performance. We then simulate a battery of psychoacoustic experiments to investigate whether specific characteristics of human behavior depend on phase locking. We then quantify overall human-model similarity for each phase locking condition by measuring correlations between analogous behavioral data points across all experiments and phase locking limits.

We previously found that many characteristics of human behavior emerge in artificial neural networks optimized under naturalistic conditions [17], [18]. The models featured here built on these results, but were improved in several respects to provide a strong test of the importance of temporal coding. Specifically, they operated on more realistic input representations (incorporating spikes and multiple types of auditory nerve fibers), were trained on more realistic datasets, and were evaluated with an expanded set of psychoacoustic experiments. The phase locking manipulation that is the central focus of this work was first introduced in our previous pitch modeling work [17]. For the voice and word recognition tasks, only the base network architecture and training dataset are borrowed from previous work [18].

If only models with high phase locking limits can account for human behavior, this would provide evidence that precise auditory nerve spike timing contributes to perception and thus physiological mechanisms for extracting it must exist. Such a result was a priori plausible for sound localization, where microsecond-level timing differences between the two ears can plausibly only be transduced via spike timing. However, it was unclear how any deficits from

impaired temporal coding should translate to sound localization behavior. It was also unclear what to expect for pitch and speech perception. We previously found that pitch perception characteristics were dependent on phase locking in a simpler model with less realistic nerve fiber responses, and here sought to test the issue more definitively. It was likewise unclear whether voice and speech recognition could be accounted for without temporal coding.

3.2.6 Model optimization – sound localization

To assess sound localization behavior, models were tasked with reporting the location of target sound sources in naturalistic auditory scenes rendered as binaural audio with a virtual acoustic room and head simulator (Fig. 3.2a). There were 1.8 million training scenes (rendered in 1800 unique rooms of varying size, wall materials, and reverberation), each consisting of a target source rendered at a single location in the presence of diffusely localized texture-like background noise. The model's task was to classify scenes according to the azimuth and elevation of the target source relative to the simulated listener's head (504-way classification into 5° azimuth by 10° elevation bins). The model operated on the auditory nerve responses from the simulated listener's left and right ears, and thus had access to the same monaural and binaural cues as a human listener in the same scene (Fig. 3.2b). Models optimized for this task with access to high-fidelity temporal coding in the peripheral representation have previously been shown to replicate characteristics of human sound localization [18], including the frequency-dependent use of interaural time and level differences for azimuth judgments [132] and the use of ear-specific spectral cues for elevation judgments [133], [134].

3.2.7 Degraded temporal coding impairs sound localization

We compared human and model sound localization accuracy for a set of 460 natural sounds presented in different levels of background noise (Fig. 3.2c). Humans were asked to report which of 95 loudspeakers (in a 19-by-5 array spanning -90° to 90° azimuth and 0° to 40° elevation in steps of 10°) produced the target sound. On each trial threshold equalizing noise [136] played diffusely from 9 other randomly selected speaker locations. Models performed the same task in a virtual rendering of the loudspeaker array room. Overall task performance was quantified with mean absolute localization error as a function of SNR. Although human listeners outperformed all models at the very lowest SNR conditions, models with access to high-frequency phase locking in their auditory nerve input produced the best match to human behavior (Fig. 3.2d). Models with 3000 and 1000 Hz phase locking limits exhibited near-human-level robustness to noise, while models with degraded temporal coding made

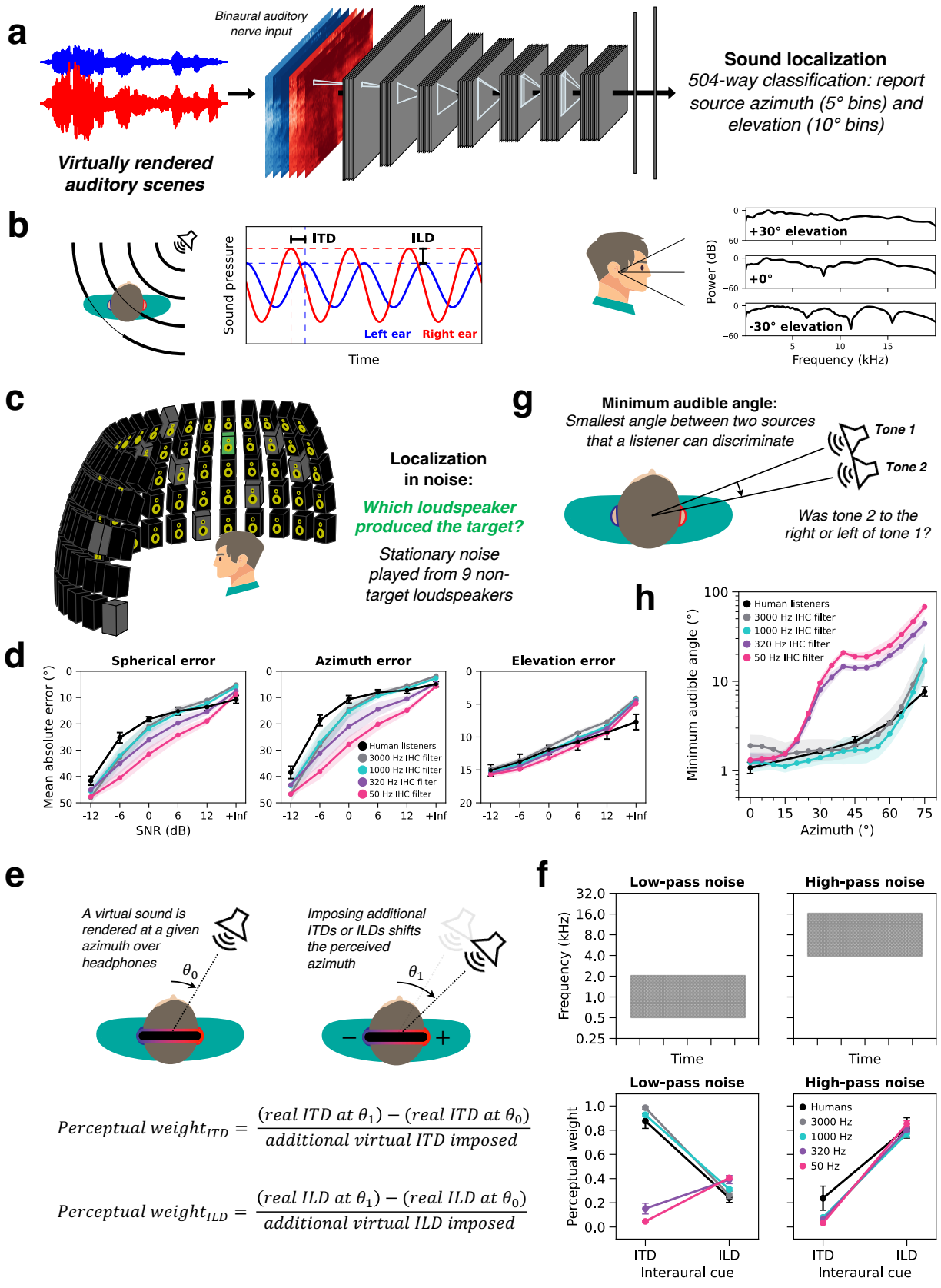

Figure 3.2: Sound localization is impaired in models with degraded auditory nerve spike timing.

a. Localization model schematic. Deep artificial neural networks optimized for sound localization operated on binaural auditory nerve representations of virtually rendered auditory scenes. b. Sound localization cues available to human listeners. Left: interaural time and level differences (ITDs and ILDs) are schematized with pure tones recorded at the left and right ear. Right: spectral differences in the anatomical transfer function provide a monaural cue to elevation. c. Schematic of the sound localization in noise experiment. d. Mean absolute localization error for humans and models localizing natural sounds in noise are plotted as a function of SNR. The three axes separately plot spherical, azimuth, and elevation errors. e. Schematic of the ITD / ILD cue weighting experiment. Computed perceptual weights indicate the extent to which imposing additional ITDs or ILDs shifts the perceived azimuth of a virtual sound presented over headphones. f. ITD and ILD perceptual weights measured with low-pass and high-pass noise from human and model listeners. g. Schematic of minimum audible angle experiment. h. Minimum audible angles were measured as a function of azimuth from human and model listeners. Model error bars always indicate ±2 standard errors of the mean across 10 network architectures per phase locking condition. In d and f, human error bars indicate ±2 standard errors of the mean across participants. In h, human error bars indicate ±2 standard errors from 1 listener averaged across 4 different pure tone frequencies. Human data in f and h are re-plotted from the original studies[132], [135].

progressively larger localization errors as the phase locking limit was lowered. Degraded temporal coding impaired localization performance in both azimuth and elevation though the effect was larger for azimuth (Fig. 3.2d, middle vs. right). These results show that precise spike timing is important for sound localization.

3.2.8 Sound localization psychoacoustics

Biological sound localization relies on three main cues. Small time and level differences between the sound at the ears are two important binaural cues to a source's location in the azimuthal plane (Fig. 3.2b, left). Before impinging on the ear drum, a sound waveform is transformed by the pinna, head, and torso, which boost some frequencies and attenuate others. This anatomical filtering is direction-specific and provides a third spectral cue to a source's location (Fig. 3.2b, right). Humans rely on these spectral cues to judge elevation [137], [138]. To investigate the contribution of temporal coding to each of these localization cues, we simulated a set of classic psychoacoustic experiments on the models.

3.2.9 Auditory nerve phase locking is critical for ITD-based sound localization

Localization in the horizontal plane relies on interaural time differences (ITDs) and interaural level differences (ILDs), but these cues are not equally informative for all sounds. Humans rely more on ITDs at low frequencies and ILDs at high frequencies. First proposed [139] as 'duplex theory' in 1907, measurements of human sensitivity to interaural cue manipulations with virtual sounds provided an elegant modern demonstration of this frequency-specific cue dependence [132] (Fig. 3.2e). In the original experiment, sounds were rendered at different azimuths using a virtual acoustic simulator. Interaural cue sensitivity was revealed by how much a sound's perceived location appeared to shift as additional ITDs or ILDs were added to the binaural waveforms. Shifts in perceived azimuth were mapped back to units

of ITD or ILD (as the ITD or ILD change corresponding to an actual shift in azimuth by the same amount), allowing interaural cue sensitivity to be quantified as a dimensionless weight: the slope of the response cue relative to the imposed cue. For low frequency sounds, ITD manipulations have a much larger effect than ILD manipulations on human azimuth judgments (ITD slope is much larger than ILD slope). The reverse is true for high frequency sounds.

Although the encoding of ITDs is thought to make use of phase locking, it was a priori not entirely clear what to expect from the models with altered phase locking limits. The ITDs of natural sounds are present in amplitude envelopes in addition to the fine structure within frequency channels [140]–[142], and because the amplitude envelope is lowpass, interaural envelope delays should in principle be detectable even if the effective sampling rate of cochlear transduction is lowered via the phase locking limit.

To investigate the contribution of phase locking to this frequency-specific cue dependence, we simulated this experiment on our models, measuring interaural cue sensitivity of models with different phase locking limits (Fig. 3.2f). Models with high phase locking limits replicated human behavior, exhibiting high ITD sensitivity only for low frequencies and high ILD sensitivity only for low frequencies. Models with degraded temporal coding (320 and 50 Hz phase locking limits) deviated from human behavior, progressively losing ITD sensitivity at all frequencies and gaining ILD sensitivity at low frequencies, suggesting phase-locked spike timing up to 1000 Hz is necessary for ITD-based sound localization. Models without access to fine timing information become increasingly reliant on ILD-based sound localization, exhibiting superhuman ILD-sensitivity for low frequencies.

3.2.10 Azimuth dependence of human localization requires phase locking

The inhuman cue dependence under degraded phase locking was also strikingly evident when measuring localization acuity as a function of azimuth. Human sound localization is best near the midline and becomes less accurate toward the periphery [143]–[145], as can be quantified by minimum audible angle thresholds [135] (the smallest detectable angular distance between two sources) (Fig. 3.2g). We simulated an experiment measuring minimum audible angle thresholds for pure tones. Thresholds measured from the 3000 and 1000 Hz phase locking models resembled those of human listeners (Fig. 3.2h). By contrast, the 320 and 50 Hz phase locking models exhibited a qualitatively different, nonmonotonic dependence on azimuth, yielding significantly higher minimum audible angle thresholds away from the midline. These results suggest that ITD cues conveyed by precise spike timing are particularly important

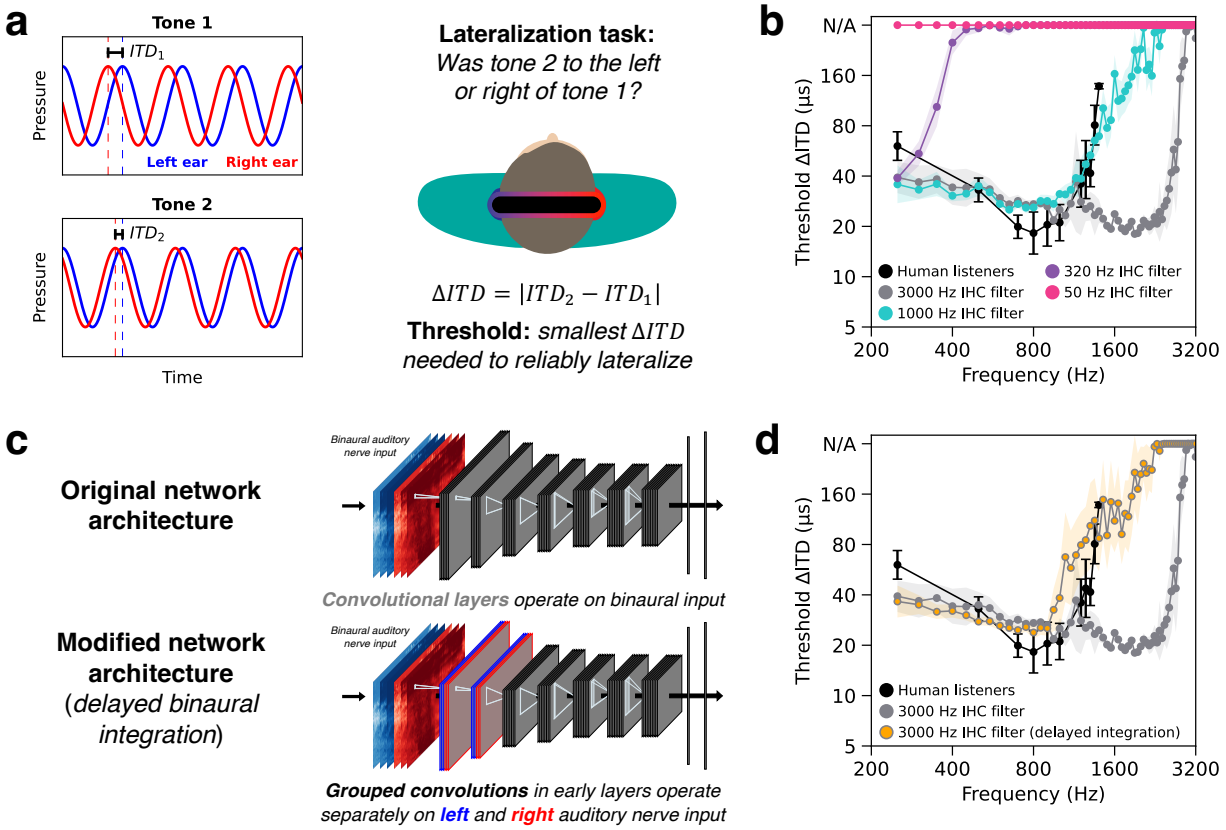

Figure 3.3: Upper frequency limit of ITD sensitivity.

a. Schematic of ITD lateralization experiment used to measure ITD sensitivity as a function of frequency. On each trial, listeners heard a pair of pure tones with two different ITDs and judged whether the second tone sounded to the right or left of the first. b. ITD lateralization thresholds measured as a function of frequency from humans and models. c. Schematic of network architecture modification to delay binaural integration. Replacing the first couple convolutional layers with grouped convolutions (1 group for each ear) forces networks to process the ears separately before binaural integration occurs in the first standard convolutional layer. Blue and red represent information from the left and right ears, respectively. d. ITD lateralization thresholds measured as a function of frequency from humans and models with and without the modified network architectures (both models had the same human-like phase locking limit). Error bars indicate ±2 standard errors of the mean across human participants or network architectures. Human data are re-plotted from the original study[146].

for accurate localization away from the midline.

3.2.11 Human ITD sensitivity is not limited only by auditory nerve phase locking

Human listeners are remarkably sensitive to ITDs at low frequencies, but this sensitivity deteriorates at higher frequencies [147]. In principle this sensitivity could be limited by the upper limit of phase locking in the auditory nerve, but human sensitivity instead declines rapidly above 1 kHz and is fully lost by 1.5 kHz [146] – well below the 3-5 kHz presumptive frequency limit of auditory nerve phase locking. To better understand this discrepancy, we studied the frequency limits of ITD sensitivity in our models.

ITD sensitivity as a function of frequency has been characterized by measuring ΔITD thresholds with pure tones (Fig. 3.3a). In such experiments, listeners judge which of two spatialized tones with different ITDs sounds further to the right. The ΔITD threshold is the smallest change in ITD needed to reliably discriminate tones by azimuth. We simulated an azimuth discrimination experiment with pure tones to measure network ΔITD thresholds as a function of frequency. Model ΔITD thresholds were unmeasurably high for frequencies above a model's respective phase locking limits (Fig. 3.3b). Thresholds measured from the 1000 Hz phase locking network produced the closest match to human behavior. The 3000 Hz phase locking model exhibited superhuman ITD sensitivity, with thresholds on the order of 20μs even up to 2.5 kHz. This discrepancy with humans suggests that some additional factor limits human ITD sensitivity.

Because input from the two ears needs to be compared to extract ITDs, perceptual sensitivity to high-frequency ITDs requires temporal coding at that frequency to persist until information from the two ears can be compared [148], [149]. One explanation for the lower limit of ITD sensitivity in humans is that there are anatomical constraints on this comparison process, whereby information from each ear passes through additional synapses before being combined, with some loss of temporal precision at each synapse. Why would the auditory system not have evolved a way to make the comparison happen earlier? When tested on naturalistic auditory scenes (natural sounds in noise), the 1000 Hz phase locking model localizes just as well as the 3000 Hz model. And across all other psychoacoustic experiments we simulated, there was no significant difference in human-model similarity between the 1000 and 3000 Hz phase locking models. These results suggest that preserving temporal coding above 1000 Hz provides little adaptive benefit, such that delaying the comparison incurs little cost.

To test the idea that "early" interaural comparisons accounted for our model's superhuman ITD sensitivity, we altered the network architectures slightly to delay binaural integration (Fig. 3.3c). We replaced the standard convolution operations in the earliest neural network stages with grouped convolution operations (2 groups), such that the resulting models must initially process information from the left and right auditory nerve separately. Reasoning that synapses introduce temporal jitter that effectively imposes low-pass filtering [150], binaural integration was only allowed to occur in the models after early temporal pooling layers. We trained these modified neural network architectures with 3000 Hz phase locking auditory nerve input and evaluated them on the full set of sound localization experiments. Consistent with our hypothesis, the models lost sensitivity to high-frequency ITDs (Fig. 3.3d) but were otherwise unaffected. These results indicate that additional physiological constraints can in some cases produce better matches to human behavior. And the model

results provide a normative explanation for the solution that biological auditory systems have arrived at over evolution.

3.2.12 Models without access to phase locking exhibit inhuman spectral cue dependence

Whereas localization in azimuth is dominated by binaural cues, localization in elevation is mediated in large part by spectral cues. Humans rely on cues specific to their own ears to make accurate elevation judgments (Fig. 3.4a). Manipulating these spectral cues (either physically by altering pinna shape with an ear mold [134] or virtually by altering the head-related transfer function used to spatialize a sound [133] via earphones) impairs elevation judgments in humans. These same manipulations have minimal effects on azimuth judgments. We tested whether this spectral cue dependence relies on fine timing information in the ear by simulating these experiments on models optimized with different phase locking cutoffs. Our models were always trained using a set of head-related transfer functions (HRTFs) measured from a standard model of the human head and torso [151] (KEMAR). We evaluated the models with virtual sounds rendered with altered HRTFs. We tested each model on 45 alternative sets of ears (HRTFs measured from 45 different people) as well on spectrally smoothed versions of the KEMAR HRTFs used to train models.

A priori it was not clear to expect. For instance, it seemed plausible that phase locking might help to obtain precise estimates of the spectral shape that underlies ear-specific localization cues, such that localization in elevation would be impaired at lower phase locking limits.

When tested with the alternative ears, elevation judgments collapsed in all models, as in human listeners with molds in their ears, indicating that spectral cues to elevation are not strongly dependent on phase locking (Fig. 3.4b). However, the effect of alternative ears on azimuth varied depending on the phase locking cutoff. In human listeners and in models with high-fidelity temporal coding, changing ears had little effect on mean azimuth error. But in models with degraded temporal coding, azimuthal localization accuracy was worse with alternative ears (Fig. 3.4c). As the phase locking limit was lowered, models became increasingly reliant on ear-specific cues to localize in azimuth as well as elevation. These results suggest that human-like dependence on ear-specific cues (i.e., only for elevation) emerges only when models have access to phase-locked spike timing.

This inhuman dependence of azimuthal localization on spectral cues was also evident in the effects of removing fine spectral details from the trained HRTFs (Fig. 3.4d). As the peaks and valleys of the trained HRTFs were parametrically smoothed away, model elevation

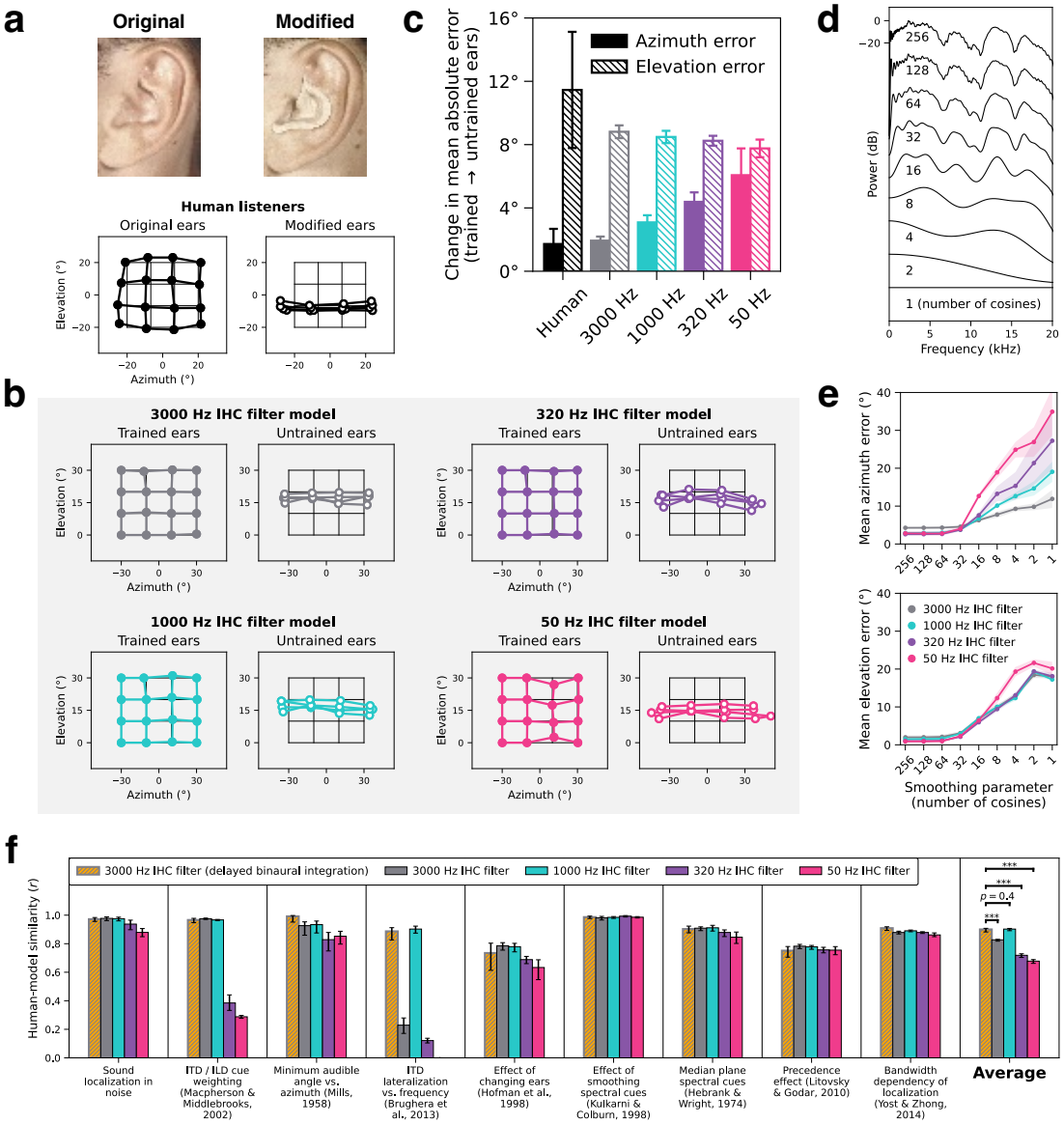

Figure 3.4: Localization models with degraded phase locking rely on spectral cues to judge azimuth as well as elevation.

a. Photographs and human data from the original study[134]. Hofman et al. (1998) measured human sound localization before and after inserting plastic molds into participants' ears to change the pinnae's direction-specific filtering. Sound localization judgments (thick lines, circle markers) with the participants original (left) and modified (ears) are plotted as a function of azimuth and elevation, superimposed on a grid of the true locations (thin lines, no markers). b. Model results for an analogous experiment, in which networks were evaluated on sounds rendered with the HRTFs used for training (trained ears) and a different set of HRTFs (untrained ears), are plotted with the same format for each of the phase locking conditions. c. The increase in mean absolute azimuth and elevation error due to changing ears is plotted side-by-side for humans and for each model. Error bars indicate ±2 standard errors of the mean across participants or network architectures. d. Power spectra of HRTFs progressively smoothed by lowering the number of cosines used to approximate the discrete cosine transform. e. Effect of the spectral smoothing manipulation on model azimuth and elevation judgments. Mean absolute azimuth (top) and elevation (bottom) errors are plotted as a function of the HRTF smoothing parameter used to render stimuli for the models. Error bars indicate ±2 standard errors of the mean across network architectures. f. Human-model similarity scores for each phase locking condition and localization experiment. Similarity scores were Pearson correlation coefficients between corresponding human and model behavioral data points. The rightmost column averages scores across all experiments. Error bars indicate 95% confidence intervals computed by bootstrapping the mean of 10 network architectures. Asterisks denote statistically significant differences between phase locking conditions ($p < 0.001$, two-tailed), evaluated by comparing mean similarity scores against a null distribution bootstrapped from the 3000 Hz condition with delayed binaural integration.

judgments progressively collapse, regardless of phase locking limit, as expected (Fig. 3.4e). By contrast, azimuth judgments are significantly more impaired by HRTF smoothing in models with lower phase locking limits, suggesting models without access to phase locking rely on fine spectral details to localize in azimuth as well as elevation (unlike humans).

3.2.13 Not all localization phenomena are inextricably linked to phase-locked spike timing

Although removing phase locking caused pronounced discrepancies with human behavior, some behaviors were relatively unaffected. All models exhibited the "precedence effect", in which localization judgments are dominated by the initial part of a sound [152] (Supplementary Fig. 3.8g). Our results suggest this effect – which likely reflects a strategy for echo-robust localization in reverberant environments [18] – does not rely on precise spike-timing. All models also exhibited human-like dependences of localization accuracy on bandwidth [153] (Supplementary Fig. 3.8h) and of elevation accuracy on high-frequency spectral cues [154] (Supplementary Fig. 3.8i).

3.2.14 Quantitative measures of human-model similarity – sound localization

To quantify human-model behavioral similarity on each localization experiment, we measured correlations between human and model results graphs (Fig. 3.4f). Of models without modified network architectures, the 1000 Hz phase locking model produced the most human-like behavior. The overall human-model similarity was lower for the 3000 Hz phase locking model because of the resulting superhuman ITD sensitivity at high frequencies. Simply modifying the network architectures to delay binaural integration in the 3000 Hz phase locking model was sufficient to restore human-model similarity to the level of the 1000 Hz phase locking model. Collectively, the results suggest that human-like sound localization relies on information conveyed by phase-locked spike timing up to, but not much beyond, 1000 Hz.

3.2.15 Model optimization – pitch perception

We modeled pitch perception as the estimation of fundamental frequency (F0). Each training stimulus consisted of a short (50ms) speech or music excerpt (selected to be periodic with a well-defined F0) embedded in aperiodic background noise taken from YouTube soundtracks (Fig. 3.5a). Models were tasked with classifying 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

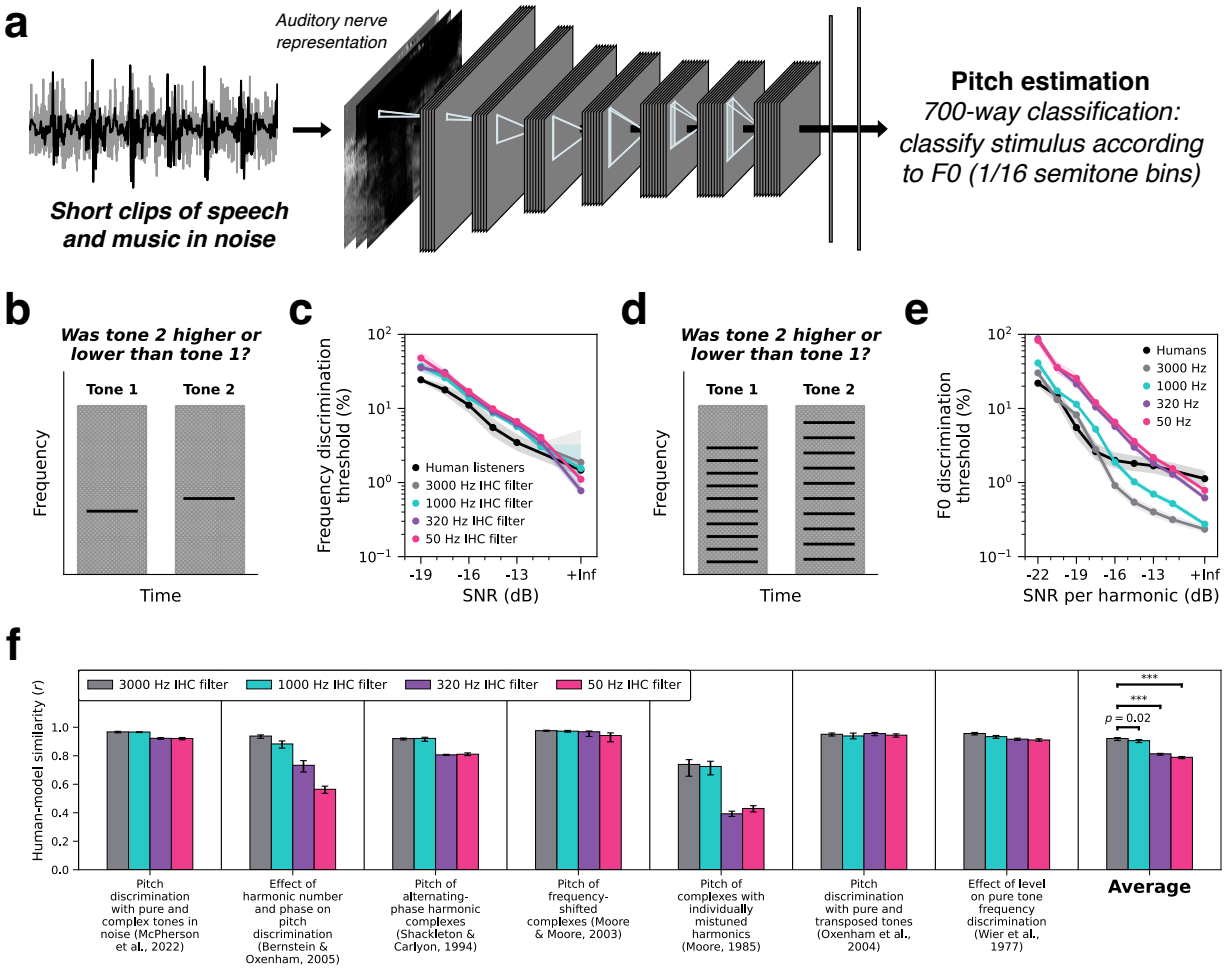

Figure 3.5: Models with degraded auditory nerve spike-timing exhibit less human-like pitch perception.

a. Pitch model schematic. Deep artificial neural networks were optimized to estimate F0 from simulated auditory nerve representations of natural sounds. b. Stimuli for frequency discrimination in noise task. Human and model listeners discriminated pairs of pure tones with similar frequencies presented in masking noise. c. Frequency discrimination thresholds measured from humans and models are plotted as a function of SNR. d. Stimuli for pitch discrimination in noise task. Human and model listeners discriminated pairs of 10-harmonic complex tones with similar F0s presented in masking noise. e. F0 discrimination thresholds measured from humans and models are plotted as a function of SNR. Error bars in c and e indicate ±2 standard errors of the mean across human participants or network architectures. Human data are re-plotted from the original study[89]. f. Human-model similarity scores for each phase locking condition and pitch experiment. Similarity scores were Pearson correlation coefficients between corresponding human and model behavioral data points. The rightmost column averages scores across all experiments. Error bars indicate 95% confidence intervals computed by bootstrapping the mean of 10 network architectures. Asterisks denote statistically significant differences between phase locking conditions ($p < 0.001$, two-tailed), evaluated by comparing mean similarity scores against a null distribution bootstrapped from the 3000 Hz condition.

previously found that models optimized for this task replicated many characteristics of human perception, but only when optimized for naturalistic stimuli with high-fidelity temporal coding in the input [17]. Here we replicate and extend these previous findings with models optimized with more realistic auditory nerve input, simulating the full set of psychoacoustic experiments from our previous work and a new experiment comparing human and model pitch discrimination in noise.

3.2.16 Human-like pitch perception depends critically on auditory nerve phase locking

To compare the pitch-related behavior of models and human listeners, we simulated a set of six experiments used in our previous work, along with measuring pitch discrimination thresholds in noise. The six original psychoacoustic experiments yielded results like those in previous work, with worse matches to human behavior in some experiments when the phase locking limit was lowered (F0 discrimination as a function of harmonic number and phase [63] (Supplementary Fig. 3.9a), pitch estimation of alternating-phase stimuli [60] (Supplementary Fig. 3.9b), pitch estimation of frequency-shifted complexes [61] (Supplementary Fig. 3.9c), pitch estimation of complexes with individually mistuned harmonics [59] (Supplementary Fig. 3.9d), frequency discrimination with pure and transposed tones [62] (Supplementary Fig. 3.9e), and pure tone frequency discrimination as a function of sound level [82] (Supplementary Fig. 3.9f). The present results show that this result holds even when a more realistic simulation of the auditory nerve is used.

We then measured network F0 discrimination thresholds (i.e., the minimum %F0 difference between two tones needed to reliably judge which of two tones had a higher F0) with pure and complex tones in noise (F0s log-uniformly distributed between 178 and 300 Hz), comparing them to previously reported human thresholds in the same conditions [89]. Model discrimination thresholds for pure tones (Fig. 3.5b) closely matched those of humans and showed little effect of phase locking limit across the SNRs tested (Fig. 3.5c). These results are consistent with the idea that human-like pure tone discrimination in this frequency range does not require phase-locked spike timing, likely because rate-place cues are sufficient for such simple stimuli. But for harmonic complex tones (Fig. 3.5d), model thresholds did benefit from phase locking, with only the 3000 and 1000 Hz phase locking models achieving human-level performance at low SNRs (Fig. 3.5e). We note that the 320 Hz model significantly underperformed humans and models with higher phase locking limits at low SNRs, even though the 320 Hz limit was above the highest F0 used in the experiment. This result suggests that human-level discrimination for complex tones in noise relies on temporal cues

conveyed by phase locking to harmonics above the fundamental. This finding is consistent with the effect of phase locking limit on the dependence on harmonic number, wherein models with lower phase locking limits were progressively more dependent on the first harmonic being present (unlike humans).

3.2.17 Quantitative measures of human-model similarity – pitch perception

We again quantified human-model behavioral similarity on each experiment by measuring correlations between analogous human and model data points (Fig. 3.5f). Collectively, the results suggest that human-like pitch perception relies on information conveyed by phase-locked spike timing, potentially as high as 1000 Hz ($p = 0.01$, $d = 2.61$ comparing 3000 vs. 1000 Hz conditions; $p < 0.001$, $d = 27.4$ and 33.2 comparing 3000 vs. 320 and 50 Hz conditions).

3.2.18 Model optimization – voice and word recognition

To model speech perception, we optimized models to recognize voices and words using the Word-Speaker-Noise dataset [93] (Fig. 3.6a). This dataset consists of 6 million 2-second speech excerpts superimposed on real-world background noise. Models were jointly optimized to classify stimuli according to the talker that produced the utterance (433-way voice recognition task) and the word that appeared in the middle of the excerpt (794-way word recognition task). The optimization objective was to simultaneously minimize voice and word classification error, and the two tasks shared all neural network stages up to the final linear read-out layers, which were task specific. We also trained models on each task individually and found similar results. We present results from the joint-task model here; results from the single-task models are shown in Supplementary Fig. 3.10.

3.2.19 Phase locking to the fundamental improves voice recognition in noise

We first measured model voice recognition performance in different types of background noise at SNRs ranging from -9 dB to Inf (quiet) (Fig. 3.6b). At low SNRs, models with access to phase locking performed better than models without. Almost all the benefit from phase locking incurred below 320 Hz, suggesting phase locking up to but not above the F0 of most human speech improves voice recognition in noise.

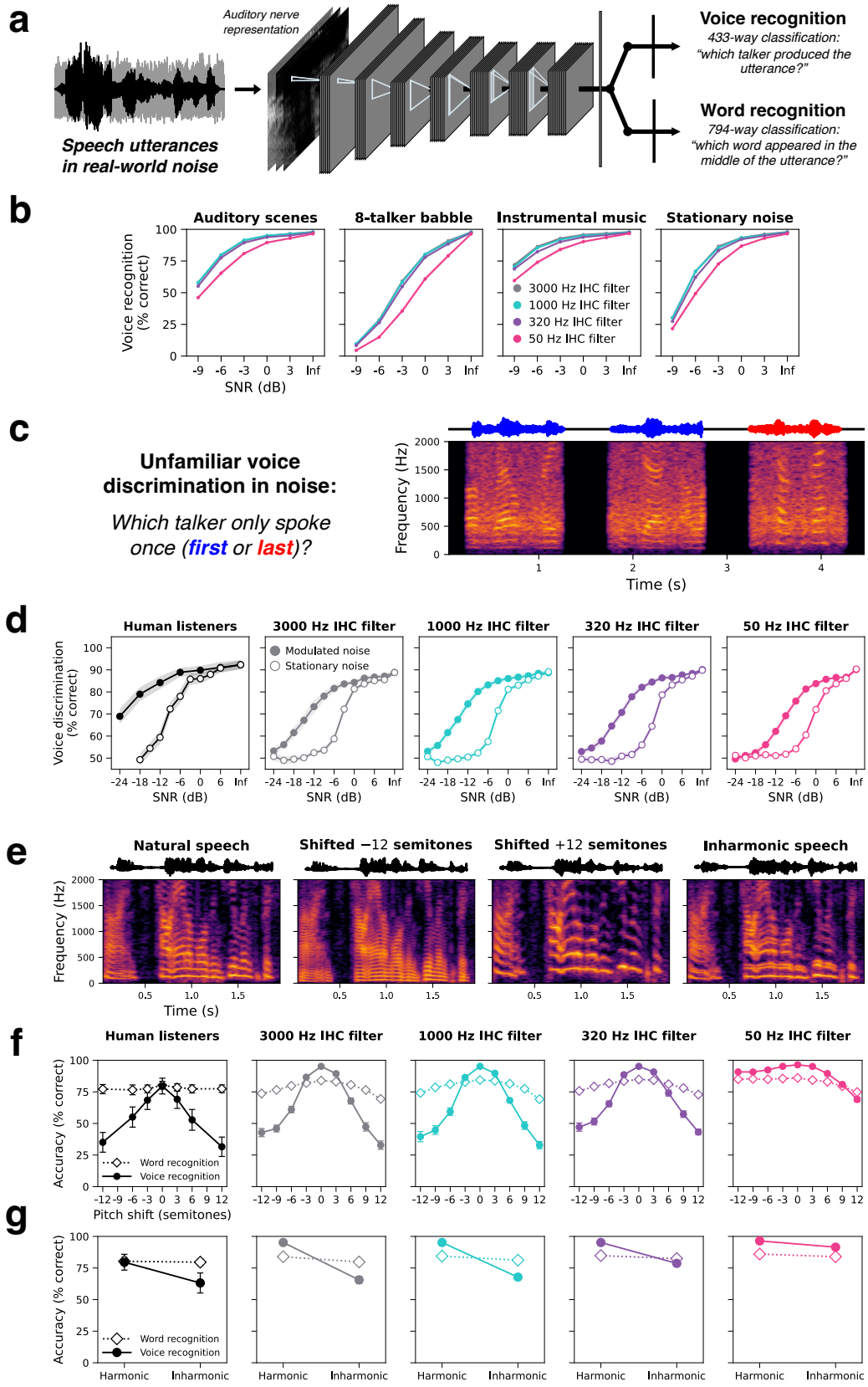

Figure 3.6: Models with degraded auditory nerve spike-timing exhibit less human-like voice recognition.

a. Speech model schematic. Deep artificial neural networks were jointly optimized to recognize voices and words from simulated auditory nerve representations of speech in noise. The two tasks shared all model stages before for the final task-specific output layers. b. Model voice recognition in noise. Different axes plot model performance as a function of SNR in different naturalistic noise conditions. c. Stimuli for the unfamiliar voice discrimination in noise task. Human and model listeners were played three 1s speech-in-noise excerpts from two unknown talkers. The task was to judge which talker only spoke once. d. Voice discrimination in noise results. Performance for human listeners and models with different phase locking limits is plotted as a function of SNR in stationary (open circles) and amplitude-modulated (closed circles) noise. e. Stimuli for pitch-altered voice and word recognition experiments. Spectrograms show the same speech clip in four different pitch conditions: unmodified (natural), pitch-shifted down 12 semitones, pitch-shifted up 12 semitones, and inharmonic. In the inharmonic condition, harmonic frequency components are randomly frequency-shifted such that they are no longer integer multiples of a common F0 and no longer linearly spaced in frequency. f. Voice and word recognition accuracy for humans and models tested on pitch-shifted speech. g. Voice and word recognition accuracy for humans and models tested on harmonic and inharmonic speech. All error bars indicate ±2 standard errors of the mean across human participants or network architectures.

To compare the models to humans we first considered absolute performance. Measuring human voice recognition is practically challenging because listeners do not all recognize the same voices and overall accuracy depends strongly on listener familiarity with voices. To compare human and model voice perception in noise, we instead simulated a voice discrimination experiment [58] (Fig. 3.6c). Participants heard three speech utterances from two different talkers and were tasked with identifying which talker (either first or last) only spoke once. We used speech excerpts from talkers that were unfamiliar to both human listeners and models (i.e., were not included in the training data). Model judgments were obtained by comparing the Kullback-Leibler divergence of the model's voice recognition output probability distributions for the first and second, and second and third voice excerpts, and taking the maximum. The experiment measured discrimination performance as a function of SNR in stationary and amplitude-modulated speech-shaped noise (Fig. 3.6d). Models, regardless of phase locking limit, achieved near human-level performance at high SNRs. At low SNRs, all models underperformed human listeners, but models with lower phase locking limits were especially impaired. Like humans, models with high phase locking limits performed substantially better when background noise was amplitude modulated (10 dB difference in threshold). This fluctuating masker benefit for voice recognition was smaller in the 50 Hz phase locking model (8 dB difference in threshold).

3.2.20 The dependence of human voice recognition on absolute pitch requires phase locking

We next considered whether phase locking would be critical for reproducing the pitch dependence of human voice recognition. Absolute pitch is an important cue humans use to recognize voices. When a familiar talker's voice is pitch-shifted or made inharmonic (by frequency jittering its harmonic components to be inconsistent with any single F0) (Fig. 3.6e), the voice is less recognizable [58]. To investigate whether this dependence on absolute pitch requires phase locking, we measured human and model voice recognition with pitch-shifted

(Fig. 3.6f, closed symbols) and inharmonic speech (Fig. 3.6g, closed symbols). Models were tested on familiar voices (but held-out speech utterances) from the training set. Humans were tested on celebrity voices [58]. Though we attempted to evaluate human performance only on celebrity voices that participants were highly familiar with (by including check trials and having participants identify which celebrities they expected to recognize by voice), it is difficult to match the relative familiarity of test voices between human and model participants, confounding comparisons of absolute performance between humans and models. Qualitatively, however, models with access to phase-locked spike timing best replicated human behavior. Human voice recognition was best for voices at their natural F0 and fell off with progressively larger shifts in either direction. Human performance was also impaired by making voices inharmonic. Networks with the 50 Hz phase locking limit exhibited superhuman robustness to these pitch manipulations, suggesting models without access to phase-locked spike timing do not rely on absolute pitch to identify voices as much as humans do. These results provide additional evidence for a role of phase locking (up to 320 Hz) in human voice recognition and pitch perception.

We also measured human and model word recognition with pitch-shifted and inharmonic speech (Fig. 3.6f and g, open symbols). Human performance was unaffected by these pitch manipulations. Model performance, regardless of phase-locking limit, was similarly robust, remaining comparable to humans for very large pitch shifts. These results suggest human word recognition in quiet is not critically dependent on either absolute pitch or phase-locked spike timing.

3.2.21 Phase locking offers little benefit for word recognition in real-world noise

Relative to model voice recognition, phase locking provided little benefit for word recognition, even in noise (Fig. 3.7a). Lowering the phase locking limit produced negligible deficits in model word recognition accuracy in any of four tested noise classes: recorded auditory scenes, speech babble, instrumental music, and stationary speech-shaped noise. Models performed comparably to humans tested on the same stimuli, regardless of phase locking limit. The one exception was amplitude-modulated speech-shaped noise, where the 50 Hz model was appreciably impaired and underperformed human listeners (Fig. 3.7b). Deficits in modulated noise are consistent with psychophysical evidence suggesting temporal fine structure cues help humans "listen in the dips" of fluctuating maskers [34], [36], [129]. Like humans, models with access to phase-locked spike timing performed better in amplitude-modulated than stationary speech-shaped noise. This fluctuating masker benefit was much smaller for the 50

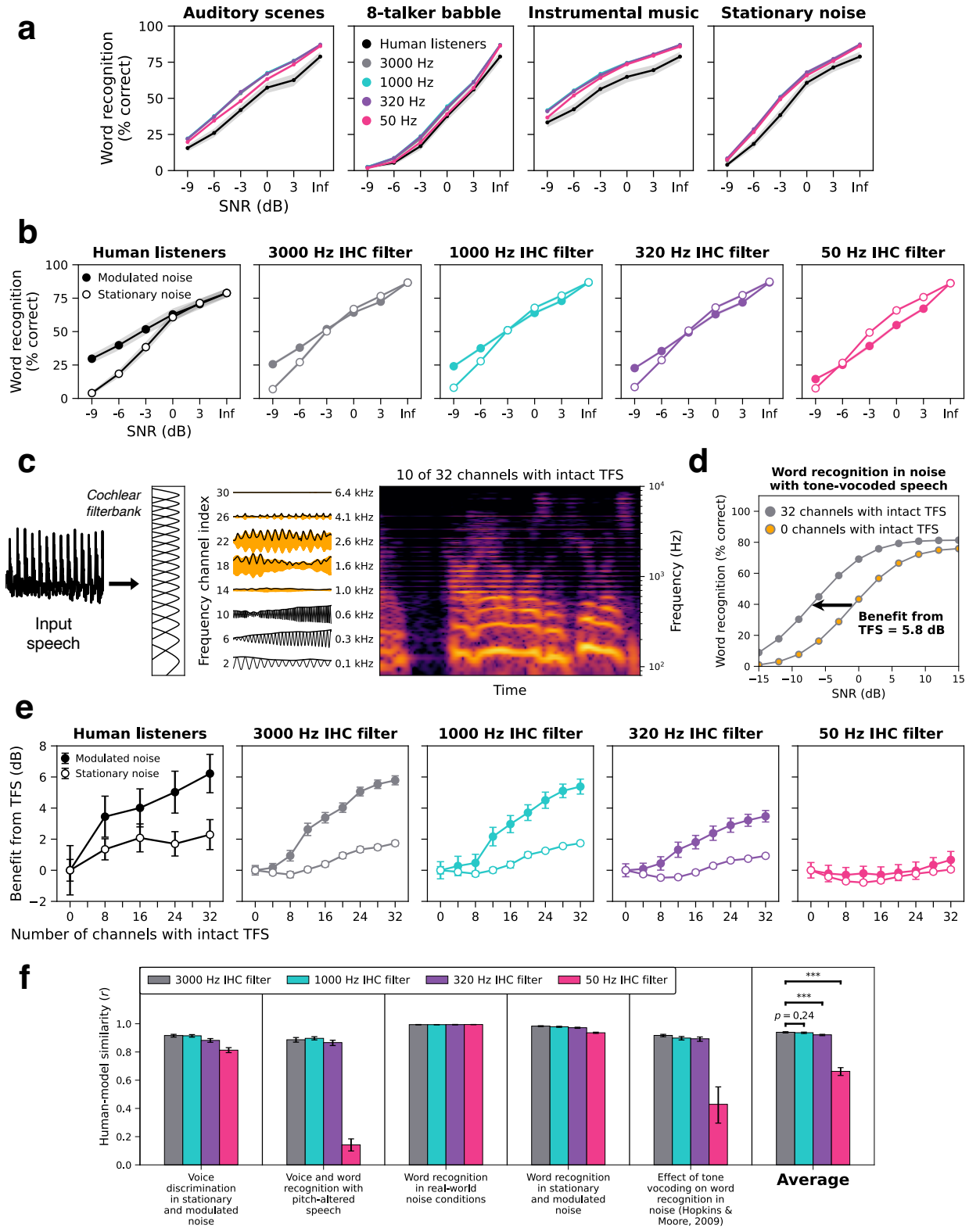

Figure 3.7: Models with degraded temporal coding achieve human-level word recognition in noise but fail to account for psychoacoustic phenomena.

a. Human and model word recognition in naturalistic noise. Different axes plot model performance as a function of SNR in different naturalistic noise conditions. b. Fluctuating masker benefit results. Word recognition performance for human listeners and models with different phase locking limits is plotted as a function of SNR in stationary (open circles) and amplitude-modulated (closed circles) noise. c. Schematic of tone vocoding stimulus manipulation with a "cutoff channel" of 10. A speech waveform was separated into 32 frequency subbands by a bank of band-pass filters that mimic the cochlea's frequency tuning. Frequency channels up to and including the cutoff channel were left intact. In the 22 higher frequency channels, temporal fine structure was disrupted by replacing the channel subband with an envelope-modulated pure tone carrier at the channel's center frequency. d. The benefit from temporal fine structure was quantified by measuring leftward shifts in psychometric functions plotting word recognition accuracy vs. SNR as the cutoff channel (i.e., the number of channels with intact temporal fine structure) is increased. All shifts are computed relative to performance with fully tone-vocoded speech (0 channels intact, orange circles). e. Tone vocoding results. The benefit from temporal fine structure – measured from humans and models – is plotted as a function of the number of channels with intact temporal fine structure. Open circles plot the benefit in stationary noise and closed circles plot the benefit in amplitude-modulated noise. Human data in e is re-plotted from the original study[36] and errors bars indicate ±1 standard error of the mean across participants. All other error bars in a, b, and e indicate ±2 standard errors of the mean across human participants or network architectures. f. Human-model similarity scores for each phase locking condition and speech experiment. Similarity scores were Pearson correlation coefficients between corresponding human and model behavioral data points. The rightmost column averages scores across all experiments. Error bars indicate 95% confidence intervals computed by bootstrapping the mean of 10 network architectures. Asterisks denote statistically significant differences between phase locking conditions ($p < 0.001$, two-tailed), evaluated by comparing mean similarity scores against a null distribution bootstrapped from the 3000 Hz condition.

Hz phase locking model. However, this deficit evidently did not substantially hinder word recognition in the real-world noise conditions. The auditory scenes and instrumental music contained drastic amplitude modulation, but are spectrally quite different from the target speech, perhaps allowing networks to solve the task without cues that depend critically on phase locking.

3.2.22 Phase locking is needed to account for vocoding effects in human word recognition

The minimal effects of phase locking limit on model word recognition in noise seemed counter to psychophysical evidence from "tone vocoding" suggesting temporal fine structure (TFS) cues contribute to human speech recognition in noise [36]. Tone vocoding is a stimulus manipulation in which a speech waveform is first decomposed into frequency channels with a cochlear filter bank (Fig. 3.7c). The temporal envelopes of each channel are then extracted and imposed on pure tone carriers at the center frequency of each channel. The resulting amplitude-modulated tones are passed through their respective cochlear filters and summed to produce a new waveform with similar envelope cues but different TFS from the original. Hopkins and Moore investigated the contribution of high frequency TFS by tone vocoding all frequency channels above a given cutoff. The authors increased this cutoff from 0 (all channels vocoded) to 32 (no channels vocoded), progressively increasing the upper frequency limit of TFS information preserved in the stimulus increased from 100 to 10000 Hz. Speech reception thresholds in stationary and modulated noise were measured as a function of this cutoff (Fig. 3.7d). In stationary noise, thresholds improved modestly but significantly as this frequency limit increased from 100 to 548 Hz (intact TFS in 8 of 32 channels). TFS

information provided a larger benefit in modulated noise, with significant improvements up to 4102 Hz (intact TFS in 24 of 32 channels) (Fig. 3.7e, left). This result was taken to suggest that humans benefit from the information in the TFS of sound waveforms, potentially upwards of 1000 Hz.

We tested our models on the same stimulus manipulation (Fig. 3.7e, right panels). Models with 3000 and 1000 Hz phase locking limits qualitatively and quantitatively replicated the human pattern of behavior, with speech reception thresholds continuing to improve as more high frequency TFS was preserved, particularly in modulated noise. The benefit from TFS information was reduced in the 320 Hz phase locking model and fully eliminated by 50 Hz. Collectively, these results suggest phase-locked spike timing is needed to comprehensively account for human word recognition behavior. Only models optimized with high-fidelity phase locking learned a word recognition strategy with human-like sensitivity to TFS manipulations. Models optimized without fine timing information learned a similarly performant (at least in the real-world noise conditions tested here) but less human-like strategy for recognizing words. This could imply the effect of vocoding on human word recognition is epiphenomenal (e.g., a consequence of sharing machinery with tasks that benefit more from phase locking) or advantageous in listening conditions beyond those considered here (e.g., multi-talker situations).

3.2.23 Quantitative measures of human-model similarity – voice and word recognition

We quantified human-model similarity for the voice and word recognition model by measuring correlations between analogous human and model data points in each experiment (Fig. 3.7f). Collectively, the results suggest that human-like speech perception relies on information conveyed by phase-locked spike timing, but virtually all the benefit incurs between 50 and 320 Hz ($p < 0.001$, $d = 27.7$ comparing 3000 vs. 50 Hz condition; $p < 0.001$, $d = 7.8$ comparing 3000 vs. 320 Hz condition).

3.2.24 Replication with simplified cochlear model

Networks equipped with the greatly simplified cochlear model qualitatively and in most cases quantitatively replicated all results from networks equipped with the highly detailed model of the auditory nerve (Supplementary Fig. 3.11, 3.12, 3.13, and 3.14). This result suggests that the effects we saw with the detailed cochlear model are not due to unintended interactions between its components. The results also indicate that future work could use the simplified model in many settings without a cost.

3.3 Discussion

We developed models of real-world sound localization, pitch perception, and speech perception by optimizing artificial neural networks to classify simulated auditory nerve representations of natural sounds. The resulting models closely replicated human behavior in simulated experiments with both real-world sounds and synthetic stimulus manipulations despite being optimized solely for natural sounds. To investigate the perceptual role of precise temporal coding in human hearing, we separately optimized models with lower auditory nerve phase locking limits. This manipulation had a larger effect on some tasks than others, impairing sound localization, pitch perception, and voice recognition, while leaving naturalistic word recognition largely intact. Accurate and human-like sound localization, pitch discrimination, and voice recognition only emerged in models optimized with access to phase-locked spike timing in the auditory nerve. Models with no access to phase locking above 50 Hz were particularly impaired for these tasks in noisy conditions. Comparatively, phase-locked spike timing offered little benefit for word recognition in noise, with models producing accurate and human-like word recognition in many types of real-world noise regardless of auditory nerve phase locking limit. Nonetheless, phase-locked spike timing was necessary to comprehensively account for speech recognition phenomena such as the fluctuating masker benefit and the effect of tone vocoding. Our results provide new evidence for a perceptual role of auditory nerve phase locking and suggest monaural mechanisms for extracting information from spike timing must exist in the auditory system.

The upper limit at which phase locking incurs a perceptual benefit places constraints on the underlying circuit mechanisms. For localization and pitch tasks, phase-locked spike timing up to 1000 Hz seems to be critical. For the voice and word tasks virtually all benefit of phase locking incurs between 50 and 320 Hz. Our model results suggest phase locking to the fundamental may provide an important pitch cue for voice recognition and word recognition in amplitude-modulated maskers, perhaps by facilitating pitch-related streaming.

Some evidence for the use of phase locking came from qualitative properties of human perception rather than absolute performance. Our speech perception model without access to phase locking has no trouble recognizing words in noise as well as humans did, but it appeared to learn a qualitatively different strategy than humans (lacking sensitivity to cues that disrupt human performance like the tone vocoding and pitch manipulations). Similarly, phase locking was not needed to accurately classify voices, but models optimized without it learned to weight cues very differently from human listeners (i.e., only weakly associating absolute F0 with voice identity).

3.3.1 Relation to prior experimental work

One strong interpretation of the tone vocoding effect in human word recognition [36] is that the human auditory system extracts information from spike timing that is phase locked to stimulus frequencies as high as 1000 to 4000 Hz. Our model results are inconsistent with this explanation. Speech reception thresholds from our models continued to benefit from increasingly high frequency TFS, even beyond their respective phase locking limits. For instance, the 1000 Hz model received a benefit from frequencies above 1000 Hz. Since our simulated auditory nerve representations cannot, by definition, encode TFS above the set phase locking limit, the task-relevant information contributed by high frequency TFS cannot be represented by high frequency phase-locked spike timing.

An alternative explanation for the effect of tone vocoding on human behavior relates to pitch. Tone vocoding is not a perfectly selective stimulus manipulation and vocoding away high frequency TFS also interferes with harmonic frequency relationships in speech. When high-numbered harmonics are vocoded, they no longer produce temporal envelope cues to pitch, which the auditory nerve encodes via phase locking to the F0 (typically below 300 Hz). Pitch is an important cue for sound segregation and disrupting the encoding of F0 could account for speech perception deficits. Consistent with this alternative explanation, network word recognition exhibited very similar deficits in noise when tested with inharmonic speech and with tone-vocoded speech (Supplementary Fig. 3.15). These manipulations both disrupt F0, but inharmonic speech preserves the TFS of non-periodic (unvoiced) phonemes [155]. Making speech inharmonic had less of an effect on models without access to phase locking. Models with degraded temporal coding performed better on inharmonic speech than models with high-fidelity phase locking. This was the case for both word recognition and voice recognition, where effect sizes were even larger. Collectively, these results suggest that F0 cues encoded in phase-locked spike timing are important for hearing in noise.

3.3.2 Relation to prior modeling work

Our approach draws inspiration from ideal observer theory, which has a rich history interrogating the role of temporal coding in hearing [9], [10]. In the 1960s, Siebert first derived the optimal solutions to simple psychoacoustic tasks given the information available in auditory nerve responses to simple stimuli. This classic approach has yielded valuable insights about which features of neural coding are needed to qualitatively account for human behavior, but the resulting models severely overestimate human performance. This is perhaps unsurprising because there is little reason to believe that the human auditory system is near-optimal for the simple psychoacoustic tasks and stimuli for which it is tractable to derive provable ideal

observers. Over evolutionary and developmental timescales, naturally behaving humans are rarely tasked with discriminating pure tones. For the perceptual tasks that humans realistically are optimized for (e.g., recognizing and localizing natural sounds) [57], deriving provably ideal observers quickly becomes intractable. Here, we showed the toolbox of contemporary deep learning provides an avenue to resurrect this classic approach for real-world perceptual tasks.

3.3.3 Limitations

Our approach is not without drawbacks. There is no guarantee that current deep optimization methods converge on optimal task solutions. Comparisons of absolute performance should thus be interpreted with caution. Does a model without access to phase locking fail to achieve human-level performance because precise spike-timing information is strictly necessary for the task or because the model is insufficiently optimized? It is not impossible that ever more expressive network architectures and optimization methods could lead to human-level performance without precise spike-timing. We hedged against this possibility in two ways. First, we used multiple network architectures for each model, ensuring reported results do not reflect the idiosyncrasies of any single network architecture. Second, our conclusions do not hinge solely on differences in absolute performance. We compared human and model behavior across a broad range of psychoacoustic experiments, allowing us to identify qualitative differences in how models solve tasks given different peripheral input. Even though our different phase locking models equally accounted for human performance in some conditions, only models with higher phase locking limits exhibited human-like cue dependencies suggesting the perceptual strategies employed by humans rely on phase locking.

Like ideal observers, artificial neural networks are also susceptible to overestimating human performance. Some of this can be attributed to human participants tiring or suffering attentional lapses during experiments. Besides spike sampling in their auditory nerve input, our network models are deterministic systems. Ideal observers often posit decision stage noise to bring model performance down to the level of humans [9], [10], and the same logic could be applied to our networks. Another possibility is optimizing with insufficient constraints. We found evidence of this in our 3000 Hz phase locking networks optimized for localization. Networks with immediate access to the left and right auditory nerve representations exhibited sensitivity to ITDs at much higher frequencies than humans. Consistent with the idea that insufficiently constrained optimization (i.e., neglecting any cost of binaural circuitry) led to this superhuman ITD sensitivity, simply altering the network architecture to require monaural processing stages before binaural integration produced more human-like

behavior. Our results suggest that one reason the human auditory system did not evolve to use ITDs much above 1000 Hz is because they do not benefit sound localization in natural settings. This manipulation provides one example of how our modeling approach enables normative analysis of sensory physiology. Additional architectural manipulations may yield further insights. For instance, restricting interactions between frequency channels might help elucidate the functional role of tonotopy throughout the auditory hierarchy.

Our models offer insight into which neural coding cues underlie perception but not how. In our framing, the deep artificial neural network portions of our models may as well be black boxes. They instantiate optimized solutions to tasks by performing opaque series of linear and nonlinear operations. They are no more a mechanistic hypothesis of how the brain works than the equations specifying an ideal observer are. Our approach is complementary to mechanistic models, however, and future work could powerfully combine the two. Mechanistic hypotheses – such as proposals for how brainstem circuitry may extract task-relevant cues [111], [114] – can be built into our models. Task-optimizing the deep network backend and asking if the resulting model still accounts for behavior would be a strong test of these hypotheses. Mechanisms that preclude human-like behavior in a backend optimized for natural tasks can potentially be ruled out as mechanisms in the brain. Building in interpretable but optimizable signal processing [156] stages could lead to mechanistic insights that are adapted to the demands of natural tasks.

Insufficient task demands may have contributed to the relatively small effect of phase locking observed for word recognition. Speech models optimized for cocktail party situations, where accurate pitch and localization cues help segregate multiple concurrent voices [155], [157], may reveal larger contributions of phase locking to speech recognition.

Model predictions are sometimes limited by the availability of naturalistic training data. Our pitch model was optimized with a limited range of F0s (constrained by available speech and music corpora), which prevented our model from making predictions about the perception of very high frequencies[28]. Similarly, the availability of low-elevation anatomical transfer functions and word- and voice-labeled speech limited our localization and speech model predictions to classes in the training datasets.

3.3.4 Future directions

Our approach has natural extensions for modeling sensorineural hearing loss and cochlear implants. The healthy auditory periphery hard-wired into our models can be altered to simulate hair cell loss [4], cochlear neuropathy [158], or electrical stimulation from a cochlear implant [159]. Optimizing models with different hearing loss etiologies may yield insights into

the diverse behavioral outcomes of people with hearing loss. We found that similarly accurate predictions of human behavior were possible with a greatly simplified cochlear through which gradients can be backpropagated. This raises the possibility of directly optimizing front-end processors [37] (or even individual sounds [93], [160]) for perceptual outcomes in the model, which could be useful for developing hearing aids, cochlear implants, and diagnostic behavioral tests.

Our results place strong constraints on future models of the auditory system. Accurate models of human localization, pitch perceptions, and speech perception should all operate on high temporal resolution input. Future work may enhance ecological validity by jointly optimizing for more tasks – for instance, training to recognize and localize words and voices. Though not the focus of this work, network representations could also be compared to brain representations, perhaps yielding insights into hierarchical processing for different tasks [13], [21]. Relating network processing stages to human EEG and ABR measures could be particularly impactful in the clinic [161].

The general approach of investigating neural coding features with models optimized for ecological tasks is not limited to hearing. Similar analysis of tactile perception could, for instance, elucidate the perceptual role of high-fidelity temporal coding in touch [128]. Optimization methods that simulate the inescapable forces of evolution and development may offer far-ranging insights into how neural codes give rise to behavior.

3.4 Methods

3.4.1 Peripheral auditory model

The Bruce et al. (2018) auditory nerve model [5] served as the peripheral front-end to our artificial neural networks. 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 nonhuman animals. 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 synapse between inner hair cells and auditory nerve fibers. Although the model's responses have only been directly compared to recordings made in nonhuman animals, some model parameters have been inferred for humans (such as the bandwidths of cochlear filters) based on behavioral and otoacoustic measurements.

The output of the auditory nerve model was a three-dimensional array of instantaneous auditory nerve firing rates with shape [N frequency channels, T timesteps, S fiber types]. We simulated instantaneous auditory nerve firing rates at N=50 (localization and speech model) or N=100 (pitch model) points along the cochlear frequency axis. Auditory nerve fiber characteristic frequencies were spaced uniformly on an ERB-number scale between 125 and 16000 Hz for the localization model. For the pitch and speech models, the highest characteristic frequency was 14000 and 8000 Hz respectively. The use of 50 to 100 frequency channels primarily reflects computational constraints (CPU time for simulating peripheral representations, storage costs, and GPU memory during training). In previous work we found that increasing the number of frequency channels tenfold had little effect on model behavior [17]. The instantaneous firing rates were downsampled from audio sampling rates to 10 kHz for the localization and speech model or 20 kHz for the pitch model. The localization model operated on 1s inputs (T=10000). The speech model operated on 2s inputs (T=20000). The pitch model operated on 50ms inputs (T=1000). The high sampling rates of auditory nerve responses ensured the information in high-frequency phase locking up to 3000 Hz could be represented. At each characteristic frequency, we simulated responses of S=3 different auditory nerve fiber types to represent canonical high (70 spikes/s), medium (4 spikes/s) and low (0.1 spikes/s) spontaneous rate fibers [130]. Fibers with different spontaneous rates vary systematically in their thresholds and dynamic ranges. High spontaneous rate fibers having the lowest thresholds but smallest dynamic ranges such that their firing rates saturate at conversational speech levels.

This array of instantaneous firing rates was then converted to an array of binomially sampled spike counts representing the population response of 32000 individual auditory nerve fibers per ear. The number of spikes occurring at each time-frequency-fiber in was sampled from a binomial distribution with p = firing rate / sampling rate and n determined by the relative numerosity of different fiber types (n = fraction of fibers × 32000 total fibers / N frequency channels). We used 60% high, 25% medium, and 15% low spontaneous rate fibers [130]. To reduce the computational cost of sampling from 1.5 million ($N \times F \times T$) independent binomial distributions per ear and stimulus, we approximated sampling from Binomial(n, p) by rounding samples from Normal$(np, np(1-p))$ in the localization model.

3.4.2 Phase locking manipulation

The phase locking manipulation was identical to that introduced in our previous work [17]. We modified the upper frequency limit of phase locking in the auditory nerve by adjusting the cutoff frequency of the inner hair cell low-pass filter within the auditory nerve model. By default, the low-pass characteristics of the inner hair cell's membrane potential are modeled as a 7th order filter with a cutoff frequency of 3000 Hz. We set this cutoff set to 3000, 1000, 320 and 50 Hz.

3.4.3 Simplified cochlear model

The Bruce et al. (2018) auditory nerve model is computationally expensive to run, requiring peripheral representations to be precomputed and stored on disk rather than generated on-the-fly during network optimization. Simulated auditory nerve representations of the training datasets alone required 12 TB (localization), 1 TB (pitch), and 26 TB (speech) per phase locking condition. We repeated experiments with a simplified cochlear model hard-wired into the network's computation graph, eliminating the need to store precomputed peripheral representations. This simplified front-end consisted of a linear cochlear filter bank followed by half-wave rectification and low-pass filtering to impose the upper limit on phase locking. Simplified cochlear models operated on 50, 32 and 20 kHz audio for the localization, pitch, and speech tasks respectively. After low-pass filtering (with a 50ms Kaiser-windowed sinc filter), cochlear representations were downsampled from audio sampling rates to 10 kHz for the localization and speech models or 20 kHz for the pitch models. The simplified localization model used a gammatone filter bank with impulse responses truncated to 50ms. The simplified pitch and speech models used a rounded exponential filter bank in the frequency domain. Half-wave rectified and low-pass filtered subbands were passed through pointwise sigmoid functions approximating the rate-level functions of high, medium, and low spon-

taneous rate fibers. These sigmoid rate-level functions ranged from a spontaneous rate to a maximum firing rate of 250 spikes/s over a dynamic range of 20, 40, or 80 dB for high, medium, and low spontaneous rate fibers with respective thresholds of 0, 12, and 28 dB SPL. These stages yielded an array of instantaneous auditory nerve firing rates with the same dimensions as the detailed auditory nerve model. The spike sampling procedure was identical between the simplified and detailed cochlear models. We repeated the temporal coding manipulation in the simplified cochlear model by setting the low-pass cutoff of the Kaiser-windowed sinc filter to 3000, 1000, 320, and 50 Hz.

3.4.4 Artificial neural network architectures

Simulated auditory nerve representations were passed as input to deep convolutional neural networks, each consisting of a series of feedforward layers. These layers were hierarchically organized and instantiated one of several simple operations: linear convolution, pointwise nonlinear rectification, pooling, normalization, linear transformation, dropout regularization, and softmax classification.

For each task, we used 10 distinct network architectures previously identified in large-scale random searches over architectural hyperparameters (e.g., number of layers, units per layer, convolutional kernel size and shape, and pooling extent). The individual network architectures for each task are summarized in Supplementary Tables 3.1, 3.2, and 3.3. For the localization model, we used the top-10 performing architectures from Francl and McDermott (2022) which implement pooling and normalization via max pooling and batch normalization operations. For the pitch models, we used the top-10 performing architectures from Saddler et al. (2021) which implement pooling and normalization via Hanning-weighted average pooling and batch normalization operations. For the speech models, we used the best-performing network architecture from Saddler and Francl et al. (2021) as a starting point. We first modified the initial network to use Hanning-weighted average pooling and layer normalization operations. We then performed a local architecture search by making 20 new architectures via single hyperparameter modifications from the starting point (e.g., add/subtract one layer or change the convolutional kernel shape in one layer at a time). We used the 10 best-performing networks from this local architecture search for the speech models in this work.

3.4.5 Localization network architecture modification

Unlike the pitch and speech models, localization models operated on binaural input. We concatenated the simulated auditory nerve representations from the left and right ear along

the last axis into a single array with shape [N frequency channels, T timesteps, 2S fiber types]. Standard convolutional layers have three-dimensional kernels, which allow the optimizable filters to integrate information across the last feature axis (i.e., between the ears). Our default convolutional neural network architectures imposed no restriction on where in the processing hierarchy interaural cues can be extracted. To test the effect of delaying binaural integration, we replaced standard convolution operations in the earliest layers with grouped convolutions. Grouped convolutions split their input representation along the feature axis and use a separate convolutional kernel filter for each group [162]. Setting the number of groups to 2 in the first convolutional layer separates the input for the left and right ear. Successive convolutional layers with 2 groups maintain separate monaural processing streams. Binaural integration only occurs at the first convolutional layer there the number of groups is set to 1 (standard convolution). To delay binaural integration in our networks until after significant temporal pooling had occurred, we set the number of groups to 2 for all convolutional layers before the representation was downsampled by a factor of at least 4 (from 10 kHz to no greater than 2.5 kHz). In Supplementary Table 3.1, the convolutional layers replaced with grouped convolutions in the delayed binaural integration models are highlighted.

3.4.6 Transposed pitch network architectures

Lowering the phase locking limit to 50 Hz eliminates virtually all temporal structure in the short-duration (50ms) stimuli used for the pitch model, leaving only rate-place information in the auditory nerve responses. In our previous work we controlled for the possibility that models with access only to rate-place information could be limited by the relatively small number of frequency channels simulated. Instead of using nerve representations with 100 frequency channels and 1000 timesteps (sampled at 20 kHz), we used 1000 frequency channels and 100 timesteps (sampled at 2 kHz) for the lowest phase locking condition. We transposed the time and frequency axes before passing these representations to the networks to maintain the expected input shape. Transposing the input representation effectively changes the orientation of a network's convolutional filters, such that kernels that were previously long in time became long in frequency. We found this manipulation yielded better absolute pitch discrimination thresholds but did not qualitatively change model behavior. To report effects of our phase locking manipulation more conservatively, we applied this same modification to our 320 and 50 Hz phase locking pitch models here.

3.4.7 Model optimization – overview

Artificial neural networks were optimized to perform real-world hearing tasks operationalized as classification tasks. The training datasets and individual tasks are described in the subsequent sections. In general, training stimuli were class-labelled (one label per task) and network parameters were iteratively updated to minimize the softmax cross-entropy loss function via stochastic gradient descent (ADAM optimizer) with gradients computed via back-propagation. Localization networks trained for 200,000 steps with a batch size of 32 and learning rate of 0.0001. Pitch networks trained for 150,000 steps with a batch size of 64 and learning rate of 0.00001. Voice and word recognition networks trained for 400,000 steps with a batch size of 32 and learning rate of 0.00001. Classification performance on held-out validation sets was recorded after every 5000 to 10000 training steps. The network weights producing the highest validation set performance during the training routine were used as the trained model. The number of steps in each model's training routine was chosen to produce plateauing validation set performance under all phase locking conditions. Network training times vary by architecture, but each model could be trained in 96 hours on a single NVIDIA A100 GPU on the MIT OpenMind Computing Cluster. Localization models trained in under 48 hours and the much smaller pitch models trained in under 12 hours.

3.4.8 Model optimization – sound localization

We used the sound localization task of Francl and McDermott (2022) in which models classified noisy 1s auditory scenes according to the azimuth and elevation of a target natural sound. The source location classes spanned 360° in azimuth (5° bin width) and 0 to 60° in elevation (10° bin width), yielding a total of 504 output classes (72 azimuth × 7 elevation classes). To ensure the task was well-defined, naturalistic auditory scenes always consisted of a single natural sound rendered at one target location and real-world noise textures diffusely localized at 3 to 12 different distractor locations. Target sounds were taken from the Glasgow Isolated Sound Events (GISE-51) [163] subset of Freesound Dataset 50k (FSD50K) [164], which consists of variable-length recordings of individual sources spanning 51 categories of everyday sounds. We only used source clips for which the original 44.1 kHz sampling rate audio could be found in FSD50K. Our training and validation datasets consisted of 12465 and 1716 unique source clips, respectively. For model evaluation and human experiments, we used 460 sounds from the GISE-51 evaluation set equally distributed across 46 sound categories (discarding 5 categories, each with fewer than 10 evaluation clips).

Texture-like background noise was sourced from a subset of the Audioset [107] corpus screened to remove nonstationary sounds (e.g., speech or music). The screening procedure

involved measuring auditory texture statistics [108] (envelope means, correlations, and modulation power in and across cochlear frequency channels) from all recordings, and discarding segments over which these statistics were not stable in time, as in previous studies. The screening procedure yielded 26515 and 562 unique 10s noise clips for the training and validation datasets, respectively. Naturalistic auditory scenes were constructed by combining randomly sampled pairs of target sounds and texture-like noise samples (sliced into 1s segments). As in previous work [18], we augmented the number of unique target waveforms by applying a randomly generated band-pass filter to the target in 50% of training and validation examples. Band-pass filter center frequencies were sampled log-uniformly between 160 and 16000 Hz. Bandwidths were sampled log-uniformly between 2 and 4 octaves and the filter order was drawn uniformly between 1 and 4. Individual target and noise sources were first spatialized and then summed together at SNRs uniformly drawn between -15 and +25 dB, except for 5% of scenes which included no noise. For all localization experiments, SNR referred to the target sound's level relative to the sum of all background noise sources.

To spatialize scenes, we used a virtual acoustic room simulator [165] to render sets of binaural room impulses responses (BRIRs) for a KEMAR in 2000 unique listener environments. The simulator used the image-source method and incorporated KEMAR's HRTFs [151]. We randomly generated 2000 unique listener environments by sampling different shoebox rooms (varying in size and wall materials) and listener positions (x, y, z coordinates and head angle) within each room. For each listener environment, we rendered BRIRs at 1008 source locations (2 distances at each of the 504 azimuth and elevation pairs). 1800 unique listener environments were included in the training set and the remaining 200 were used for validation. The final training and validation datasets consisted of 1,814,400 and 201,600 binaural auditory scenes, respectively. Target natural sounds were placed once at each of the 2000 $\times$ 1008 source locations to ensure the dataset was balanced across the 504 target location classes. Auditory scenes were presented to the model during training at sound levels drawn uniformly between 30 and 90 dB SPL.

3.4.9 Model optimization – pitch perception

The pitch task and training dataset used here were unchanged from our previous work [17]. In short, we operationalized pitch perception as F0 estimation by having networks classify 50ms excerpts of speech and music embedded in nonperiodic texture-like background noise according to F0. There were 700 F0 bins spanning 80 to 1000 Hz on a logarithmic scale (bin width = 1/16 semitones = 0.36% F0). F0 labels for the training stimuli were computed via the STRAIGHT algorithm [104] applied to clean speech and music excerpts. The final

training consisted of 2.1 million exemplars sampled at 32kHz with SNRs drawn uniformly between -10 and +10 dB and sound levels drawn uniformly between 30 and 90 dB SPL. Models were trained on 80% of the dataset and 20% was used for validation.

3.4.10 Model optimization – voice and word recognition

Voice and word recognition are intertwined in this work because the same dataset was used to train both tasks. We used an augmented version of the Word-Speaker-Noise dataset [93], which consists of 230,356 unique speech clips embedded in 718,625 unique nonspeech background noise clips from Audioset [107]. Randomly sampled pairs of 2s speech and noise clips were combined to yield a training dataset of 5.8 million examples. A validation set of 370,000 examples was similarly constructed from speech and noise clips excluded from training. Each example is labeled with the talker that produced the speech utterance and the word that appeared in the middle of the utterance (i.e., overlapped the 1s mark of the 2s utterance). The datasets contain 433 unique talker labels and 794 unique word labels.

Training models for robust voice recognition is complicated by the fact that large speech corpora are often crowd-sourced online, with individuals contributing recordings of themselves reading passages or responding to prompts. The resulting dataset is accurately talker-labeled but models optimized for talker classification may pick up on non-voice cues that predict these labels (e.g., characteristics of the recording device or environment). To ensure our models learned robust voice representations, we applied a set of randomly sampled audio manipulations to the speech to approximate the variable conditions in which human listeners may encounter the same voices. In 25% of the dataset, speech clips were augmented to increase natural voice variability by applying small pitch ($\pm$ 0.5 semitones) and tempo shifts ($\pm$ 20%) or simulating whispering (less than 0.5% of examples) via the STRAIGHT algorithm [104]. In an independently drawn 25%, we applied commonly encountered audio distortions like band-pass / equalization filters, lossy audio compression / transmission, and dynamic range companding. In another independently drawn 5%, we replaced background noise with 12 to 36-talker babble to give the model some exposure to multi-talker situations. SNRs for the augmented speech clips were drawn uniformly between -10 and +10 dB and sound levels were drawn uniformly between 30 and 90 dB SPL.

We jointly optimized individual networks to recognize both voices and words using the augmented dataset. Multi-task optimization was accomplished with a separate output layer for each task. All other network stages were shared between the two tasks and parameters were updated to minimize the sum of the softmax cross entropy loss from both tasks.

3.4.11 Model evaluation – overview

For each task, we compared naturalistic human and model behavior in noise. Humans and models were evaluated on the same tasks with the same stimuli wherever possible. For the localization and speech models, we collected new human behavioral data to compare against. For the pitch task model, we measured model F0 discrimination thresholds in noise using stimuli from a recent human study. For each task, we also tested models on classic stimulus manipulations from the psychoacoustics literature and compared behavior to published human results. Human-model behavioral similarity was quantified for each experiment and model by measuring Pearson correlation coefficients between analogous human and model data points.

3.4.12 Localization model evaluation – sound localization in noise

Human experiment

We measured human's ability to localize natural sounds in noise using a 19-by-5 array of loudspeakers arranged on a hemisphere with 2m radius. The array spanned the 180° in azimuth (frontal hemifield) and 0° to 40° in elevation (10° spacing in both azimuth and elevation) relative to the listener's head at the center. 11 normal hearing listeners (5 female) with ages between 10 and 30 each performed 460 trials with 460 unique target natural sounds from the GISE-51 evaluation dataset. On each trial a target natural sound was played from one of the 95 loudspeakers and threshold equalizing noise played from 9 distinct loudspeakers. Target and noise locations were randomly sampled each trial. The listener's task was to report which loudspeaker produced the target by entering the loudspeaker's identifier on a keypad. Target sounds were presented at 60 dBA and noise levels were determined such that the SNR of the target relative to the sum of the 9 noise sources was -12, -6, 0, +6, +12 dB or infinite (no noise). All stimuli were sampled at 44.1kHz and were 1s in duration, including 15ms onset and offset ramps (Hanning window).

Model experiment

Models were tested on all combinations of the 460 target natural sounds, 6 SNRs, and 95 target locations (262,200 total stimuli) used in the human experiments. Sources were spatialized in a virtual rendering of the speaker array room human listeners were evaluated in. To match the task between human and models, we restricted network localization judgments to just azimuth and elevations corresponding to the 95 loudspeaker locations.

Human-model comparison

Human and model performance was quantified by measuring mean absolute spherical error (great circle distance), azimuth error, and elevation between the true and reported target sound location. Human-model similarity scores report the correlation between these human and model error metrics as a function of SNR. We also reported mean absolute azimuth and elevation errors as a function of SNR.

3.4.13 Localization model evaluation – psychoacoustics

We simulated an expanded version of the battery of localization experiments used in Francl and McDermott (2022). The stimuli and analysis of 6 of 8 psychoacoustic experiments were unchanged from previous work. The minimum audible angle and ITD lateralization experiments are new additions. All psychoacoustic stimuli for model localization experiments were sampled at 44.1 kHz.

3.4.14 Minimum audible angle vs. azimuth

Human experiment

Mills (1958) measured human localization acuity as a function of frequency and azimuth by playing pure tones to a blindfolded listener from a rotating boom in an anechoic chamber. Minimum audible angle thresholds were defined as the smallest change in azimuth required for the listener to discriminate whether a tone's location shifted left or right between two presentations. Though Mills (1958) only reported thresholds measured from a single human listener, the key result that localization acuity is best near the midline and decays steadily towards the periphery is well-established and holds across different experimental paradigms.

Model experiment

We measured model thresholds by simulating a left/right lateralization experiment. Pure tones (1s duration including 70ms onset and offset Hanning window ramps) were spatialized in a virtual anechoic room at 0° elevation and azimuths of -90 to +90° in steps of 0.5° (using linear interpolations of BRIRs spaced 5° apart). For each tone, we collected model predicted location probability distributions. These distributions were then multiplied by a mask assigning zero probability to nonzero elevations and azimuths outside the frontal hemifield. This resulted in probability distributions over predicted azimuth in the frontal hemifield for each stimulus. Left/right discrimination trials were simulated by comparing

the means of these distributions for pairs of stimuli rendered at different azimuths. Trials in which the signed predicted azimuth of the second tone was larger than the signed predicted azimuth of the first were considered rightward shifts. Minimum audible angle thresholds for different frequencies (250, 500, 750 and 1000 Hz) and reference azimuths (0° to 75° in steps of 5°) were inferred from separate psychometric functions (proportion of rightward shifts as a function of azimuth difference) constructed from all possible trials within ±10° azimuth of the reference. Model minimum audible angle thresholds were defined as the azimuth difference that yielded 70.7% rightward shifts (calculated by fitting Normal cumulative distribution functions to the psychometric functions).

Human-model comparison

We averaged human and model thresholds across pure tone frequencies of 250, 500, 750, and 1000 Hz. Human-model similarity was quantified by correlating average model thresholds with linearly interpolated human thresholds as a function of absolute azimuth between 0° and 75°.

3.4.15 ITD / ILD cue weighting

Human experiment

We simulated the experiment of Macpherson and Middlebrooks (2002), which measured shifts in perceived azimuth for virtual sounds with additional ITDs and ILDs imposed. In the original experiment, 13 participants (5 female) were played sounds over headphones and reported perceived azimuth by turning their head to face the virtual source. The experiment took place in an anechoic chamber and used both low-pass (0.5 to 2 kHz) and high-pass (4 to 16 kHz) 100ms noise bursts with 1ms squared-cosine ramps at the onset and offset.

Model experiment

We used identical stimuli spatialized in a virtual anechoic room at 0° elevation and 0° to 360° azimuth in steps of 5°. For each of the source locations and noise bands, we also separately created ITD- and ILD-biased versions of the stimuli. ITD-biased versions were modified imposed additional ±300µs and ±600µs time delays between the two ears. ILD-biased versions imposed additional ±10 and ±20 dB level differences between the two ears. We collected model azimuth predictions for each stimulus. Azimuth predictions in the rear hemifield were mapped to the frontal hemifield by reflecting across the coronal plane. We paired the model azimuth prediction for each ITD- and ILD-biased stimulus ("the biased azimuth") with the

azimuth prediction for the corresponding unbiased stimulus ("the unbiased azimuth"). We then computed shifts in the biased azimuth relative to the unbiased azimuth. Azimuth shifts for ITD-biased stimuli were expressed in µs by subtracting the ITD of a real source at the biased azimuth from the ITD of a real source at the unbiased azimuth. Azimuth shifts for ILD-based stimuli were expressed in dB by subtracting the ILD of a real source at the biased azimuth from the ILD of a real source at the unbiased azimuth. Expressing azimuth shifts in cue units enables calculation of a dimensionless perceptual weight by dividing the azimuth shift by the imposed cue amount. Separate ITD and ILD perceptual weights were computed for low-pass and high-pass noise by averaging across all azimuths and bias magnitudes. An ITD perceptual weight of 1 indicates that, for a given virtual stimulus, imposing 600µs of additional ITD shifts the perceived azimuth by an angle corresponding to a 600µs ITD change between two real source locations. Perceptual weights of 0 indicates that imposing additional ITDs or ILDs has no effect on the perceived the azimuth.

Human-model comparison

Human-model similarity was quantified by correlating ITD and ILD perceptual weights measured for low-pass and high-pass noise between humans and models.

3.4.16 ITD lateralization vs. frequency

Human experiment

The upper frequency limit of fine structure ITD use in humans has classically been measured by asking listeners to make left/right lateralization judgments with pure tones presented over headphones. The pure tones have variable fine structure ITDs but identical envelopes (fixed window to eliminate onset ITDs) and zero ILD (identical amplitude between the two ears). Listeners hear pairs of tones with different ITDs and judge whether the second tone sounded left or right of the first. The ΔITD threshold is the smallest change in ITD between two tones that a listener can reliably lateralize. Brughera et al. (2013) measured ΔITD thresholds of 4 young adult listeners (1 female) with 250, 500, 700, 800, 900, 1000, 1200, 1250, 1300, 1350, and 1400 Hz pure tones.

Model experiment

We simulated the experiment on our models by collecting predicted location probability distributions from our networks tested on 500ms pure tone stimuli (including 100ms linear onset and offset ramps). Fine structure ITDs ranged from -160µs to 160µs in steps of 1µs.

Frequencies ranged from 50 to 4000 Hz in steps of 50 Hz. Lateralization ΔITD thresholds were inferred from model judgments using the same method as the minimum audible angle experiment. Model predictions were compared for pairs of stimuli with different ITD, and rightward shifts were defined as trials in which the signed azimuth prediction was larger for the second tone than the first. Psychometric functions were constructed for each frequency (proportion of rightward shifts as a function of ΔITD) and the ΔITD threshold was defined as the difference in azimuth yielding 70.7% rightward shifts.

Human-model comparison

Human-model similarity was quantified by correlating log-transformed model thresholds with linearly interpolated human thresholds as a function of frequency between 250 and 1500 Hz.

3.4.17 Effect of changing ears

Human experiment

We simulated a change in our models' ears analogous to the manipulation of Hofman et al. (1998). In the original experiment, 4 participants localized white noise bursts presented in a 4-by-4 grid uniformly tiling ±20° in azimuth and ±20° elevation. Participants reported perceived locations by making eye movements to the source. After collecting baseline azimuth and elevation judgments, plastic molds were inserted in the participants ears, which altered the direction-specific filtering of their pinnae. Participants then repeated the localization task with modified ears.

Model experiment

We simulated the experiment by collecting baseline model azimuth and elevation judgments with the same stimuli (500ms noise bursts with a frequency band of 0.2 to 2 kHz) and then switching out KEMAR's HRTFs for 45 different sets of HRTFs from the CIPIC dataset. Model azimuth and elevation predictions were collected for stimuli spatialized on 4-by-4 grid uniformly tiling ±30° in azimuth and 0° to 30° in elevation. Azimuths and elevations were not matched exactly to the human experiments due to constraints of the available HRTFs. We averaged model judgments across the 45 different sets of HRTFs not used to train the model to compare against human judgments with modified ears.

Human-model comparison

To summarize the effects of changing ears on azimuth and elevation accuracy, we computed changes in mean absolute azimuth and elevation error with the untrained ears relative to the trained ears, averaging across all 16 source locations. Human-model similarity was quantifying by correlating absolute azimuth and elevation errors as a function of grid position and ear condition between humans and models.

3.4.18 Effect of smoothing spectral cues

Human experiment

We simulated a modified version of the experiment by Kulkarni and Colburn (1998), which measured the effect of HRTF spectral details on sound localization. In the original experiment, 4 listeners were played white noise bursts in an anechoic chamber. Sounds were presented from either a physical loudspeaker in the room or virtually over open-backed headphones. The virtual sounds were spatially rendered at the loudspeaker's location using the participant's own HRTFs. Participants were tasked with reporting whether the sound came from the loudspeaker or the headphones. When the participants' full HRTFs were used, performance was at chance (50%). As the spectral details of the HRTFs were smoothed out by approximating the HRTF's discrete cosine transform with progressively fewer cosines, performance rose above chance as participants no longer perceived the virtual stimuli at the loudspeaker's location.

Model experiment

We applied the same smoothing manipulation to KEMARs HRTFs (approximating the discrete cosine transform with 256, 128, 64, 32, 16, 8, 2, and 1 cosines) and evaluated model performance in a virtual anechoic room using 1s broadband (0.2 to 20 kHz) noise bursts. Model localization judgments were collected for each smoothing condition at 413 locations spanning 0° to 60° in elevation and 0° to 360° in azimuth (spacing determined by the locations of the measured KEMAR HRTFs). We computed mean absolute spherical, azimuth, and elevation errors as a function of the number of cosines used to approximate the HRTFs.

Human-model comparison

Reasoning that higher absolute localization errors in the model would correspond to better performance on the human real/virtual discrimination experiment, we quantified human-

model similarity by correlating model absolute error with human percent correct scores as a function of the smoothing parameter.

3.4.19 Median plane spectral cues

Human experiment

We simulated a modified version of the experiment by Hebrank and Wright (1974), which measured the accuracy of human elevation judgments as a function of noise burst frequency content. In the original experiment, 10 participants were played 1s noise bursts from a vertical array of speakers along the median plane spanning -30° to +210° in elevation with 30° spacing (0° is frontal). The experiment took place in an anechoic chamber and participants were tasked with reporting which speaker produced the noise burst. Noise bursts were either low-pass or high-pass with varying cutoff frequencies: 3.9, 6.0, 8.0, 10.3, 12.0, 14.5 or 16.0 kHz for the low-pass noise and 3.8, 5.8, 7.5, 10.0, 13.2 or 15.3 kHz for the high-pass noise.

Model experiment

We evaluated our model on noise bursts with the same cutoff frequencies rendered in a virtual anechoic room at elevations of 0°, 30°, 60°, 120°, 150°, and 180° along the median plane. To match the task between human and models, we restricted network localization judgments to just azimuth and elevations along the median plane.

Human-model comparison

Human-model similarity was quantified by correlating human and model percent correct scores as a function of noise type and frequency cutoff.

3.4.20 Precedence effect

Human experiment

To quantitatively compare the precedence effect between human and models we simulated the experiment of Litovsky and Godar (2010), which measured localization accuracy for 25ms (including 2ms cosine onset and offset ramps) pink noise bursts played at two different locations. The bursts were played from two loudspeakers in an array spanning -60° to +60° in azimuth (20° spacing, 0° elevation) and were delayed relative to one another by 5, 10, 25, 50, or 100ms. 10 listeners (all female) with ages between 19 and 26 were tasked with reporting whether they heard one or two sounds as well the loudspeaker that produced each

sound. Root-mean-squared azimuth errors were calculated separately for the lead and lag noise burst and reported as a function of the delay between the lead and lag bursts.

Model experiment

We evaluated models on the same stimuli rendered in a virtual anechoic room. Our models always reported a single location which we used to compute the root-mean-squared azimuth error relative to both the lead and lag burst.

Human-model comparison

Human-model similarity was quantified by correlating human and model azimuth error for both the lead and lag burst as a function of the inter-burst delay.

3.4.21 Bandwidth dependency of localization

Human experiment

We simulated the experiment of Yost and Zhong (2014), measuring the effect of bandwidth on localization accuracy with an array of 8 loudspeaker positioned between -15° to +90° in azimuth (15° spacing) relative to the midline. 33 participants (26 female) with ages between 18 and 36 were tasked with reporting which loudspeaker produced a 200ms (including 20ms squared cosine onset and offset ramps) sound. Stimuli were pure tones or band-pass filtered white noise bursts with bandwidths of 1/20, 1/10, 1/6, 1/3, 1, and 2 octaves. Pure tone and center frequencies were set to 250, 2000, and 4000 Hz. Human listeners made 20 localization judgments per bandwidth, center frequency, and loudspeaker position.

Model experiment

We evaluated our models on the same stimuli rendered in a virtual anechoic room at azimuth -90° to +90° in steps of 5°. Model localization judgments were restricted to the frontal hemifield and 0° elevation.

Human-model comparison

Human-model similarity was quantified by correlating human and model root-mean-squared error as a function of bandwidth (averaged across center frequencies).

3.4.22 Pitch model evaluation – pitch discrimination in noise

Human experiment

We simulated the pitch discrimination experiment of McPherson et al. (2022) which measured human discrimination thresholds as a function of SNR using pure and harmonic complex tones. The original experiment also included inharmonic complex tones, but we excluded these because our models were optimized for F0 estimation. All stimuli were 500ms in duration including 10ms Hanning window onset and offset ramps. Harmonic complex tones contained the first 10 harmonics in random phase with equal amplitudes. F0s and pure tone frequencies were drawn log-uniformly between 200 and 267 Hz. Tones were presented in threshold equalizing noise low-pass filtered above 6 kHz with a 6th-order Butterworth filter. Participants heard pairs of tones in noise and judged whether the second tone had a higher or lower pitch than the first. F0 discrimination thresholds were measured with a 2-down-1-up adaptive procedure that tracked 70.7% of trials correct. Thresholds were measured at SNRs of -19, -17.5, -16, -14.5, and -13 dB for pure tones and -22, -20.5, -19, -17.5, -16, -14.5, and -13 dB per component for complex tones. SNRs for the harmonic complexes are expressed per individual harmonic component such that pure and complex tone thresholds can be meaningfully plotted on the same axes. The experiment was run online and included 52 participants (20 female) who passed a headphone check.

Model experiment

We measured model F0 discrimination thresholds by evaluating networks on a superset of the stimuli from the human experiment. For each of 3329 F0s log-uniformly distributed between 158 and 337 Hz ($\pm 1/3$ octave relative to the human experiment), we generated a pure tone and a random-phase 10-harmonic complex. We embedded these stimuli in the modified threshold equalizing noise from the human experiment at SNRs between -23.5 and -11.5 dB per component in 1.5 dB increments. We simulated a two-alternative forced choice paradigm by making pairwise comparisons between model F0 predictions (restricted to be within a 1-octave band centered at the true F0) for stimuli within the same condition. 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). 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. Normal cumulative distribution functions were fit to the psychometric functions for each condition and thresholds were defined as the F0 difference (in percent, capped at 100%) that

yielded 70.7% of trials correct.

Human-model comparison

Human-model similarity was quantified by correlating log-transformed discrimination thresholds as a function of SNR and tone condition. To ensure our similarity metric was dominated by performance in noise and because the human experiment was run online, we excluded noiseless conditions from the correlation.

3.4.23 Pitch model evaluation – psychoacoustics

We simulated the full battery of pitch psychoacoustic experiments from our previous work with only minor changes to simplify presentation. All stimuli for pitch model experiments were sampled at 32 kHz.

3.4.24 Effect of harmonic number and phase on pitch discrimination

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

Stimuli

Stimuli were harmonic complex tones, band-pass 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 band-pass 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, 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. Each participant completed 4 adaptive tracks per condition.

Model experiment

The F0 discrimination experiment we ran on our networks had 600 conditions corresponding to all combinations of 2 harmonic phases (sine or random), 30 lowest harmonic numbers ($n_{low} = 1, 2, 3, \ldots 30$), 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 $\pm 6\%$ of $F_{0,ref}$) but the same band-pass 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. 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, 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 band-pass 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). We then computed the Pearson correlation coefficient between log-transformed vectors of human and network thresholds.

3.4.25 Pitch of alternating-phase harmonic complexes

We reproduced the stimulus manipulation of Shackleton and Carlyon (1994) to test if our models 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 the prior experiment, these harmonic tones were band-pass 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 two different 4th-order Butterworth filters specified by their -3 dB passbands: 125 to 625 Hz ("low 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), but we restricted our analysis to just the 125 Hz condition to simplify presentation. We generated 354 stimuli with F0s between 120 and 130 Hz for eac filter condition.

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 the two filter conditions.

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 the Pearson correlation coefficient between analogous human and network histograms (after re-binning both to use 4% bin widths).

3.4.26 Pitch of frequency-shifted complexes

We 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 two 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 "high harmonics") had a flat 5-harmonic wide passband centered at the 16th harmonic. Both 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 $x = 1 - |(F - F_e)/(1.5 \times F0)|$ 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. 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 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.

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 14 median shifts, one for each combination of spectral envelope and component shift. To obtain a human vector with 14 analogous pitch shifts, we linearly interpolated the human data between component shifts.

3.4.27 Pitch of complexes with individually mistuned harmonics

We 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 178 tones with F0s near 200 Hz (logarithmic scale within ±4% of 200 Hz). 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 8722 stimuli (178 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 from the previous experiment to measure shifts in the network's predicted F0 for all stimuli. F0 shifts were reported as a function of component shift and harmonic number.

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 49 mean shift values corresponding to the 49 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 49 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.

3.4.28 Pitch discrimination with pure and transposed tones

To investigate the necessity of correct tonotopic representation for accurate pitch perception, we 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 low-pass 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 modified uniform masking noise. The original study embedded only the transposed tones in low-pass 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 other F0 discrimination experiments. 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 averaged transposed tone thresholds 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 10 discrimination thresholds corresponding to pure tones and transposed tones at each of the 5 reference frequencies. 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.

3.4.29 Effect of level on pure tone 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.

Stimuli

We generated pure tones with 6144 frequencies spaced uniformly on a logarithmic scale between 200 and 800 Hz. Tones were embedded in modified uniform masking noise. The signal-to-noise ratio was fixed at 20 dB and the overall stimulus levels were varied between 10 and 80 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, low-pass filtered at 10000 Hz). We averaged the reported thresholds across four test frequencies (200, 400, 600, and 800 Hz).

Model experiment

We used the procedure from previously described F0 discrimination experiments 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

Human-model similarity was quantified my measuring the Pearson correlation coefficient between log-transformed human thresholds as a function of dB sensation level and log-transformed model thresholds as a function of dB SPL. Human thresholds were linearly interpolated between 10 and 80 dB sensation level to facilitate this.

3.4.30 Speech model evaluation – voice and word recognition in noise

Model experiment

We measured model voice and word recognition accuracy as a function of SNR in five different types of background noise using an evaluation set of 376 speech clips not used for training or validation. Each clip had a voice and word label included in the training dataset (376 unique word labels from 164 unique talkers). Unique background noise clips (376 per condition) were sourced from IEEE AASP CASA Challenge [166] (auditory scenes), CommonVoice [167] (8-talker speech babble), and MUSDB18 [168] (instrumental music). Stationary speech-shaped noise was synthesized by imposing the power spectrum of each evaluation speech clip on white noise. Amplitude-modulated noise was synthesized by imposing the temporal envelope from a randomly selected speech clip on the stationary speech-shaped noise [169]. Speech clips were combined with background noise from each condition at 6 SNRs (noiseless and -9, -6, -3, 0, +3 dB) yielding 11280 stimuli (376 speech clips $\times$ 5 noise type $\times$ 6 SNRs). Stimuli were 2s in duration and sampled at 20 kHz. The speech level was held fixed at 60 dB SPL and noise levels were adjusted to the desired SNRs. Networks were evaluated on the full evaluation set.

Human experiment

We measured human word recognition as a function of SNR and noise type using same stimuli and task as the model experiment. Humans were presented 2s stimuli and asked to report

which word (from a list of 793) appeared in the middle of the utterance (overlapped the 1s mark). Participants typed responses into a textbox and, as they typed, the displayed list of 793 was filtered to include only words that matched the entered string. Only responses from the word list could be submitted. The experiment was run online and included 82 participants (21 female, 60 male, 1 nonbinary) who passed a headphone check, completed at least 100 trials, and responded correctly to at least 75% of a attention check trials (isolated words presented in silence). Participants ages were between 18 and 64 (median 35) years.

Human-model comparison

We quantified human-model similarity by correlating human and model percent words correct as a function of SNR and noise types. We report correlations for the experiment in two parts: once for the auditory scenes, 8-talker babble, instrumental music, and stationary speech-shaped noise and once for just stationary and modulated speech-shaped noise. This allowed our human-model similarity metrics to distinguish between the four more naturalistic noise conditions (where we saw little effect of phase locking) and the fluctuating masker benefit experiment (where an effect was evident).

3.4.31 Voice discrimination in noise

Human experiment

To evaluate human voice recognition in noise without the confound of voice familiarity, we performed an unfamiliar voice discrimination experiment. In the experiment human listeners heard 3 different 1s speech clips spoken by two different talkers with 500ms silent gaps between. Listeners were asked to judge which talker (always either the first or last) only spoke once. We used from 252 unique voices (126 female, 126 male) from TIMIT [170] which our models were never trained on. Listeners attempted up to 504 trials such that no voice was the correct response more than once. For each participant, the 504 trials were randomly distributed across 16 noise conditions and 1 attention check condition (in which there was no noise and the same clip played twice). The 16 noise conditions were, stationary speech-shaped noise at 9 SNRs (-18 to +6 dB in increments of 3 dB), amplitude-modulated speech-shaped noise at 6 SNRs (-24 to +6 dB in increments of 6 dB), and noiseless. Speech-shaped noise for the discrimination experiment was synthesized by imposing the mean power spectrum of all speech stimuli on white noise to ensure the noise spectrum could not be used to solve the task. The experiment was conducted online and included 44 participants (19 female, 25 male) who passed a headphone check, passed at least 95% of attention check

trials, and completed at least 20 trials per noise condition. Participant ages were between 19 and 71 (median 36) years.

Model experiment

We evaluated our models on all 12096 stimuli (252 voices × 3 unique utterances × 16 noise conditions) generated for the human experiment and stored the network predicted voice class probabilities. For each noise condition, we simulated the same set of 504 voice discrimination trials human listeners were tested on by comparing triplets of these probability distributions. On each trial, we measured the Kullback-Leibler divergence between the predicted distributions for the first and second voice and the third and second voice. The voice producing a larger divergence from the second was deemed the model's judgment as to which talker only spoke once.

Human-model comparison

We quantified human-model similarity by measuring the correlation between human and model voice discrimination accuracy (percent trials correct) as a function of SNR and noise condition.

3.4.32 Voice and word recognition with pitch-altered speech

Human voice recognition experiment

We replicated the voice recognition experiment of McPherson and McDermott (2018), which measured listeners' ability to recognize pitch-altered voices from famous celebrities. Stimuli were 4s speech clips from 37 recognizable politicians, actors, singers, and television hosts. In the first block of 37 trials, each participant heard all 37 voices randomly assigned to one of 8 pitch-manipulation conditions (inharmonic or shifted ±12, ±6, ±3, 0 semitones from the original F0). In the second block of 37 trials, each participant heard different excerpts of the same 37 voices with no pitch shift. Each participants results were analyzed only for first the block, limited to just the celebrity voices they successfully recognized in the second block and identified as familiar in a pre-experiment survey. All stimuli were resynthesized with the STRAIGHT algorithm [104]. Voices were made inharmonic by shifting harmonic frequency components above the fundamental by random amounts uniformly sampled between -50% and +50% of F0 [58], [155]. Jitter values were sampled independently for each harmonic frequency and voice clip but were constrained (via rejection sampling) such that adjacent harmonics were always separated by at least 30 Hz. The experiment was a 100-alternative

forced-choice task. Participants entered responses into a textbox which filtered a displayed list of 100 celebrity names and descriptors (e.g., "Dolly Parton (country singer-songwriter)") until there was only a single match, which the participant could submit. The experiment included 112 participants (45 female, 65 male, 1 nonbinary) with ages between 20 and 73 (median 39) years. Because analysis was limited to voices for which participants demonstrated familiarity and each voice could only be assigned to one pitch condition, the number of participants for each condition ranged from 87 to 95.

Human word recognition experiment

We measured human word recognition accuracy for the same 8 pitch conditions using the 376 model evaluation set speech clips. The experiment was identical to our word recognition in noise experiment, except clips were randomly assigned to one of 8 pitch conditions and presented in quiet. The online experiment included 31 participants (10 female, 21 male) who passed a headphone check, passed at least 95% of attention check trials, and completed at least 20 trials condition. Participants ages were between 24 and 71 (median 36) years.

Model voice and word recognition experiment

We measured model voice and word recognition accuracy on the pitch-manipulated evaluation set used in the human word recognition experiment. We collect model voice and word predictions for the 376 speech clips in each of 10 pitch conditions (inharmonic or shifted ± 12, ± 9, ± 6, ± 3, 0 semitones from the original F0).

Human-model comparison

Human-model similarity across all pitch conditions and both tasks was quantified with a single correlation coefficient. We concatenated voice and word recognition scores as a function of pitch condition into a single vector of length 16 (8 voice + 8 word recognition scores) and correlated analogous vectors between humans and models.

3.4.33 Word recognition with tone-vocoded speech

Human experiment

We simulated the word recognition in noise experiment of Hopkins and Moore (2009), which measured speech reception thresholds in stationary and modulated noise using progressively tone-vocoded speech. In the original experiment, speech stimuli were split into frequency subbands with a 32-channel cochlear band-pass filter bank with center frequencies equally

spaced on an ERB-number scale between 100 and 10000 Hz. Frequency channels above a set cutoff channel (which determined the "number of channels with intact TFS") were tone vocoded to disrupt TFS. Channels were tone vocoded by imposing the temporal envelope (absolute value of the Hilbert transform) of the subband on a pure tone carrier at the channel's center frequency. Tone-vocoded subbands were band-pass filtered using the cochlear filter bank and summed together with the unmodified subbands. The resulting stimuli were presented to listeners in both stationary and modulated speech-shaped noise. Noise was amplitude-modulated with an 8 Hz sinusoid on a decibel scale with a peak-to-valley ratio of 30 dB to match the human experiment. Human speech reception thresholds were measured from 10 normal hearing participants using an adaptive procedure that tracked the SNR needed to achieve 50% of words correct. Hopkins and Moore (2009) reported speech reception thresholds with the cutoff channel set to 0, 8, 16, 24, and 32.

Model experiment

We applied the same stimulus manipulation to our 376 evaluation set speech clips and measured model word recognition accuracy in stationary and modulated noise at SNRs of -15 to +15 dB in increments of 3 dB. We used the stationary speech-shaped noise from our previous word recognition in noise experiment. Amplitude-modulated noise was matched to the human experiment by applying the same 8 Hz sinusoidal envelope to our stationary noise. Speech reception thresholds were calculated for the model by fitting a sigmoid to the psychometric metric function (word recognition accuracy as a function of SNR) for each condition and selecting the SNR that yielded half-maximal performance. We measured model speech reception thresholds with the TFS cutoff channel set between 0 (all channels tone vocoded) and 32 (all channels intact) in steps of 4. Because our models were trained with speech sampled at 20 kHz, the 32-channel Gammatone filter bank used to synthesize model stimuli had center frequencies equally spaced on an ERB-number scale between 80 and 8000 Hz rather than 100 to 10000 Hz.

Human-model comparison

Human and model speech reception thresholds for both noise types were expressed relative to speech with no intact TFS (i.e., subtracted from the threshold with cutoff channel set to 0). Human-model similarity was quantified by correlating this "benefit from TFS" as a function of SNR and noise type between humans and models.

3.4.34 Statistics

Human-model behavioral similarity was quantified separately for each model and experiment with a Pearson correlation coefficient. In each case, we compared mean model behavior (averaged across 10 network architectures) with mean human behavior (averaged across experiment participants). We calculated 95% confidence intervals for each human-model comparison by bootstrapping the model mean (sampling 10 network architectures with replacement 2000 times). The statistical significance of the effect of degraded phase locking was assessed by comparing human-model similarity scores against null distributions from the 3000 Hz phase locking models. Null distributions were Gaussian fits to the histograms of the bootstrapped human-model similarity scores for the 3000 Hz phase locking models. We then calculated the two-tailed p-values of obtaining the mean similarity scores from each degraded phase locking model under the null distribution. Effect sizes were quantified by measuring Cohen's d between bootstrapped distributions of human-model similarity scores from different phase locking conditions.

3.5 Supplementary Information

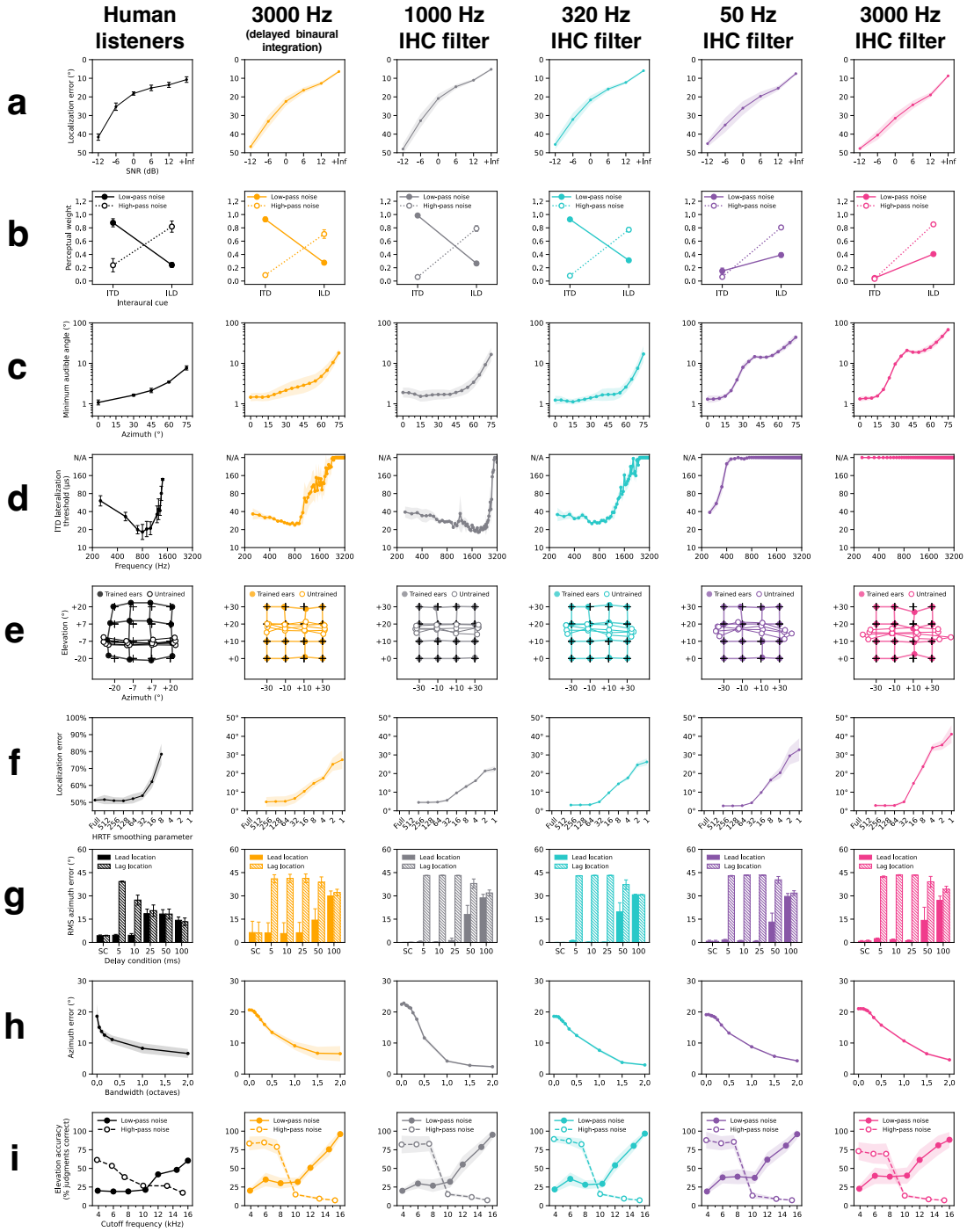

Figure 3.8: Effect of phase locking manipulation on all localization experiments.

This grid summarizes the behavioral data used to measure human-model similarity scores for the localization models. Columns correspond to human listeners and models optimized with different phase locking limits. The second (orange) column corresponds to the 3000 Hz phase locking model with network architectures modified to delay binaural integration. Rows correspond to 9 different sound localization experiments. a. Sound localization in noise. b. Minimum audible angle vs. frequency. c. ITD / ILD cue weighting. d. ITD lateralization vs. frequency. e. Effect of changing ears. f. Effect of smoothing spectral cues. g. Precedence effect. h. Bandwidth dependency of localization. i. Median plane spectral cues. All model error bars indicate ±2 standard errors of the mean across 10 network architectures. Human data in b-i is re-plotted from the original studies [132]–[135], [146], [153], [154], [171].

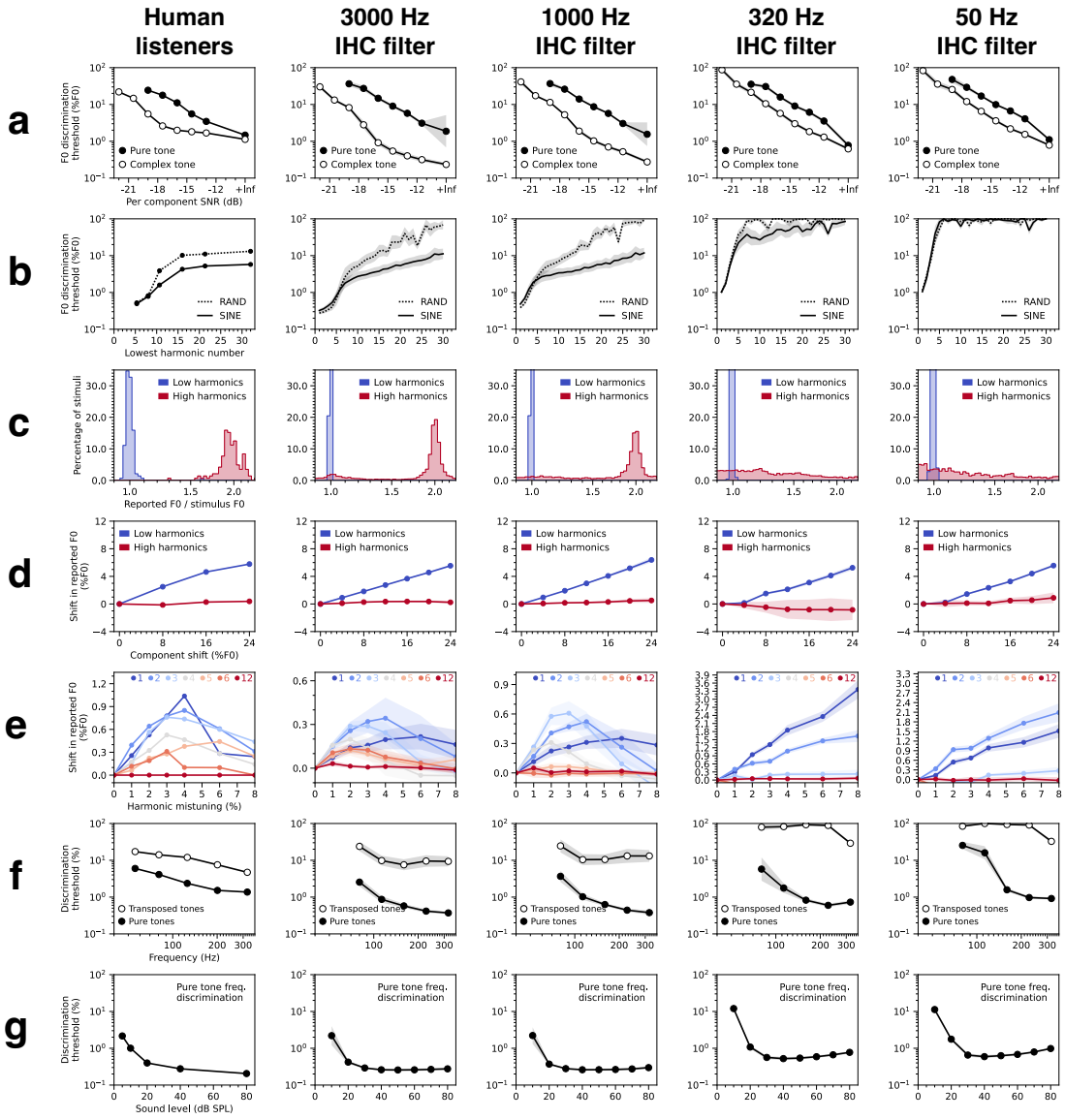

Figure 3.9: Effect of phase locking manipulation on all pitch experiments.

This grid summarizes the behavioral data used to measure human-model similarity scores for the pitch models. Columns correspond to human listeners and models optimized with different phase locking limits. Rows correspond to 7 different pitch perception experiments. a. Pitch discrimination with pure and complex tones in noise. b. Effect of harmonic number and phase on pitch discrimination. c. Pitch of alternating-phase harmonic complexes. d. Pitch of frequency-shifted complexes. e. Pitch of complexes with individually mistuned harmonics. f. Pitch discrimination with pure and transposed tones. g. Effect of level on pure tone frequency discrimination. All model error bars indicate ±2 standard errors of the mean across 10 network architectures. Human data is re-plotted from the original studies [59]–[63], [82], [89].

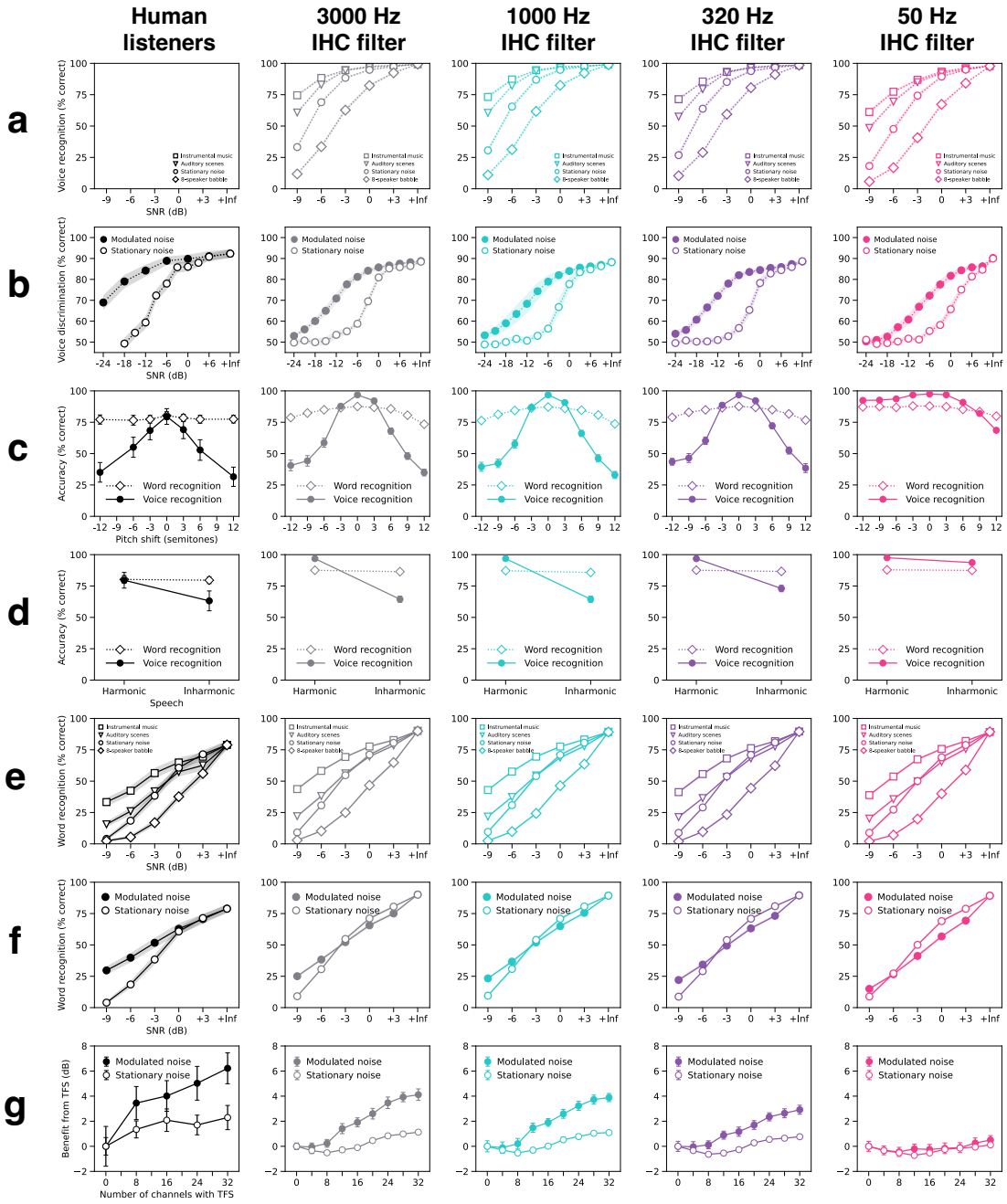

Figure 3.10: Models optimized separately for voice and word recognition – effect of phase locking manipulation on all speech experiments.

The same network architectures optimized jointly for voice and word recognition in the main text were also optimized separately for the voice and word recognition tasks for each phase locking condition. This yielded similar results to the jointly optimized models. Columns correspond to human listeners and models optimized with different phase locking limits. Rows correspond to 7 speech experiments. a. Voice recognition in real-world noise conditions (model only experiment). b. Voice discrimination in stationary and modulated noise. c. Voice and word recognition with pitch-shifted speech. d. Voice and word recognition with harmonic and inharmonic speech. e. Word recognition in real-world noise conditions. f. Word recognition in stationary and modulated noise. g. Effect of tone vocoding on word recognition in stationary and modulated noise. All model error bars indicate ±2 standard errors of the mean across 10 network architectures. Human data in g is re-plotted from the original study [36].

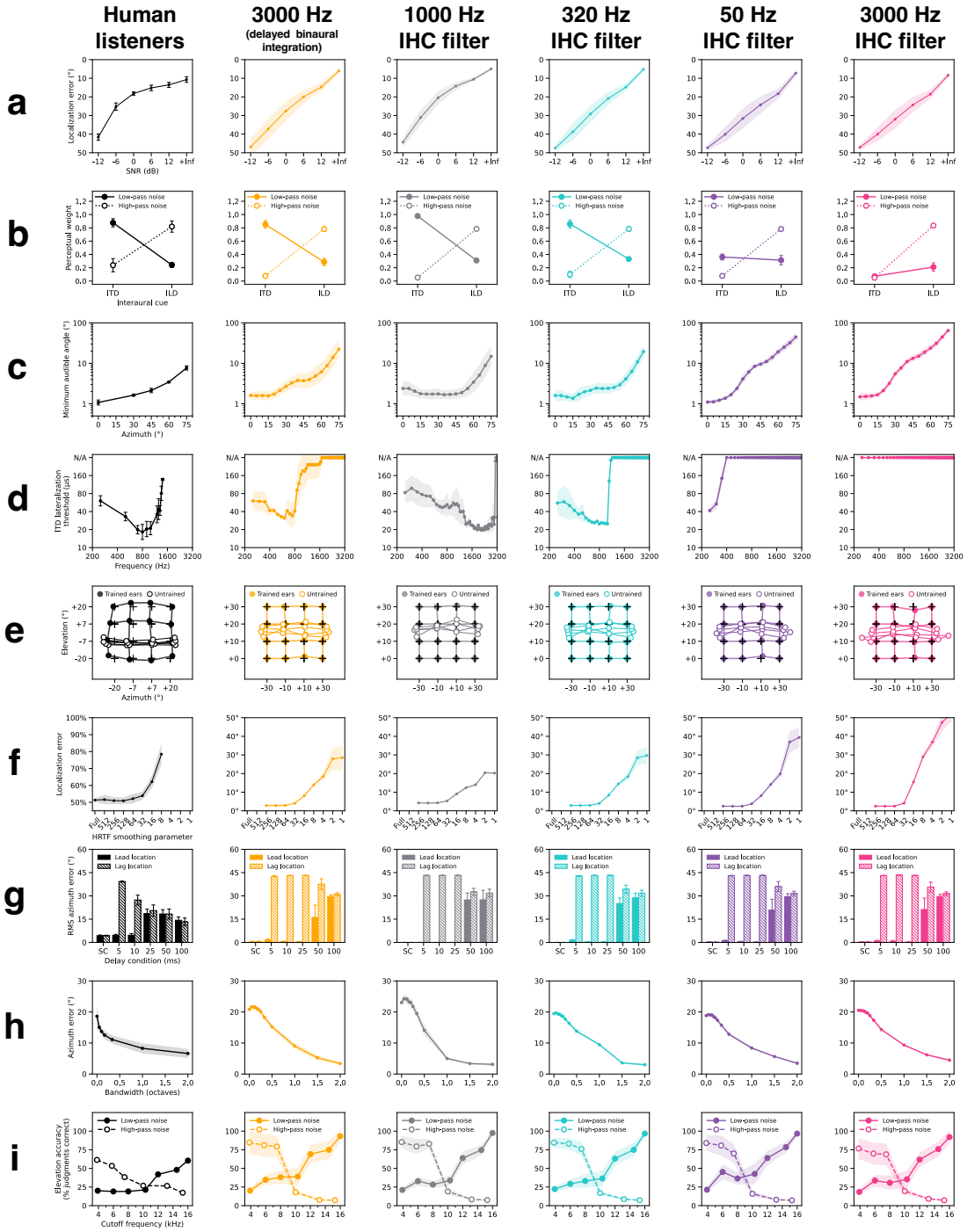

Figure 3.11: Simplified cochlear model – effect of phase locking manipulation on all localization experiments.

Columns correspond to human listeners and models optimized with different phase locking limits. The second (orange) column corresponds to the 3000 Hz phase locking model with network architectures modified to delay binaural integration. Rows correspond to 9 different sound localization experiments. a. Sound localization in noise. b. Minimum audible angle vs. frequency. c. ITD / ILD cue weighting. d. ITD lateralization vs. frequency. e. Effect of changing ears. f. Effect of smoothing spectral cues. g. Precedence effect. h. Bandwidth dependency of localization. i. Median plane spectral cues. All model error bars indicate ±2 standard errors of the mean across 10 network architectures. Human data in b-i is re-plotted from the original studies [132]–[135], [146], [153], [154], [171].

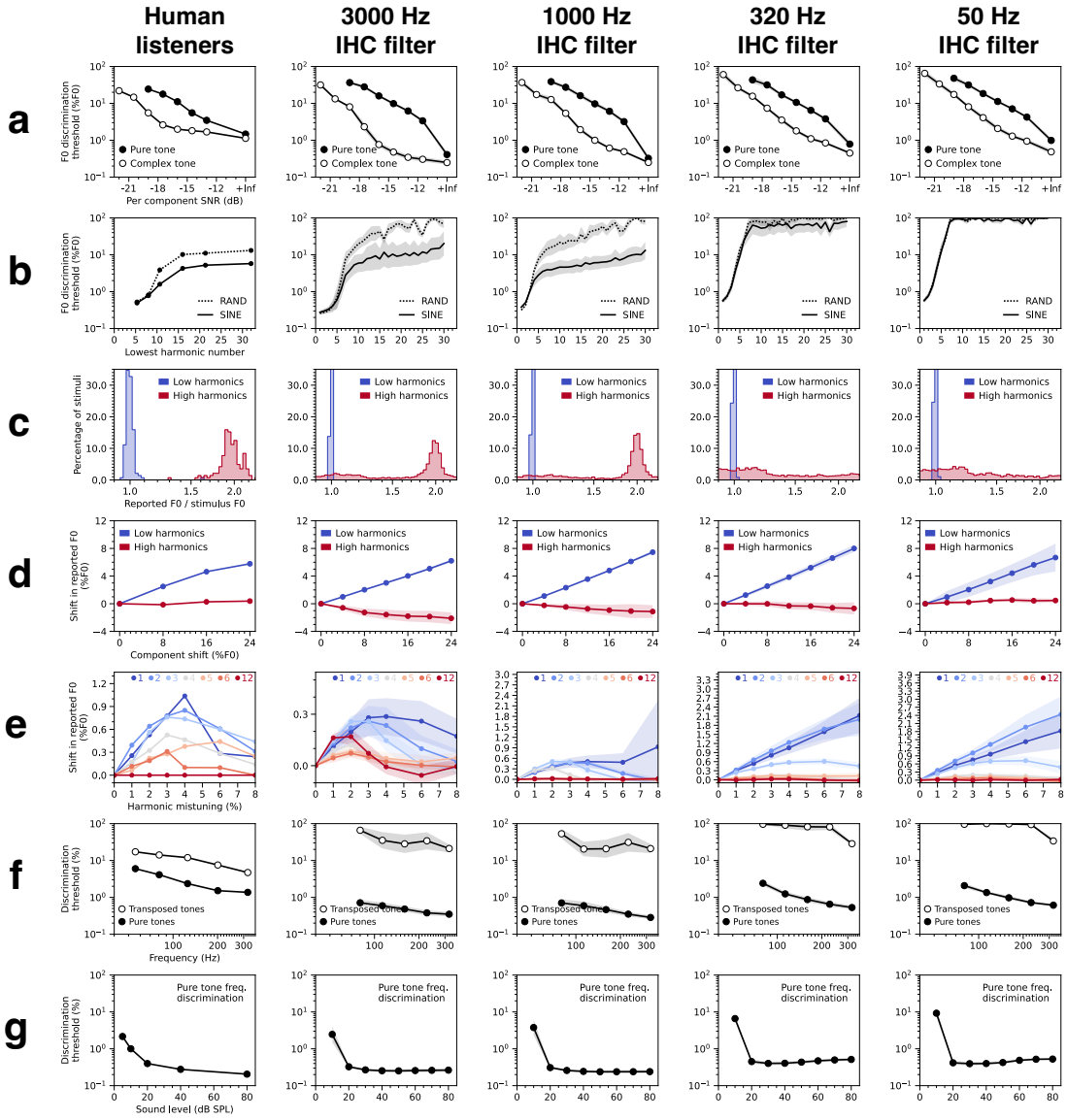

Figure 3.12: Simplified cochlear model – effect of phase locking manipulation on all pitch experiments.

Columns correspond to human listeners and models optimized with different phase locking limits. Rows correspond to 7 different pitch perception experiments. a. Pitch discrimination with pure and complex tones in noise. b. Effect of harmonic number and phase on pitch discrimination. c. Pitch of alternating-phase harmonic complexes. d. Pitch of frequency-shifted complexes. e. Pitch of complexes with individually mistuned harmonics. f. Pitch discrimination with pure and transposed tones. g. Effect of level on pure tone frequency discrimination. All model error bars indicate ±2 standard errors of the mean across 10 network architectures. Human data is re-plotted from the original studies [59]–[63], [82], [89].

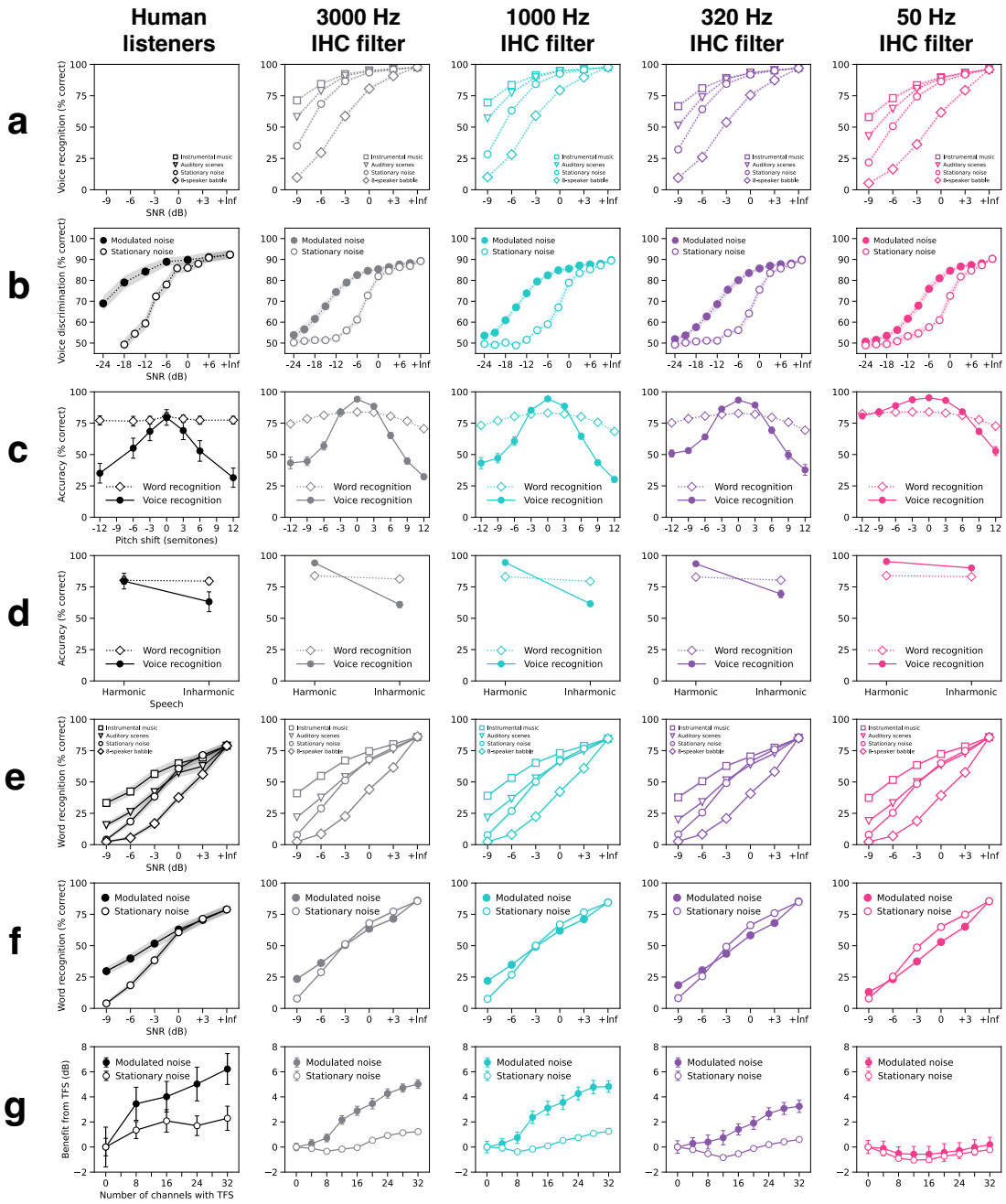

Figure 3.13: Simplified cochlear model – effect of phase locking manipulation on all speech experiments.

Columns correspond to human listeners and models optimized with different phase locking limits. Rows correspond to 7 speech experiments. a. Voice recognition in real-world noise conditions (model only experiment). b. Voice discrimination in stationary and modulated noise. c. Voice and word recognition with pitch-shifted speech. d. Voice and word recognition with harmonic and inharmonic speech. e. Word recognition in real-world noise conditions. f. Word recognition in stationary and modulated noise. g. Effect of tone vocoding on word recognition in stationary and modulated noise. All model error bars indicate ±2 standard errors of the mean across 10 network architectures. Human data in g is re-plotted from the original study [36].

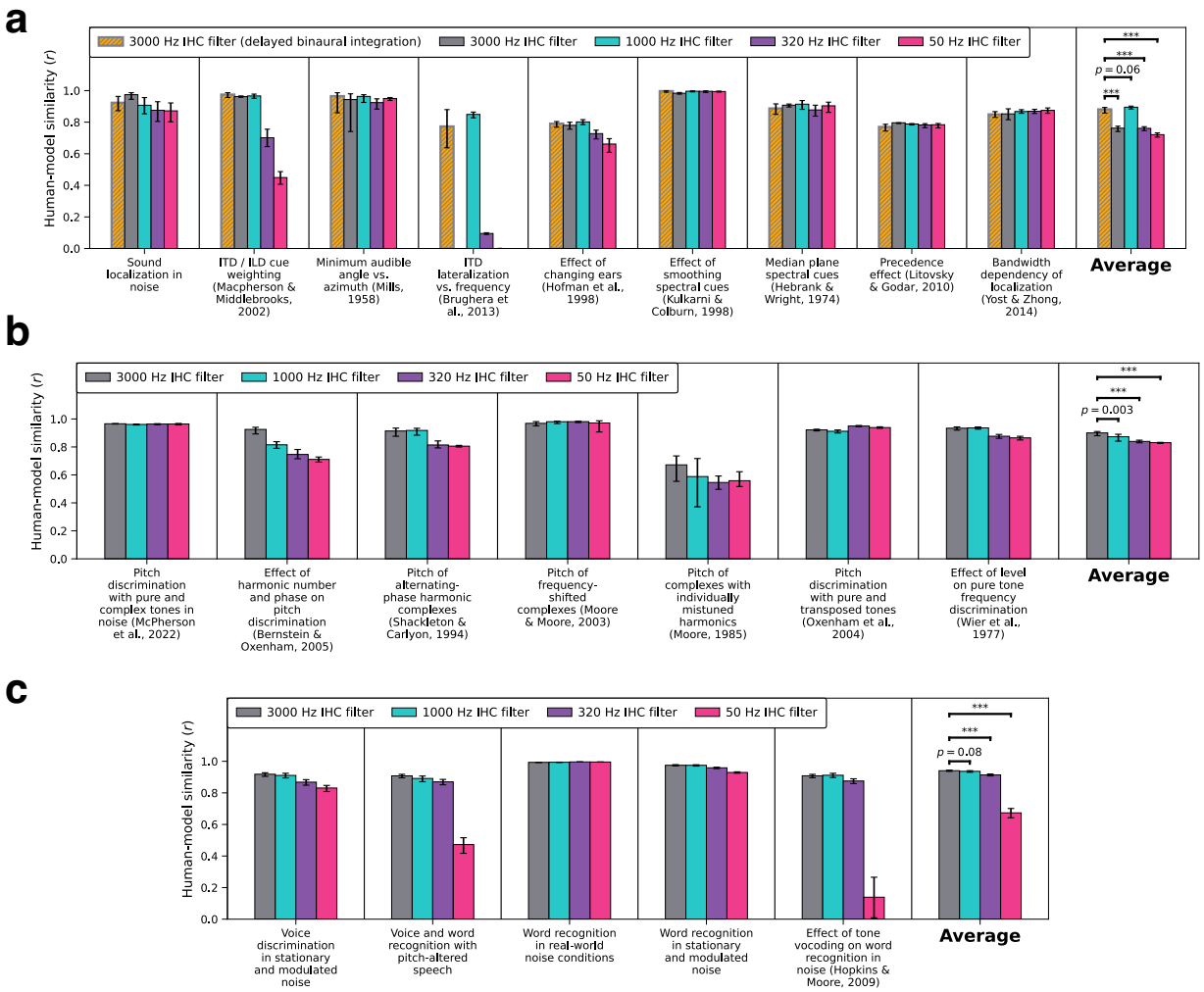

Figure 3.14: Simplified cochlear model – human-model similarity scores.

Analogous to Fig. 3.4f, Fig. 3.5f, and Fig. 3.7f in the main text but for models operating on simplified cochlear input representations. Human-model similarity scores for each phase locking condition are plotted for every (a.) localization, (b.) pitch, and (c.) speech experiment. Similarity scores were Pearson correlation coefficients between corresponding human and model behavioral data points. The rightmost columns present similarity scores averaged across all experiments within the same model. Error bars indicate 95% confidence intervals computed by bootstrapping the mean of 10 network architectures. Asterisks denote statistically significant differences between phase locking conditions ($p < 0.001$, two-tailed), evaluated by comparing mean similarity scores against null distributions bootstrapped from the 3000 Hz conditions.

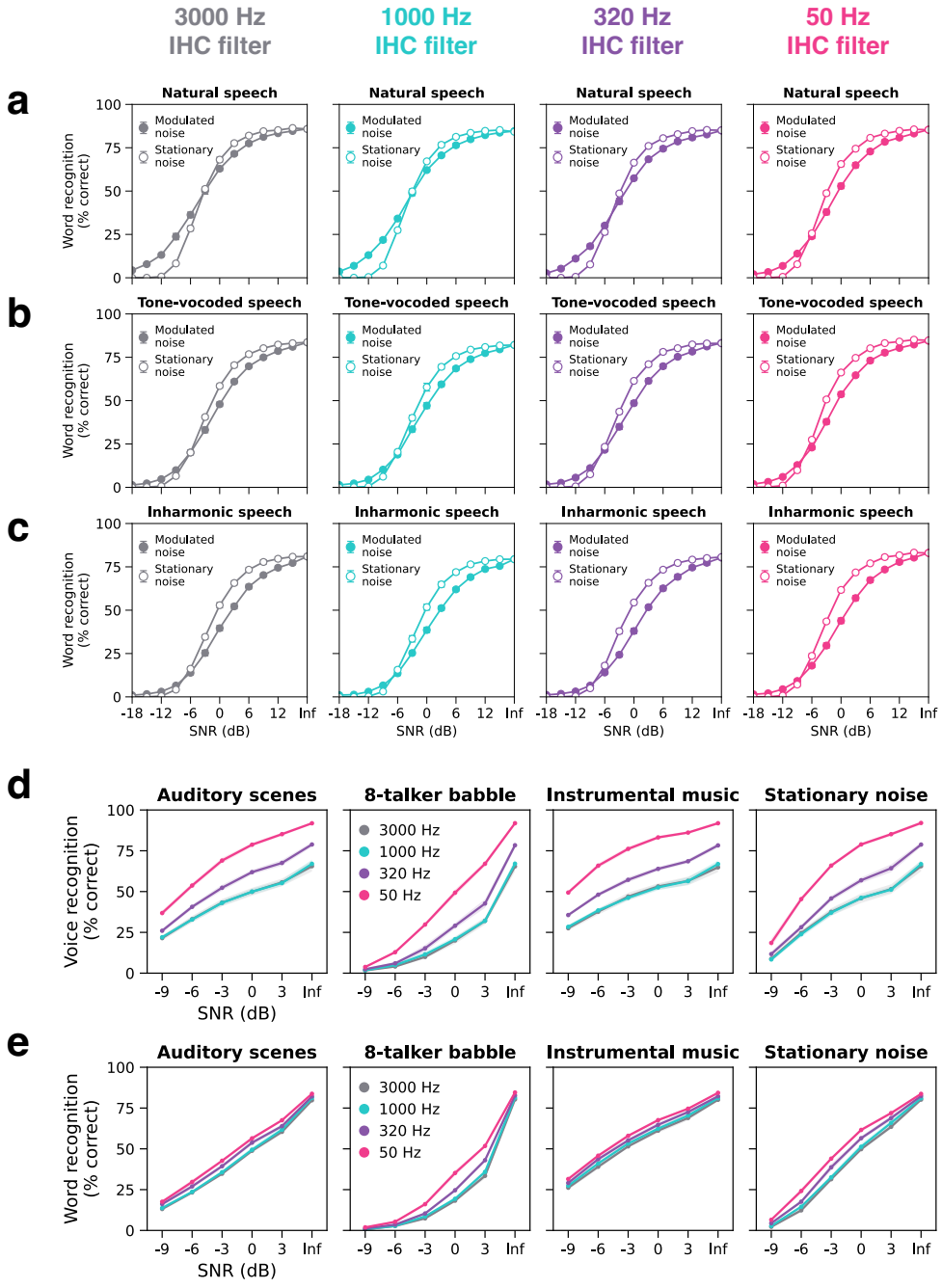

Figure 3.15: Model voice and word recognition with inharmonic speech in noise.

To directly compare effects of the inharmonicity and tone vocoding stimulus manipulations on model word recognition in noise, we measured word recognition accuracy in stationary and modulated speech-shaped noise at SNRs between -18 and +15 dB in 3 dB increments using (a.) natural, (b.) tone-vocoded, and (c.) inharmonic versions of the same speech. The tone-vocoded speech was fully vocoded (0 channels with intact TFS). d. Model voice recognition with inharmonic speech as a function of SNR in four different types of real-world noise. e. Model word recognition with inharmonic speech as a function of SNR in four different types of real-world noise. All error bars indicate ±2 standard errors of the mean across 10 network architectures.

Architecture	arch_01	arch_02	arch_03	arch_04	arch_05	arch_06	arch_07	arch_08	arch_09	arch_10
Operation	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]
1	conv0 [1, 8, 32]	conv0 [2, 8, 32]	conv0 [1, 4, 32]	conv0 [3, 8, 32]	conv0 [2, 32, 32]	conv0 [1, 64, 32]	conv0 [1, 16, 32]	conv0 [1, 64, 32]	conv0 [3, 32, 32]	conv0 [2, 4, 32]
2	mpool0 [1, 1]	mpool0 [1, 1]	mpool0 [1, 1]	mpool0 [1, 1]	mpool0 [1, 2]	mpool0 [1, 8]	mpool0 [1, 1]	mpool0 [1, 1]	mpool0 [1, 1]	mpool0 [2, 2]
3	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0
4	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0
5	conv1 [1, 64, 32]	conv1 [3, 16, 32]	conv1 [3, 32, 32]	conv1 [3, 8, 32]	conv1 [1, 4, 64]	conv1 [2, 4, 64]	conv1 [1, 8, 32]	conv1 [2, 16, 32]	conv1 [2, 16, 32]	conv1 [2, 4, 32]
6	mpool1 [1, 1]	mpool1 [1, 1]	mpool1 [1, 8]	mpool1 [1, 2]	mpool1 [1, 4]	mpool1 [1, 1]	mpool1 [1, 2]	mpool1 [1, 8]	mpool1 [1, 4]	mpool1 [1, 4]
7	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1
8	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1
9	conv2 [1, 64, 32]	conv2 [2, 4, 32]	conv2 [3, 32, 64]	conv2 [1, 32, 64]	conv2 [3, 2, 64]	conv2 [1, 32, 64]	conv2 [2, 4, 64]	conv2 [2, 4, 64]	conv2 [2, 32, 64]	conv2 [3, 16, 64]
10	mpool2 [1, 8]	mpool2 [1, 8]	mpool2 [1, 1]	mpool2 [1, 1]	mpool2 [1, 1]	mpool2 [2, 4]	mpool2 [1, 1]	mpool2 [1, 1]	mpool2 [1, 1]	mpool2 [1, 2]
11	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2
12	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2
13	conv3 [2, 4, 64]	conv3 [3, 16, 64]	conv3 [1, 8, 64]	conv3 [3, 8, 64]	conv3 [2, 8, 64]	conv3 [3, 4, 128]	conv3 [2, 32, 64]	conv3 [2, 16, 64]	conv3 [3, 4, 64]	conv3 [1, 2, 128]
14	mpool3 [2, 4]	mpool3 [1, 1]	mpool3 [1, 4]	mpool3 [2, 4]	mpool3 [1, 1]	mpool3 [1, 1]	mpool3 [1, 4]	mpool3 [1, 1]	mpool3 [1, 4]	mpool3 [1, 2]
15	relu3	relu3	relu3	relu3	relu3	relu3	relu3	relu3	relu3	relu3
16	bnorm3	bnorm3	bnorm3	bnorm3	bnorm3	bnorm3	bnorm3	bnorm3	bnorm3	bnorm3
17	conv4 [3, 8, 128]	conv4 [1, 8, 64]	conv4 [3, 8, 64]	conv4 [2, 2, 128]	conv4 [1, 16, 64]	conv4 [2, 16, 128]	conv4 [3, 2, 64]	conv4 [1, 16, 64]	conv4 [3, 8, 128]	flatten
18	mpool4 [1, 1]	mpool4 [1, 4]	mpool4 [1, 1]	mpool4 [1, 4]	mpool4 [1, 4]	mpool4 [1, 2]	mpool4 [1, 1]	mpool4 [1, 2]	mpool4 [1, 4]	fc0 [512]
19	relu4	relu4	relu4	relu4	relu4	relu4	relu4	relu4	relu4	relu_fc0
20	bnorm4	bnorm4	bnorm4	bnorm4	bnorm4	bnorm4	bnorm4	bnorm4	bnorm4	norm_fc0
21	conv5 [3, 32, 128]	conv5 [3, 8, 128]	conv5 [1, 2, 64]	conv5 [1, 4, 256]	conv5 [3, 4, 128]	conv5 [1, 2, 256]	conv5 [1, 2, 64]	conv5 [2, 32, 128]	conv5 [3, 2, 256]	dropout
22	mpool5 [1, 4]	mpool5 [1, 4]	mpool5 [1, 1]	mpool5 [1, 1]	mpool5 [1, 2]	mpool5 [1, 1]	mpool5 [2, 4]	mpool5 [1, 4]	mpool5 [1, 2]	fc [504]
23	relu5	relu5	relu5	relu5	relu5	relu5	relu5	relu5	relu5	
24	bnorm5	bnorm5	bnorm5	bnorm5	bnorm5	bnorm5	bnorm5	bnorm5	bnorm5	
25	conv6 [3, 4, 256]	conv6 [2, 2, 128]	conv6 [2, 2, 64]	conv6 [3, 2, 256]	conv6 [3, 4, 256]	conv6 [3, 4, 256]	conv6 [1, 8, 128]	conv6 [2, 16, 128]	conv6 [2, 8, 512]	
26	mpool6 [1, 1]	mpool6 [1, 2]	mpool6 [2, 4]	mpool6 [1, 1]	mpool6 [1, 1]	mpool6 [1, 2]	mpool6 [1, 1]	mpool6 [1, 1]	mpool6 [1, 1]	
27	relu6	relu6	relu6	relu6	relu6	relu6	relu6	relu6	relu6	
28	bnorm6	bnorm6	bnorm6	bnorm6	bnorm6	bnorm6	bnorm6	bnorm6	bnorm6	
29	conv7 [3, 8, 256]	conv7 [3, 2, 256]	conv7 [2, 4, 128]	conv7 [2, 2, 256]	conv7 [3, 4, 256]	flatten	flatten	conv7 [1, 2, 128]	conv7 [3, 4, 512]	
30	mpool7 [1, 2]	mpool7 [1, 2]	mpool7 [1, 1]	mpool7 [1, 2]	mpool7 [1, 1]	fc0 [512]	fc0 [512]	mpool7 [1, 1]	mpool7 [1, 2]	
31	relu7	relu7	relu7	relu7	relu7	relu_fc0	relu_fc0	relu7	relu7	
32	bnorm7	bnorm7	bnorm7	bnorm7	bnorm7	norm_fc0	norm_fc0	bnorm7	bnorm7	
33	flatten	conv8 [1, 8, 512]	conv8 [1, 8, 128]	flatten	conv8 [2, 4, 256]	dropout	dropout	conv8 [3, 16, 128]	conv8 [1, 3, 512]	
34	fc0 [512]	mpool8 [1, 2]	mpool8 [1, 1]	fc0 [512]	mpool8 [1, 2]	fc [504]	fc [504]	mpool8 [1, 4]	mpool8 [1, 1]	
35	relu_fc0	relu8	relu8	relu_fc0	relu8			relu8	relu8	
36	norm_fc0	bnorm8	bnorm8	norm_fc0	bnorm8			bnorm8	bnorm8	
37	dropout	flatten	conv9 [3, 2, 128]	dropout	flatten			flatten	flatten	
38	fc [504]	fc0 [512]	mpool9 [1, 4]	fc [504]	fc0 [512]			fc0 [512]	fc0 [512]	
39		relu_fc0	relu9		relu_fc0			relu_fc0	relu_fc0	
40		norm_fc0	bnorm9		norm_fc0			norm_fc0	norm_fc0	
41		dropout	flatten		dropout			dropout	dropout	
42		fc [504]	fc0 [512]		fc [504]			fc [504]	fc [504]	
43			relu_fc0							
44			norm_fc0							
45			dropout							
46			fc [504]							
47										

Table 3.1: Neural network architectures for sound localization models.

Grey bands indicate blocks of convolution, pooling, nonlinear rectification, and normalization operations. The convolution operations highlighted in orange were replaced with grouped convolutions (2 groups for the left and right ear) when network architectures were modified to delay binaural integration. Legend:

- $conv[h, w, k]$: convolutional layer with h = kernel height (frequency dimension), w = kernel width (time dimension), and k = number of kernels
- $relu$: rectified linear unit activation function
- $mpool[s_f, s_t]$: max pooling operation with stride s_f in the frequency dimension and stride s_t in the time dimension
- $bnorm$: batch normalization operation
- $flatten$: multidimensional representation reshaped to a vector
- $fc[N]$: fully-connected layer with N units
- $dropout$: dropout regularization with 50% dropout rate

Architecture	arch_0083	arch_0154	arch_0190	arch_0191	arch_0286	arch_0288	arch_0302	arch_0335	arch_0338	arch_0346
Operation	input [100, 1000, 3]	input [100, 1000, 3]	input [100, 1000, 3]	input [100, 1000, 3]	input [100, 1000, 3]	input [100, 1000, 3]	input [100, 1000, 3]	input [100, 1000, 3]	input [100, 1000, 3]	input [100, 1000, 3]
1	conv0 [2, 82, 32]	conv0 [1, 70, 64]	conv0 [2, 97, 32]	conv0 [2, 83, 32]	conv0 [1, 180, 64]	conv0 [3, 77, 64]	conv0 [1, 250, 32]	conv0 [1, 71, 32]	conv0 [2, 110, 64]	conv0 [3, 53, 32]
2	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0
3	hpool0 [1, 2]	hpool0 [1, 5]	hpool0 [1, 6]	hpool0 [1, 2]	hpool0 [2, 6]	hpool0 [1, 2]	hpool0 [1, 5]	hpool0 [1, 1]	hpool0 [3, 2]	hpool0 [1, 2]
4	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0
5	conv1 [1, 162, 64]	conv1 [7, 21, 128]	conv1 [5, 11, 64]	conv1 [1, 164, 64]	conv1 [2, 37, 128]	conv1 [1, 193, 128]	conv1 [19, 11, 64]	conv1 [1, 114, 32]	conv1 [1, 126, 128]	conv1 [1, 60, 64]
6	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1
7	hpool1 [2, 2]	hpool1 [4, 3]	hpool1 [1, 1]	hpool1 [3, 7]	hpool1 [1, 1]	hpool1 [4, 3]	hpool1 [1, 7]	hpool1 [1, 3]	hpool1 [3, 3]	hpool1 [2, 4]
8	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1
9	conv2 [1, 72, 128]	conv2 [4, 26, 256]	conv2 [1, 56, 128]	conv2 [5, 9, 128]	conv2 [15, 10, 128]	conv2 [8, 10, 128]	conv2 [12, 9, 128]	conv2 [1, 86, 64]	conv2 [4, 30, 256]	conv2 [3, 46, 128]
10	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2
11	hpool2 [4, 2]	hpool2 [1, 6]	hpool2 [4, 7]	hpool2 [1, 7]	hpool2 [1, 1]	hpool2 [2, 6]	hpool2 [3, 1]	hpool2 [4, 1]	hpool2 [2, 5]	hpool2 [1, 6]
12	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2
13	conv3 [6, 3, 128]	conv3 [2, 1, 512]	conv3 [8, 5, 256]	conv3 [4, 3, 256]	flatten	conv3 [2, 2, 256]	conv3 [7, 7, 256]	conv3 [13, 13, 128]	conv3 [1, 5, 256]	conv3 [8, 1, 256]
14	relu3	relu3	relu3	relu3	fc0 [512]	relu3	relu3	relu3	relu3	relu3
15	hpool3 [2, 5]	hpool3 [2, 1]	hpool3 [1, 2]	hpool3 [2, 1]	relu_fc0	hpool3 [2, 2]	hpool3 [1, 1]	hpool3 [1, 8]	hpool3 [1, 3]	hpool3 [2, 2]
16	bnorm3	bnorm3	bnorm3	bnorm3	norm_fc0	bnorm3	bnorm3	bnorm3	bnorm3	bnorm3
17	flatten	conv4 [5, 3, 256]	conv4 [1, 3, 256]	conv4 [5, 2, 512]	dropout	conv4 [2, 1, 512]	conv4 [5, 3, 512]	conv4 [2, 10, 256]	conv4 [2, 2, 512]	conv4 [7, 2, 256]
18	fc0 [128]	relu4	relu4	relu4	fc [700]	relu4	relu4	relu4	relu4	relu4
19	relu_fc0	hpool4 [1, 1]	hpool4 [2, 1]	hpool4 [1, 1]		hpool4 [1, 1]	hpool4 [3, 1]	hpool4 [1, 3]	hpool4 [1, 1]	hpool4 [1, 1]
20	norm_fc0	bnorm4	bnorm4	bnorm4		bnorm4	bnorm4	bnorm4	bnorm4	bnorm4
21	dropout	conv5 [2, 2, 256]	flatten	flatten		conv5 [2, 4, 1024]	flatten	conv5 [3, 2, 512]	conv5 [1, 1, 1024]	conv5 [2, 2, 512]
22	fc [700]	relu5	dropout	fc0 [256]		relu5	fc0 [1024]	relu5	relu5	relu5
23		hpool5 [1, 1]	fc [700]	relu_fc0		hpool5 [1, 1]	relu_fc0	hpool5 [2, 1]	hpool5 [1, 1]	hpool5 [2, 1]
24		bnorm5		norm_fc0		bnorm5	norm_fc0	bnorm5	bnorm5	bnorm5
25		conv6 [3, 1, 256]		dropout		flatten	dropout	flatten	conv6 [1, 1, 1024]	conv6 [1, 1, 512]
26		relu6		fc [700]		fc0 [256]	fc [700]	dropout	relu6	relu6
27		hpool6 [1, 1]				relu_fc0		fc [700]	hpool6 [1, 1]	hpool6 [1, 1]
28		bnorm6				norm_fc0			bnorm6	bnorm6
29		flatten				dropout			flatten	flatten
30		dropout				fc [700]			fc0 [256]	fc0 [512]
31		fc [700]							relu_fc0	relu_fc0
32									norm_fc0	norm_fc0
33									dropout	dropout
34									fc [700]	fc [700]
35										

Table 3.2: Neural network architectures for pitch models.

Grey bands indicate blocks of convolution, pooling, nonlinear rectification, and normalization operations. Legend:

- $conv[h, w, k]$: convolutional layer with h = kernel height (frequency dimension), w = kernel width (time dimension), and k = number of kernels
- $relu$: rectified linear unit activation function
- $hpool[s_f, s_t]$: Hanning window weighted average pooling operation with stride s_f in the frequency dimension and stride s_t in the time dimension
- $bnorm$: batch normalization operation
- $flatten$: multidimensional representation reshaped to a vector
- $fc[N]$: fully-connected layer with N units
- $dropout$: dropout regularization with 50% dropout rate

Architecture	arch0_0000	arch0_0001	arch0_0002	arch0_0004	arch0_0006	arch0_0007	arch0_0008	arch0_0009	arch0_0016	arch0_0017
Operation	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]
1	input_lnorm	input_lnorm	input_lnorm	input_lnorm	input_lnorm	input_lnorm	input_lnorm	input_lnorm	input_lnorm	input_lnorm
2	conv0 [2, 42, 32]	conv0 [1, 84, 32]	conv0 [4, 21, 32]	conv0 [2, 42, 32]	conv0 [2, 42, 32]	conv0 [2, 42, 32]	conv0 [2, 42, 32]	conv0 [2, 42, 32]	conv0 [2, 42, 32]	conv0 [2, 42, 32]
3	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0
4	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]
5	lnorm0	lnorm0	lnorm0	lnorm0	lnorm0	lnorm0	lnorm0	lnorm0	lnorm0	lnorm0
6	conv1 [2, 18, 64]	conv1 [2, 18, 64]	conv1 [2, 18, 64]	conv1 [4, 9, 64]	conv1 [2, 18, 64]	conv1 [2, 18, 64]	conv1 [2, 18, 64]	conv1 [2, 18, 64]	conv1 [2, 18, 64]	conv1 [2, 18, 64]
7	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1
8	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]
9	lnorm1	lnorm1	lnorm1	lnorm1	lnorm1	lnorm1	lnorm1	lnorm1	lnorm1	lnorm1
10	conv2 [6, 6, 128]	conv2 [6, 6, 128]	conv2 [6, 6, 128]	conv2 [6, 6, 128]	conv2 [12, 3, 128]	conv2 [6, 6, 128]	conv2 [6, 6, 128]	conv2 [6, 6, 128]	conv2 [6, 6, 128]	conv2 [6, 6, 128]
11	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2
12	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]
13	lnorm2	lnorm2	lnorm2	lnorm2	lnorm2	lnorm2	lnorm2	lnorm2	lnorm2	lnorm2
14	conv3 [6, 6, 256]	conv3 [6, 6, 256]	conv3 [6, 6, 256]	conv3 [6, 6, 256]	conv3 [6, 6, 256]	conv3 [3, 12, 256]	conv3 [12, 3, 256]	conv3 [6, 6, 256]	conv3 [6, 6, 256]	conv3 [6, 6, 256]
15	relu3	relu3	relu3	relu3	relu3	relu3	relu3	relu3	relu3	relu3
16	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]
17	lnorm3	lnorm3	lnorm3	lnorm3	lnorm3	lnorm3	lnorm3	lnorm3	lnorm3	lnorm3
18	conv4 [8, 8, 512]	conv4 [8, 8, 512]	conv4 [8, 8, 512]	conv4 [8, 8, 512]	conv4 [8, 8, 512]	conv4 [8, 8, 512]	conv4 [8, 8, 512]	conv4 [4, 16, 512]	conv4 [8, 8, 512]	conv4 [8, 8, 512]
19	relu4	relu4	relu4	relu4	relu4	relu4	relu4	relu4	relu4	relu4
20	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]
21	lnorm4	lnorm4	lnorm4	lnorm4	lnorm4	lnorm4	lnorm4	lnorm4	lnorm4	lnorm4
22	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]
23	relu5	relu5	relu5	relu5	relu5	relu5	relu5	relu5	relu5	relu5
24	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]
25	lnorm5	lnorm5	lnorm5	lnorm5	lnorm5	lnorm5	lnorm5	lnorm5	lnorm5	lnorm5
26	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]
27	relu6	relu6	relu6	relu6	relu6	relu6	relu6	relu6	relu6	relu6
28	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]
29	lnorm6	lnorm6	lnorm6	lnorm6	lnorm6	lnorm6	lnorm6	lnorm6	lnorm6	lnorm6
30	flatten	flatten	flatten	flatten	flatten	flatten	flatten	flatten	conv7 [2, 8, 512]	conv7 [8, 2, 512]
31	fc0 [512]	fc0 [512]	fc0 [512]	fc0 [512]	fc0 [512]	fc0 [512]	fc0 [512]	fc0 [512]	relu7	relu7
32	relu_fc0	relu_fc0	relu_fc0	relu_fc0	relu_fc0	relu_fc0	relu_fc0	relu_fc0	hpool7 [1, 1]	hpool7 [1, 1]
33	norm_fc0	norm_fc0	norm_fc0	norm_fc0	norm_fc0	norm_fc0	norm_fc0	norm_fc0	lnorm7	lnorm7
34	dropout	dropout	dropout	dropout	dropout	dropout	dropout	dropout	flatten	flatten
35	fc [433, 794]	fc [433, 794]	fc [433, 794]	fc [433, 794]	fc [433, 794]	fc [433, 794]	fc [433, 794]	fc [433, 794]	fc0 [512]	fc0 [512]
36									relu_fc0	relu_fc0
37									norm_fc0	norm_fc0
38									dropout	dropout
39									fc [433, 794]	fc [433, 794]
40										

Table 3.3: Neural network architectures for voice and word recognition models.

Grey bands indicate blocks of convolution, pooling, nonlinear rectification, and normalization operations. Legend:

- $conv[h, w, k]$: convolutional layer with h = kernel height (frequency dimension), w = kernel width (time dimension), and k = number of kernels
- $relu$: rectified linear unit activation function
- $hpool[s_f, s_t]$: Hanning window weighted average pooling operation with stride s_f in the frequency dimension and stride s_t in the time dimension
- $lnorm$: layer normalization operation
- $flatten$: multidimensional representation reshaped to a vector
- $fc[N]$: fully-connected layer with N units
- $fc[N_{voice}, N_{word}]$: two parallel fully-connected layers operating on the same input, one with N_voice units and one with N_word units
- $dropout$: dropout regularization with 50% dropout rate

Chapter 4

Speech denoising with auditory models

Mark R. Saddler, Andrew Francl, Jenelle Feather, Kaizhi Quan, Yang Zhang, Josh H. McDermott

Abstract

Contemporary speech enhancement predominantly relies on audio transforms that are trained to reconstruct a clean speech waveform. The development of high-performing neural network sound recognition systems has raised the possibility of using deep feature representations as 'perceptual' losses with which to train denoising systems. We explored their utility by first training deep neural networks to classify either spoken words or environmental sounds from audio. We then trained an audio transform to map noisy speech to an audio waveform that minimized the difference in the deep feature representations between the output audio and the corresponding clean audio. The resulting transforms removed noise substantially better than baseline methods trained to reconstruct clean waveforms, and also outperformed previous methods using deep feature losses. However, a similar benefit was obtained simply by using losses derived from the filter bank inputs to the deep networks. The results show that deep features can guide speech enhancement, but suggest that they do not yet outperform simple alternatives that do not involve learned features.

Index Terms: speech enhancement, denoising, deep neural networks, cochlear model, perceptual metrics

4.1 Introduction

Recent advances in speech enhancement have been driven by neural network models trained to reconstruct speech sample-by-sample [38]–[45]. These methods provide substantial benefits over previous approaches, but nonetheless leave room for improvement. The resulting processed speech usually contains audible artifacts, and noise removal is usually incomplete at lower SNRs.

A parallel line of work has explored the use of deep artificial neural networks as models of sensory systems [19], [57]. Although substantial discrepancies remain [15], [93], such trained neural networks currently provide the best predictive models of brain responses and behavior in both the visual and auditory systems [13], [19]. The apparent similarities between deep supervised feature representations and representations in the brain raises the possibility that such representations could be used as perceptual metrics. Such metrics have been successfully employed in image processing [172], but are not widely used in audio applications.

Deep feature losses for denoising were previously proposed in [46], [173]–[176], but were explored only for relatively high signal-to-noise ratios (SNRs), a single task and network, or were not compared to baseline methods using the same transform architecture. Additionally, direct comparisons have not been made to simpler losses derived from conventional filter banks. It was thus unclear the extent to which deep feature losses could improve on simpler approaches, and what choices in the feature training would produce the best results. The goal of this paper was to directly compare deep perceptual losses to alternative losses, and to explore the conditions in which benefits might be achieved. We found that deep feature losses produced more natural denoising compared to waveform losses, but that a similar benefit could be achieved using a loss derived from standard filter bank representations.

4.2 Methods

There were two components to our denoising approach (Figure 4.1). The first component was a recognition network trained to recognize either speech or environmental sounds. Once trained, this network was used to define deep feature losses. Speech recognition is a natural choice in this context, but it also seemed plausible that more general-purpose audio features learned for environmental sound recognition might help to achieve natural-sounding audio even in speech applications. The input to the network was the output of a filter bank modeled on the human cochlea.

The second component was a waveform-to-waveform audio transform whose parameters were adjusted via gradient descent to minimize a loss function (evaluated on features of the

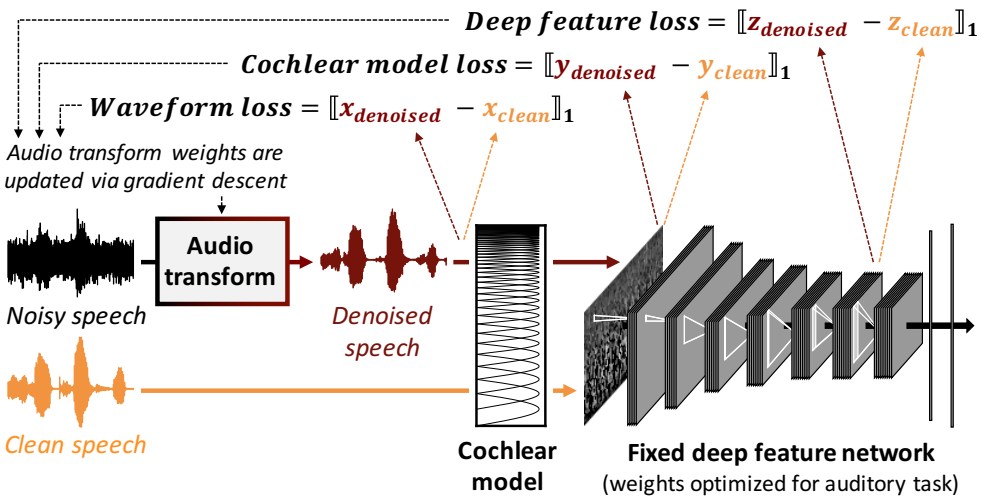

Figure 4.1: Schematic of audio transform training.

recognition network, or the outputs of a filter bank, or on the waveform). We used a Wave-U-Net [177], which has been found to perform comparably to WaveNet [178] based on objective metrics of noise reduction, but which can be specified with many fewer parameters and run with a much lower memory footprint. Code, models, and audio examples are available at: https://mcdermottlab.mit.edu/denoising/demo.html.

4.2.1 Recognition networks

The recognition networks took as input simulated cochlear representations of 2s sound clips (audio sampled at 20 kHz). The cochlear model consisted of a bank of 40 bandpass filters whose frequency tuning mimics that of the human ear (evenly spaced on an Equivalent Rectangular Bandwidth scale [112]), followed by half-wave rectification, downsampling to 10 kHz, and 0.3 power compression [108].

Recognition network architectures

We used three feed-forward CNN architectures for the recognition networks. Each consisted of stages of convolution, rectification, batch normalization, and weighted average pooling with a hanning kernel to minimize aliasing [93], [179]. The three architectures were selected based on word recognition task performance from 3097 randomly-generated architectures varying in number of convolutional layers (from 4 to 8), size and shape of convolutional kernels, and extent of pooling. The selected architectures had 6 (arch1) or 7 (arch2,3) convolutional layers.

Recognition network training

The recognition networks were trained to perform either word recognition or environmental sound recognition. For the speech task, each training example was a speech excerpt (from the Wall Street Journal [180] or Spoken Wikipedia Corpora [105]). The task was to recognize the word overlapping with the center of the clip [13], [93] (out of 793 word classes sourced from 432 unique speakers, with 230,357 unique clips in the training set and 40,651 segments in the validation set). For the environmental sound recognition task, each training example was a non-speech YouTube soundtrack excerpt (from a subset of 718,625 AudioSet examples [107]), and the task was to predict the AudioSet labels (spanning 516 categories in our dataset).

The three network architectures were trained on each task until performance on the validation set task plateaued. Word task classification accuracies for the three architectures were: arch1 = 90.4%, arch2 = 88.5%, and arch3 = 80.6%. AudioSet task AUC values were: arch1 = 0.845, arch2 = 0.861, and arch3 = 0.869.

4.2.2 Audio transforms

Wave-U-Net architecture

The Wave-U-Net architecture was the same as in [178]: 12 layers in the contracting path, a 1-layer bottleneck, and 12 layers in the expanding path. All layers utilized 1D convolutions with learned filters and LeakyReLU activation functions. There were 24 filters in the first layer, and the number of filters increased by a factor of 2 with each successive layer prior to the bottleneck.

Deep feature Losses

The recognition networks were used to define a deep feature loss function as the $L1$ distance between network representations of noisy speech and clean speech. The total loss for a single recognition network and single training example was the sum of the $L1$ distances between the noisy speech and clean speech activations for each convolutional layer, weighted to approximately balance the contribution of each layer.

Cochlear model losses

We also trained transforms using losses derived from the cochlear model that provided input to the recognition network, as well as variants of the model that varied in i) the number of filters (5, 10, 20, 40, 80 and 160 filters, evenly spaced on an ERB-scale [112], with bandwidths scaled to tile the spectrum in all cases), ii) the dependence of filter bandwidth on frequency

(linearly-spaced and 'reversed', with broad low-frequency filters and narrow high-frequency filters, opposite to what is found in the ear), and iii) in their phase invariance (subband envelopes computed by lowpass-filtering the rectified subbands; cutoff of 100 Hz).

Wave-U-Net training

Out of concern that the audio transform might overfit to idiosyncrasies of any individual recognition network, we trained some transforms on losses computed simultaneously from an ensemble of three different networks (arch1,2,3), and some on just a single network (arch1).

In all cases the Wave-U-Net was trained on speech superimposed on non-speech AudioSet excerpts (the same corpora used to train the recognition networks) with SNR drawn uniformly from $[-20, +10]$ dB. AudioSet excerpts were used as the training 'noise' as they were highly varied and diverse. All Wave-U-Net models were trained with the ADAM optimizer for 600,000 steps (batch size=8, learning rate=10^{-4}).

Baselines

We used two baseline models, both trained to explicitly reconstruct clean speech waveforms from noisy speech waveforms drawn from the same training set described above. The first was a previously described WaveNet [42] and the second was the Wave-U-Net [178] used with the deep feature and filter losses.

We also compared our results to those of a previously published denoising transform trained with a deep feature loss [46], using both the pre-trained model made available by Germain et al. and a Wave-U-Net that we trained on our dataset using the feature loss from [46] (deep network features trained on the DCASE 2016 [181] environmental sound challenge).

4.2.3 Evaluation

We evaluated the trained models on 40 speech excerpts (from a separate validation set) superimposed separately on each of four types of noise signals: speech-shaped Gaussian noise, auditory scenes from the DCASE 2013 dataset [166], instrumental music from the Million Song Dataset [182], and 8-speaker babble made from public-domain audiobooks (librivox.org). These noise sources were chosen to be distinct from those in the training set, and to span a variety of noise types to assess the generality of the trained transforms.

Table 4.1: Experiment 1 results. Reported metrics are averaged across the five tested SNR levels. Higher is better for all metrics.

Model name	Loss function	Natural.	PESQ	STOI	SDR
Cochlear model (N=40; human)	40 ERB-spaced subbands	**4.43**	1.55	0.75	7.16
A123	AudioSet features (arch123)	**4.43**	1.66	0.77	4.06
A1+W1	AudioSet + Word features (arch1)	4.36	**1.68**	0.79	6.18
A123+W123	AudioSet + Word features (arch123)	4.33	1.67	0.77	4.18
A1	AudioSet features (arch1)	4.33	1.65	0.78	3.63
W123	Word features (archs123)	4.24	1.67	**0.79**	6.64
W1	Word features (arch1)	4.22	1.63	0.77	3.30
Random1	Random features (arch1)	3.91	1.57	0.78	5.64
Random123	Random features (arch123)	3.84	1.57	0.77	5.08
Germain DeepFeatures	DCASE features from [46]	3.83	1.47	0.77	6.72
Germain (pre-trained)	DCASE features from [46]	2.36	1.14	0.64	0.93
Waveform (Wave-U-Net)	Waveform	4.17	1.51	0.76	**7.35**
Waveform (WaveNet)	Waveform	3.72	1.40	0.75	6.00
Unprocessed input		2.67	1.15	0.70	0.21

Human perceptual evaluation and objective metrics

We evaluated the audio transforms by conducting perceptual experiments on Amazon Mechanical Turk. Participants first completed a screening task to help ensure that they were wearing headphones or earphones [183]. The participants who passed this screening task then rated the naturalness of a set of processed speech signals, presented seven at a time in a MUSHRA-like paradigm. Listeners could listen to each clip as many times as they wished and then gave each a numerical rating on a scale of 1-7. Listeners were provided with anchors corresponding to the ends of the rating scale (1 and 7). The anchor at the high end was always the original clean speech. The low-end anchor was 4-bit-quantized speech (an example of very high distortion). To help ensure that participants were using the scale as instructed, each experiment included 3 catch trials where two of the stimuli were the two anchors. In order to be included in the analysis, participants had to rate all instances of the high and low anchors as 7 and 1, respectively.

We ran two identically structured experiments to evaluate all of our audio transforms. Experiment 1 compared various deep feature losses to baselines and contained all of the conditions listed in Table 4.1. Experiment 2 compared losses derived from different cochlear filter banks and contained all of the conditions listed in Table 4.2. 54 and 105 participants met the inclusion criteria for Experiments 1 and 2, respectively.

We also used three standard objective measures for evaluation: perceptual evaluation of

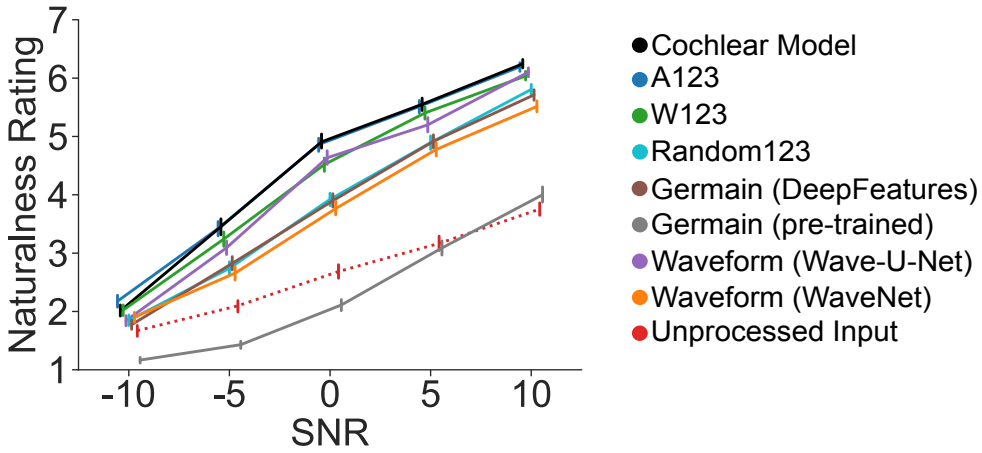

Figure 4.2: Rated naturalness vs. SNR for speech processed by Wave-U-Nets trained on deep feature losses, in addition to baseline models trained to reconstruct clean speech waveforms, and two versions of a related prior method [46]. Error bars plot SEM (across 54 participants).

speech quality (PESQ) [184], short-time objective intelligibility measure (STOI) [185], and the signal-to-distortion ratio (SDR) [186].

4.3 Results

4.3.1 Deep feature losses yield improved denoising

The best-performing systems trained with deep perceptual feature losses outperformed both waveform-based baselines. The average objective and subjective evaluation results are shown in Table 4.1. Human listeners found the speech processed by the deep feature models to be more natural than the speech processed by the baseline models. We plot the naturalness results in more detail (Figure 4.2) for two of the best-performing models trained on each of AudioSet features (A123) and word recognition features (W123), as well as a model trained on random features (Random123), the two baselines, and the two versions of the denoising network from [46].

4.3.2 Learned vs. random deep features

The benefit of deep feature losses was specific to models trained with learned features. Audio transforms trained to reconstruct random features did not produce better naturalism than the baseline WaveNet, and performed worse overall than the baseline Wave-U-Net (Figure 4.2; Table 4.1).

Table 4.2: Experiment 2 results. Reported metrics are averaged across the five tested SNR levels. Higher is better for all metrics.

Model name	Loss function	Natural.	PESQ	STOI	SDR
Cochlear model (N=20)	20 ERB-spaced subbands	**4.33**	1.54	0.77	**7.61**
Cochlear model (N=40; human)	40 ERB-spaced subbands	4.30	1.55	0.75	7.16
Cochlear model (N=160)	160 ERB-spaced subbands	4.26	1.60	0.77	7.51
Cochlear model (N=10)	10 ERB-spaced subbands	4.22	1.49	0.76	7.08
Cochlear model (N=80)	80 ERB-spaced subbands	4.21	1.53	0.74	6.69
Cochlear model (N=5)	5 ERB-spaced subbands	3.93	1.42	0.75	6.02
Cochlear model (N=40; linear)	40 linearly-spaced subbands	4.32	1.51	0.76	6.82
Cochlear model (N=40; env.)	Envelopes of 40 ERB subbands	4.16	1.59	0.75	6.94
Cochlear model (N=40; reverse)	40 reverse-ERB-spaced subbands	4.08	1.47	0.73	4.73
A123	AudioSet features (arch123)	4.27	**1.66**	**0.77**	4.06
Waveform (Wave-U-Net)	Waveform	4.17	1.51	0.76	7.35
Unprocessed input		2.47	1.15	0.70	0.21

4.3.3 Comparison to previous deep feature systems

Our best-performing deep feature-based systems also outperformed previously published systems with deep feature losses. The pre-trained system from Germain et al. [46] generalized poorly to our test set. Furthermore, the Wave-U-Net we trained using the deep feature loss from [46] also performed worse than the baseline Wave-U-Net. These findings suggest that the features used for the perceptual loss are important, and that the DCASE task used in [46] may not have produced sufficiently general features.

4.3.4 Effect of task used to train deep features

The best results occurred for features trained on the environmental sound recognition task – naturalism was consistently higher than for features trained on word recognition (Figure 4.2; Table 4.1). However, all of the models trained with feature losses from our recognition networks produced more natural-sounding speech than the baselines, and than the systems trained with DCASE features based on [46]. There was no obvious benefit from training on features from three different networks.

4.3.5 Cochlear model losses match deep feature losses

Although deep features produced better performance than baselines trained using waveform losses, we found that we could reproduce their benefit using losses derived from the cochlear model inputs to the recognition networks. Based on rated naturalness, the transform trained

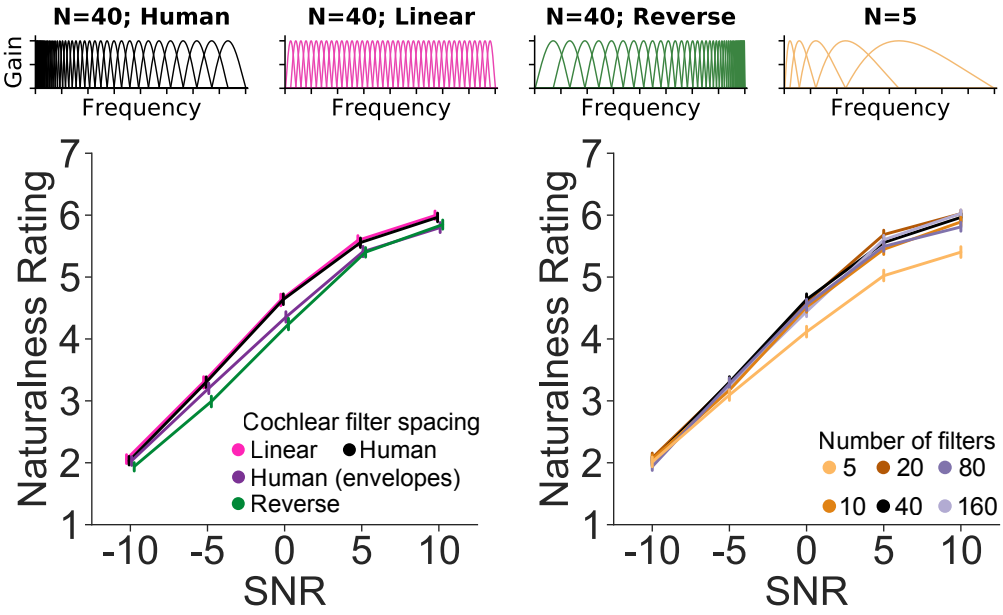

Figure 4.3: Rated naturalness vs. SNR for speech processed by Wave-U-Nets trained on cochlear model losses with different filter banks (select examples depicted above). Error bars plot SEM (across 105 participants).

with this 'cochlear' loss performed just as well as our best model trained with deep feature losses (Table 4.1).

4.3.6 Effect of filter bank characteristics

The benefit of the cochlear loss depended to some extent on the filter characteristics (Table 4.2; Figure 4.3, left). Worse performance was obtained with a 'reversed' filter bank, with wide filters at low frequencies and narrower filters at high frequencies, opposite to that of the ear. Using the envelope of the filter outputs also produced worse performance (counter to the hypothesis that phase invariance might be critical). However, filters that were linearly spaced along the frequency axis worked about as well as those modeled on the ear.

Worse performance was also obtained using only five filters (scaled to cover the frequency spectrum), but good results were obtained provided at least 10 filters were used (Figure 4.3, right). This result suggests that splitting the audio up into multiple frequency channels is sufficient to replicate the benefit of deep features provided there are enough channels with reasonably sensible frequency tuning.

4.3.7 Objective metrics

The models trained on deep recognition features also performed better than the baselines according to PESQ and STOI. Notably, this advantage was not evident when measured with SDR. The filter bank-trained models showed the opposite trend – better performance as measured by SDR, and worse via PESQ and STOI (Table 4.2). These differences suggest that the filter bank and deep feature losses are not fully interchangeable despite having similar effects on overall naturalness. The results also underscore the limitations of objective metrics for capturing human perception of altered speech.

4.4 Discussion

Prior work has proposed denoising based on deep feature losses [46], [173]–[176], but has not evaluated it relative to methods using simpler waveform- or subband-based losses. We found that deep recognition features could be used to train denoising systems that outperform waveform-based methods, but that their benefit could be matched using a standard one-layer auditory filter bank. The results thus provide no evidence that deep features provide a unique benefit for denoising.

Although deep neural networks yield the best current models of biological sensory systems [19], [57], our results indicate that these similarities are not yet sufficient to produce audio enhancement algorithms above and beyond what can be obtained from simple filter bank models. However, it is possible that building better models of human perceptual systems will also yield feature losses [187], [188] that would better transfer their perceptual benefits to humans, and produce benefits relative to simpler approaches. It also remains possible that the audio quality is limited more by the audio transform than the feature loss. More expressive transforms, or transforms with stronger generative constraints, might yield a clearer benefit of deep features.

The benefits of deep feature and cochlear model losses relative to waveform-based losses were clear from the ratings of human listeners, but were less evident in the objective metrics we tested (PESQ, STOI, SDR). This result indicates that optimizing for auditory model-based losses may provide perceptual benefits that conventional objective metrics are poorly suited to measuring, and suggests to the potential value of auditory model features as new objective metrics.

In sum, we found that audio transforms trained to modify noisy speech so as to reconstruct deep feature representations of clean speech produce better denoising performance than transforms trained to reconstruct clean speech waveforms, as measured by the ratings

of human listeners. However, a similar benefit was obtained using one-layer auditory filter banks, suggesting the importance of multi-channel, overcomplete representations rather than learned features per se.

4.5 Acknowledgements

The authors thank Ray Gonzalez for developing the training dataset, John Cohn and the Oak Ridge National Laboratory for use of Summit, and the MIT-IBM Watson AI Lab for funding.

Chapter 5

Conclusion

5.1 Contributions

We developed new models of hearing by combining detailed signal processing descriptions of the auditory periphery with deep artificial neural networks. When optimized for real-world hearing tasks, these models accounted for a broad array of human pitch perception, sound localization, voice recognition, and word recognition behavior. We used these models to investigate the underpinnings of human behavior, identifying environmental and peripheral neural coding factors that constrain auditory perception.

In Chapter 2, we introduced this modeling approach in the domain of pitch perception by optimizing networks to estimate F0 from speech and music in noise. The resulting pitch model innovated on prior work in two key regards. First, the model was optimized for naturalistic stimuli and listening conditions. By contrast, most prior models instantiate mechanistic hypotheses for how pitch information might be extracted from specific cues in the periphery [24], [63], [67]–[72]. Our model had access to all the cues available in the periphery and was free to use whichever combination of cues maximized task performance. Our optimization approach allowed us to explicitly test how human pitch computations may be adapted to the constraints of natural sounds. Second, our model produced quantitative matches to human behavior across a suite of classic psychoacoustic experiments. Importantly, the model was never trained on the psychoacoustic stimuli or fit to human data in any way. The characteristics of human pitch perception simply emerged when the model was optimized to estimate F0 from natural sounds in natural listening conditions heard through a human-like cochlea. To probe the origins of these characteristics, we optimized models with altered environmental and neural coding constraints. The results suggest pitch perception is critically shaped by the constraints of natural environments in addition to those of the cochlea.

In Chapter 3, we extended the approach to investigate the widely debated role of precise auditory nerve spike timing across hearing more broadly. Models were optimized to perform real-world hearing tasks using simulated auditory nerve input with different phase locking limits. Models with high phase locking limits replicated human behavior across all 21 experiments considered. Models with lower phase locking limits produced less human-like behavior, but some tasks were more sensitive than others. Degraded phase locking impaired sound localization, pitch perception, and voice recognition but left word recognition largely intact. The results suggest the information encoded in precise auditory nerve spike timing contributes to perception and therefore must be extracted in the auditory system. These findings place strong constraints on future models of the auditory system and clarify conditions in which prostheses that fail to restore high-fidelity temporal coding (e.g., contemporary cochlear implants) could in principle restore near-normal hearing.

More generally, these studies illustrate how highly expressive and optimizable models from deep learning can enable normative analysis in domains where traditional ideal observers are intractable. Models optimized for perceptual tasks under different constraints can provide insights into why human behavior exhibits certain characteristics [57], [189].

Our approach draws inspiration from classic ideal observer theory but differs in the important regard that current deep optimization methods provide no guarantee of optimality. Comparisons of absolute model performance should thus be interpreted with caution. Does a model optimized with altered neural coding fail to achieve human-level performance because a critical cue is disrupted or simply because the model is insufficiently optimized? Without a provably optimal task solution, the second possibility is theoretically impossible to rule out. In practice, we hedged against this possibility by ensuring conclusions generalized across neural network architectures and by comparing model behavior across many experiments. Simulating suites of psychophysical experiments probing different aspects of human perception allowed us to identify qualitative differences between human and model perceptual strategies in addition to any quantitative differences in overall task performance.

This thesis argues that foregoing provably optimal solutions has an invaluable upside: the ability to analyze real-world tasks. There is little reason to believe human perceptual systems are optimized for the simple tasks and stimuli for which provably optimal solutions can be derived. More plausibly, human perceptual systems are optimized for important tasks in the natural environment [18]. Our results suggest that many of the psychophysical characteristics of human hearing can be understood as consequences of this optimization.

5.2 Future directions

The general approach of this thesis was to relate environmental and peripheral neural coding features to perception by comparing models in terms of behavioral similarity to humans. The power of our approach is tied to the sensitivity of our behavioral metrics. We aimed to evaluate human-model similarity across as many psychoacoustic experiments as we could plausibly simulate in our networks. In addition to simply increasing the number of experiments, future work should increase the power to adjudicate between models by making finer-grained behavioral comparisons within the same experiments [15]. Rather than comparing overall task performance, finer-grained behavioral judgments like word and voice-level confusions or localization biases could be compared at the level of stimuli or individual trials. Finer-grained behavioral comparisons may elucidate links between progressively finer-grained details of neural coding and perception.

Despite their evocative name, the artificial neural networks in this work share little in common with biological neural networks in the brain. Under our approach, artificial neural networks may be viewed as just one class of highly expressive mathematical functions that can be optimized to perform tasks. At some level, our networks are no more a mechanistic hypothesis of brain function than the set of equations specifying an ideal observer are. However, this view should be tempered when considering the reality that deep artificial neural networks, despite their implausible learning rules and coarse approximations of biology, are the current state of the art for predicting brain responses in sensory systems [20]. Model manipulations that better align network and brain representations may in turn improve behavioral predictions. We saw evidence for this in Chapter 3, where incorporating a simple architectural constraint that made networks more brain-like (requiring localization networks to initially process inputs from the two ears separately) yielded a more human-like task solution. Additional constraints that future models could incorporate include more plausible learning rules [190], biological resource limitations [191], and joint optimization for many tasks.

Our approach has natural extensions to modeling hearing loss and cochlear implants. The detailed auditory nerve model used as input to our networks can be altered to simulate peripheral consequences of hair cell loss [4] or electrical stimulation by a cochlear implant [159]. Cochlear neuropathy can realistically be simulated in our models by adjusting the numbers of different nerve fiber types [158]. Models optimized for auditory nerve representations with different impairments could clarify the patterns of behavior to expect under different hearing loss etiologies or cochlear implant processors [192]. Models optimized with either healthy or impaired input and then evaluated with impaired input (potentially with

partial re-optimization for impaired input) could yield insights into the role of plasticity after hearing loss. Such experiments may shed light on hearing aid acclimatization or cochlear implant outcomes.

The causes and consequences of hearing loss are highly variable. Distinguishing different types of sensorineural hearing loss is important for remediation but remains challenging because the state of the auditory periphery cannot be directly observed in humans. Diagnostic tests for inferring peripheral state from behavior are needed in the clinic but have been difficult to identify [193]. Task-optimized models that predict behavioral consequences of different hearing loss etiologies could enable efficient searches over tasks and stimuli to identify such tests. For instance, stimuli could be synthesized to produce maximally disparate perceptual judgments between auditory models with different peripheral states [160].

Task-optimized models of impaired auditory perception could potentially be used to develop hearing aid and cochlear implant processing strategies. In Chapter 3, we found that the human-like behavior of our networks did not critically depend on many of the complex, nonlinear response properties of the auditory nerve. Very similar behavior emerged in models optimized with a much simpler cochlear front-end through which gradients can be back-propagated. This raises the possibility of using model-based loss functions to artificially learn novel hearing aid strategies [194]. Chapter 4 explored a precursor to this, investigating the utility of task-optimized auditory models as perceptual metrics for speech denoising. We found that loss functions based on deep network features outperformed standard waveform-based losses but not simple cochlear model-based losses. However, it remains to be seen whether our networks confer benefit for hearing aids, which must contend with altered cochlear processing. A hearing aid processor optimized to reconstruct healthy auditory nerve representations may be suboptimal for a person with a severely degraded auditory nerve. Processors optimized for perceptual outcomes in a network equipped with impaired ears may yet provide benefit and warrant future investigation.

References

[1] J. J. Zwislocki, "Five decades of research on cochlear mechanics," *The Journal of the Acoustical Society of America*, vol. 67, no. 5, pp. 1679–1685, May 1, 1980.

[2] M. A. Ruggero, "Physiology and coding of sound in the auditory nerve," in *The Mammalian Auditory Pathway: Neurophysiology*, ser. Springer Handbook of Auditory Research, A. N. Popper and R. R. Fay, Eds., New York, NY: Springer, 1992, pp. 34–93.

[3] M. G. Heinz, X. Zhang, I. C. Bruce, and L. H. Carney, "Auditory nerve model for predicting performance limits of normal and impaired listeners," *Acoustics Research Letters Online*, vol. 2, no. 3, pp. 91–96, Jun. 8, 2001.

[4] M. S. A. Zilany and I. C. Bruce, "Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery," *The Journal of the Acoustical Society of America*, vol. 120, no. 3, pp. 1446–1466, Sep. 1, 2006.

[5] I. C. Bruce, Y. Erfani, and M. S. A. Zilany, "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites," *Hearing Research*, Computational models of the auditory system, vol. 360, pp. 40–54, Mar. 1, 2018.

[6] A. Ciorba, C. Bianchini, S. Pelucchi, and A. Pastore, "The impact of hearing loss on the quality of life of elderly adults," *Clinical Interventions in Aging*, vol. 7, pp. 159–163, Jun. 15, 2012, Publisher: Dove Medical Press _eprint: https://www.tandfonline.com/doi/pdf/10.2147/CIA.S26059.

[7] W. H. Organization, *World report on hearing*. World Health Organization, Mar. 3, 2021, 272 pp., Google-Books-ID: zMRqEAAAQBAJ.

[8] M. G. Heinz, "Computational modeling of sensorineural hearing loss," in *Computational Models of the Auditory System*, ser. Springer Handbook of Auditory Research, Springer, Boston, MA, 2010, pp. 177–202.

[9] W. Siebert, "Frequency discrimination in the auditory system: Place or periodicity mechanisms?" *Proceedings of the IEEE*, vol. 58, no. 5, pp. 723–730, May 1970, Conference Name: Proceedings of the IEEE.

[10] M. G. Heinz, H. S. Colburn, and L. H. Carney, "Evaluating auditory performance limits: II. one-parameter discrimination with random-level variation," *Neural Computation*, vol. 13, no. 10, pp. 2317–2338, Oct. 1, 2001.

[11] W. S. Geisler, "Contributions of ideal observer theory to vision research," *Vision Research*, Vision Research 50th Anniversary Issue: Part 1, vol. 51, no. 7, pp. 771–781, Apr. 13, 2011.

[12] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8619–8624, Jun. 10, 2014.

[13] A. J. E. Kell, D. L. K. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott, "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy," *Neuron*, vol. 98, no. 3, 630–644.e16, May 2, 2018.

[14] K. M. Jozwik, N. Kriegeskorte, K. R. Storrs, and M. Mur, "Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments," *Frontiers in Psychology*, vol. 8, 2017, Publisher: Frontiers.

[15] R. Rajalingham, E. B. Issa, P. Bashivan, K. Kar, K. Schmidt, and J. J. DiCarlo, "Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks," *The Journal of Neuroscience*, vol. 38, no. 33, pp. 7255–7269, Aug. 15, 2018.

[16] M. L. King, I. I. A. Groen, A. Steel, D. J. Kravitz, and C. I. Baker, "Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images," *NeuroImage*, vol. 197, pp. 368–382, Aug. 15, 2019.

[17] M. R. Saddler, R. Gonzalez, and J. H. McDermott, "Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception," *Nature Communications*, vol. 12, no. 1, p. 7278, Dec. 14, 2021, Number: 1 Publisher: Nature Publishing Group.

[18] A. Francl and J. H. McDermott, "Deep neural network models of sound localization reveal how perception is adapted to real-world environments," *Nature Human Behaviour*, vol. 6, no. 1, pp. 111–133, Jan. 2022, Number: 1 Publisher: Nature Publishing Group.

[19] D. L. K. Yamins and J. J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature Neuroscience*, vol. 19, no. 3, pp. 356–365, Mar. 2016.

[20] M. Schrimpf, J. Kubilius, M. J. Lee, N. A. R. Murty, R. Ajemian, and J. J. DiCarlo, "Integrative benchmarking to advance neurally mechanistic models of human intelligence," *Neuron*, vol. 108, no. 3, pp. 413–423, Nov. 11, 2020, Publisher: Elsevier.

[21] G. Tuckute, J. Feather, D. Boebinger, and J. H. McDermott, *Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions*, Pages: 2022.09.06.506680 Section: New Results, Jun. 14, 2023.

[22] A. d. Cheveigné, "Pitch perception models," in *Pitch*, ser. Springer Handbook of Auditory Research, Springer, New York, NY, 2005, pp. 169–233.

[23] A. J. Oxenham, "Questions and controversies surrounding the perception and neural coding of pitch," *Frontiers in Neuroscience*, vol. 16, 2023.

[24] P. A. Cariani and B. Delgutte, "Neural correlates of the pitch of complex tones. i. pitch and pitch salience," *Journal of Neurophysiology*, vol. 76, no. 3, pp. 1698–1716, Sep. 1, 1996.

[25] S. Shamma and D. Klein, "The case of the missing pitch templates: How harmonic templates emerge in the early auditory system," *The Journal of the Acoustical Society of America*, vol. 107, no. 5, pp. 2631–2644, May 2000.

[26] R. D. Patterson, S. Uppenkamp, I. S. Johnsrude, and T. D. Griffiths, "The processing of temporal pitch and melody information in auditory cortex," *Neuron*, vol. 36, no. 4, pp. 767–776, Nov. 14, 2002.

[27] A. de Cheveigné and D. Pressnitzer, "The case of the missing delay lines: Synthetic delays obtained by cross-channel phase interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 6, pp. 3908–3918, Jun. 1, 2006, Publisher: Acoustical Society of America.

[28] A. J. Oxenham, C. Micheyl, M. V. Keebler, A. Loper, and S. Santurette, "Pitch perception beyond the traditional existence region of pitch," *Proceedings of the National Academy of Sciences*, vol. 108, no. 18, pp. 7629–7634, May 3, 2011, Publisher: Proceedings of the National Academy of Sciences.

[29] D. A. Schwartz and D. Purves, "Pitch is determined by naturally occurring periodic sounds," *Hearing Research*, vol. 194, no. 1, pp. 31–46, Aug. 1, 2004.

[30] J. E. Rose, J. F. Brugge, D. J. Anderson, and J. E. Hind, "Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey.," *Journal of Neurophysiology*, vol. 30, no. 4, pp. 769–793, Jul. 1967, Publisher: American Physiological Society.

[31] D. H. Johnson, "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," *The Journal of the Acoustical Society of America*, vol. 68, no. 4, pp. 1115–1122, Oct. 1, 1980, Publisher: Acoustical Society of America.

[32] A. R. Palmer and I. J. Russell, "Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells," *Hearing Research*, vol. 24, no. 1, pp. 1–15, Jan. 1, 1986.

[33] E. Verschooten, S. Shamma, A. J. Oxenham, B. C. J. Moore, P. X. Joris, M. G. Heinz, and C. J. Plack, "The upper frequency limit for the use of phase locking to code temporal fine structure in humans: A compilation of viewpoints," *Hearing Research*, vol. 377, pp. 109–121, Jun. 1, 2019.

[34] C. Lorenzi, G. Gilbert, H. Carn, S. Garnier, and B. C. J. Moore, "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *Proceedings of the National Academy of Sciences*, vol. 103, no. 49, pp. 18 866–18 869, Dec. 5, 2006, Publisher: Proceedings of the National Academy of Sciences.

[35] B. C. J. Moore, "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people," *Journal of the Association for Research in Otolaryngology*, vol. 9, no. 4, pp. 399–406, Dec. 1, 2008.

[36] K. Hopkins and B. C. J. Moore, "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 442–446, Jan. 1, 2009.

[37] M. R. Saddler, A. Francl, J. Feather, K. Qian, Y. Zhang, and J. H. McDermott, "Speech denoising with auditory models," in *Interspeech 2021*, ISCA, Aug. 30, 2021, pp. 2681–2685.

[38] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016.

[39] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[40] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *arXiv preprint arXiv:1708.07524*, 2017.

[41] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using bayesian wavenet.," in *Interspeech*, 2017, pp. 2013–2017.

[42] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ISSN: 2379-190X, Apr. 2018, pp. 5069–5073.

[43] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019, Publisher: IEEE.

[44] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019, Publisher: IEEE.

[45] T.-A. Hsieh, H.-M. Wang, X. Lu, and Y. Tsao, "WaveCRN: An efficient convolutional recurrent neural network for end-to-end speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 2149–2153, 2020, Publisher: Institute of Electrical and Electronics Engineers (IEEE).

[46] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," in *Proc. Interspeech 2019*, 2019, pp. 2723–2727.

[47] B. A. Wandell, *Foundations of Vision*. Sinauer Associates, 1995.

[48] J. G. Hildebrand and G. M. Shepherd, "Mechanisms of olfactory discrimination: Converging evidence for common principles across phyla," *Annual Review of Neuroscience*, vol. 20, no. 1, pp. 595–631, 1997.

[49] R. K. Yin, "Looking at upside-down faces," *Journal of Experimental Psychology*, vol. 81, no. 1, pp. 141–145, 1969, Place: US Publisher: American Psychological Association.

[50] F. Attneave and R. K. Olson, "Pitch as a medium: A new approach to psychophysical scaling," *The American Journal of Psychology*, vol. 84, no. 2, pp. 147–166, 1971, Publisher: University of Illinois Press.

[51] E. Javel and J. B. Mott, "Physiological and psychophysical correlates of temporal processes in hearing," *Hearing Research*, vol. 34, no. 3, pp. 275–294, Aug. 1, 1988.

[52] N. Jacoby, E. A. Undurraga, M. J. McPherson, J. Valdés, T. Ossandón, and J. H. McDermott, "Universal and non-universal features of musical pitch perception revealed by singing," *Current Biology*, vol. 29, no. 19, 3229–3243.e12, Oct. 2019.

[53] Y. Weiss, E. P. Simoncelli, and E. H. Adelson, "Motion illusions as optimal percepts," *Nature Neuroscience*, vol. 5, no. 6, pp. 598–604, Jun. 2002.

[54] J. Burge and W. S. Geisler, "Optimal defocus estimation in individual natural images," *Proceedings of the National Academy of Sciences*, vol. 108, no. 40, pp. 16 849–16 854, Oct. 4, 2011, Publisher: National Academy of Sciences Section: Biological Sciences.

[55] A. R. Girshick, M. S. Landy, and E. P. Simoncelli, "Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics," *Nature Neuroscience*, vol. 14, no. 7, pp. 926–932, Jul. 2011, Number: 7 Publisher: Nature Publishing Group.

[56] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[57] A. J. Kell and J. H. McDermott, "Deep neural network models of sensory systems: Windows onto the role of task constraints," *Current Opinion in Neurobiology*, Machine Learning, Big Data, and Neuroscience, vol. 55, pp. 121–132, Apr. 1, 2019.

[58] M. J. McPherson and J. H. McDermott, "Diversity in pitch perception revealed by task dependence," *Nature Human Behaviour*, vol. 2, no. 1, pp. 52–66, Jan. 2018, Number: 1 Publisher: Nature Publishing Group.

[59] B. C. J. Moore, B. R. Glasberg, and R. W. Peters, "Relative dominance of individual partials in determining the pitch of complex tones," *The Journal of the Acoustical Society of America*, vol. 77, no. 5, pp. 1853–1860, May 1, 1985.

[60] T. M. Shackleton and R. P. Carlyon, "The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination," *The Journal of the Acoustical Society of America*, vol. 95, no. 6, pp. 3529–3540, Jun. 1994.

[61] G. A. Moore and B. C. J. Moore, "Perception of the low pitch of frequency-shifted complexes," *The Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 977–985, Jan. 28, 2003.

[62] A. J. Oxenham, J. G. W. Bernstein, and H. Penagos, "Correct tonotopic representation is necessary for complex pitch perception," *Proceedings of the National Academy of Sciences*, vol. 101, no. 5, pp. 1421–1425, Feb. 3, 2004.

[63] J. G. W. Bernstein and A. J. Oxenham, "An autocorrelation model with place dependence to account for the effect of harmonic number on fundamental frequency discrimination," *The Journal of the Acoustical Society of America*, vol. 117, no. 6, pp. 3816–3831, May 31, 2005.

[64] D. Bendor and X. Wang, "The neuronal representation of pitch in primate auditory cortex," *Nature*, vol. 436, no. 7054, pp. 1161–1165, Aug. 2005.

[65] S. Norman-Haignere, N. Kanwisher, and J. H. McDermott, "Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex," *Journal of Neuroscience*, vol. 33, no. 50, pp. 19 451–19 469, Dec. 11, 2013.

[66] A. H. Mehta and A. J. Oxenham, "Effect of lowest harmonic rank on fundamental-frequency difference limens varies with fundamental frequency," *The Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2314–2322, Apr. 2020.

[67] J. C. R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 7, no. 4, pp. 128–134, Apr. 1, 1951.

[68] J. L. Goldstein, "An optimum processor theory for the central formation of the pitch of complex tones," *The Journal of the Acoustical Society of America*, vol. 54, no. 6, pp. 1496–1516, Dec. 1, 1973.

[69] E. Terhardt, "Calculating virtual pitch," *Hearing Research*, vol. 1, no. 2, pp. 155–182, Mar. 1, 1979.

[70] M. Slaney and R. Lyon, "A perceptual pitch detector," in *International Conference on Acoustics, Speech, and Signal Processing*, ISSN: 1520-6149, Apr. 1990, 357–360 vol.1.

[71] R. Meddis and L. O'Mard, "A unitary model of pitch perception," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1811–1820, Sep. 1, 1997.

[72] J. Laudanski, Y. Zheng, and R. Brette, "A structural theory of pitch," *eNeuro*, vol. 1, no. 1, Nov. 1, 2014, Publisher: Society for Neuroscience Section: Theory/New Concepts.

[73] N. Ahmad, I. Higgins, K. M. M. Walker, and S. M. Stringer, "Harmonic training and the formation of pitch representation in a neural network model of the auditory brain," *Frontiers in Computational Neuroscience*, vol. 10, 2016, Publisher: Frontiers.

[74] O. Barzelay, M. Furst, and O. Barak, "A new approach to model pitch perception using sparse coding," *PLOS Computational Biology*, vol. 13, no. 1, e1005338, Jan. 18, 2017.

[75] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ISSN: 2379-190X, Apr. 2018, pp. 161–165.

[76] A. J. M. Houtsma and J. Smurzynski, "Pitch identification and discrimination for complex tones with many harmonics," *The Journal of the Acoustical Society of America*, vol. 87, no. 1, pp. 304–310, Jan. 1, 1990, Publisher: Acoustical Society of America.

[77] B. C. J. Moore, B. R. Glasberg, and G. M. Proctor, "Accuracy of pitch matching for pure tones and for complex tones with overlapping or nonoverlapping harmonics," *The Journal of the Acoustical Society of America*, vol. 91, no. 6, pp. 3443–3450, Jun. 1, 1992, Publisher: Acoustical Society of America.

[78] J. Mehrer, C. J. Spoerer, N. Kriegeskorte, and T. C. Kietzmann, "Individual differences among deep neural network models," *Nature Communications*, vol. 11, no. 1, p. 5725, Nov. 12, 2020, Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cognitive neuroscience;Computational neuroscience;Network models;Neuroscience Subject_term_id: cognitive-neuroscience;computational-neuroscience;network-models;neuroscience.

[79] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, vol. 30, Long Beach, CA, Nov. 3, 2017, pp. 6402–6413.

[80] A. G. Wilson, "The case for bayesian deep learning," *arXiv:2001.10995 [cs, stat]*, Jan. 29, 2020.

[81] J. G. Bernstein and A. J. Oxenham, "Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number?" *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3323–3334, May 29, 2003.

[82] C. C. Wier, W. Jesteadt, and D. M. Green, "Frequency discrimination as a function of frequency and sensation level," *The Journal of the Acoustical Society of America*, vol. 61, no. 1, pp. 178–184, Jan. 1, 1977, Publisher: Acoustical Society of America.

[83] K. H. Arehart and E. M. Burns, "A comparison of monotic and dichotic complex-tone pitch perception in listeners with hearing loss," *The Journal of the Acoustical Society of America*, vol. 106, no. 2, pp. 993–997, Aug. 1999.

[84] J. G. W. Bernstein and A. J. Oxenham, "The relationship between frequency selectivity and pitch discrimination: Sensorineural hearing loss," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 3929–3945, Dec. 1, 2006.

[85] C. A. Shera, J. J. Guinan, and A. J. Oxenham, "Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements," *Proceedings of the National Academy of Sciences*, vol. 99, no. 5, pp. 3318–3323, Mar. 5, 2002.

[86] C. Micheyl, K. Delhommeau, X. Perrot, and A. J. Oxenham, "Influence of musical and psychoacoustical training on pitch discrimination," *Hearing Research*, vol. 219, no. 1, pp. 36–47, Sep. 1, 2006.

[87] M. J. McPherson and J. H. McDermott, "Time-dependent discrimination advantages for harmonic sounds suggest efficient coding for memory," *Proceedings of the National Academy of Sciences*, vol. 117, no. 50, pp. 32 169–32 180, Dec. 15, 2020, Publisher: National Academy of Sciences Section: Biological Sciences.

[88] D. Bendor, M. S. Osmanski, and X. Wang, "Dual-pitch processing mechanisms in primate auditory cortex," *Journal of Neuroscience*, vol. 32, no. 46, pp. 16 149–16 161, Nov. 14, 2012, Publisher: Society for Neuroscience Section: Articles.

[89] M. J. McPherson, R. C. Grace, and J. H. McDermott, "Harmonicity aids hearing in noise," *Attention, Perception, & Psychophysics*, vol. 84, no. 3, pp. 1016–1042, Apr. 1, 2022.

[90] N. I. Durlach and L. D. Braida, "Intensity perception. i. preliminary theory of intensity resolution," *The Journal of the Acoustical Society of America*, vol. 46, no. 2, pp. 372–383, Aug. 1, 1969, Publisher: Acoustical Society of America.

[91] C. Micheyl, P. R. Schrater, and A. J. Oxenham, "Auditory frequency and intensity discrimination explained using a cortical population rate code," *PLOS Computational Biology*, vol. 9, no. 11, e1003336, Nov. 14, 2013, Publisher: Public Library of Science.

[92] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, pp. 356–363, Apr. 2002, Number: 4 Publisher: Nature Publishing Group.

[93] J. Feather, A. Durango, R. Gonzalez, and J. McDermott, "Metamers of neural networks reveal divergence from human perceptual systems," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 10 078–10 089.

[94] G. W. Lindsay, "Convolutional neural networks as a model of the visual system: Past, present, and future," *Journal of Cognitive Neuroscience*, vol. 33, no. 10, pp. 2017–2031, Sep. 1, 2021.

[95] C. Tang, L. S. Hamilton, and E. F. Chang, "Intonational speech prosody encoding in the human auditory cortex," *Science*, vol. 357, no. 6353, pp. 797–801, Aug. 25, 2017, Publisher: American Association for the Advancement of Science Section: Report.

[96] W. J. Dowling and D. S. Fujitani, "Contour, interval, and pitch recognition in memory for melodies," *The Journal of the Acoustical Society of America*, vol. 49, no. 2, pp. 524–531, Feb. 1971.

[97] E. J. Allen and A. J. Oxenham, "Symmetric interactions and interference between pitch and timbre," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1371–1379, Mar. 2014.

[98] J. K. Bizley, K. M. M. Walker, F. R. Nodal, A. J. King, and J. W. H. Schnupp, "Auditory cortex represents both pitch judgments and the corresponding acoustic cues," *Current Biology*, vol. 23, no. 7, pp. 620–625, Apr. 8, 2013.

[99] K. Gfeller, C. Turner, J. Oleson, X. Zhang, B. Gantz, R. Froman, and C. Olszewski, "Accuracy of cochlear implant recipients on pitch perception, melody recognition, and speech reception in noise," *Ear and Hearing*, vol. 28, no. 3, pp. 412–423, Jun. 2007.

[100] W. P. Shofner and M. Chaney, "Processing pitch in a nonhuman mammal (chinchilla laniger)," *Journal of Comparative Psychology*, vol. 127, no. 2, pp. 142–153, 2013, Place: US Publisher: American Psychological Association.

[101] K. M. Walker, R. Gonzalez, J. Z. Kang, J. H. McDermott, and A. J. King, "Across-species differences in pitch perception are consistent with differences in cochlear filtering," *eLife*, vol. 8, C. E. Carr, E. Marder, and D. Bendor, Eds., e41626, Mar. 15, 2019.

[102] P. X. Joris, C. Bergevin, R. Kalluri, M. M. Laughlin, P. Michelet, M. v. d. Heijden, and C. A. Shera, "Frequency selectivity in old-world monkeys corroborates sharp cochlear tuning in humans," *Proceedings of the National Academy of Sciences*, vol. 108, no. 42, pp. 17 516–17 520, Oct. 18, 2011.

[103] L. J. White and C. J. Plack, "Temporal processing of the pitch of complex tones," *The Journal of the Acoustical Society of America*, vol. 103, no. 4, pp. 2051–2063, Apr. 1, 1998, Publisher: Acoustical Society of America.

[104] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3933–3936, 2008.

[105] A. Köhn, F. Stegen, and T. Baumann, "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), May 2016, pp. 4644–4647.

[106] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proceedings of the 34th International Conference on Machine Learning*, ISSN: 2640-3498, PMLR, Jul. 17, 2017, pp. 1068–1077.

[107] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[108] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," *Neuron*, vol. 71, pp. 926–940, 2011.

[109] R. McWalter and J. H. McDermott, "Adaptive and selective time averaging of auditory scenes," *Current Biology*, vol. 28, no. 9, 1405–1418.e10, May 7, 2018.

[110] M. C. Liberman, "Central projections of auditory-nerve fibers of differing spontaneous rate. i. anteroventral cochlear nucleus," *Journal of Comparative Neurology*, vol. 313, no. 2, pp. 240–258, 1991, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cne.903130205.

[111] L. H. Carney, "Supra-threshold hearing and fluctuation profiles: Implications for sensorineural and hidden hearing loss," *Journal of the Association for Research in Otolaryngology*, vol. 19, no. 4, pp. 331–352, Aug. 1, 2018.

[112] B. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

[113] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, Aug. 1, 2005, Publisher: Acoustical Society of America.

[114] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. II. spectral and temporal integration," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2906–2919, Nov. 1, 1997, Publisher: Acoustical Society of America.

[115] D. Pressnitzer, R. D. Patterson, and K. Krumbholz, "The lower limit of melodic pitch," *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2074–2084, May 1, 2001, Publisher: Acoustical Society of America.

[116] D. M. Green and J. A. Swets, *Signal detection theory and psychophysics* (Signal detection theory and psychophysics). Oxford, England: John Wiley, 1966, xi, 455, Pages: xi, 455.

[117] H. B. Barlow, "The efficiency of detecting changes of density in random dot patterns," *Vision Research*, vol. 18, no. 6, pp. 637–650, Jan. 1, 1978.

[118] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, Jan. 2002, Number: 6870 Publisher: Nature Publishing Group.

[119] D. Kersten, P. Mamassian, and A. Yuille, "Object perception as bayesian inference," *Annual Review of Psychology*, vol. 55, no. 1, pp. 271–304, Feb. 1, 2004.

[120] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. i. model structure," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, Jun. 1, 1996.

[121] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, Jan. 2011, Conference Name: IEEE Transactions on Audio, Speech, and Language Processing.

[122] D. R. Guest and A. J. Oxenham, "Human discrimination and modeling of high-frequency complex tones shed light on the neural codes for pitch," *PLOS Computational Biology*, vol. 18, no. 3, e1009889, Mar. 3, 2022, Publisher: Public Library of Science.

[123] A. Goldstein, Z. Zada, E. Buchnik, *et al.*, "Shared computational principles for language processing in humans and deep language models," *Nature Neuroscience*, vol. 25, no. 3, pp. 369–380, Mar. 2022, Number: 3 Publisher: Nature Publishing Group.

[124] Z. F. Mainen and T. J. Sejnowski, "Reliability of spike timing in neocortical neurons," *Science*, vol. 268, no. 5216, pp. 1503–1506, Jun. 9, 1995, Publisher: American Association for the Advancement of Science.

[125] F. Marion-Poll and T. R. Tobin, "Temporal coding of pheromone pulses and trains in manduca sexta," *Journal of Comparative Physiology A*, vol. 171, no. 4, pp. 505–512, Nov. 1, 1992.

[126] J. D. Victor and K. P. Purpura, "Nature and precision of temporal coding in visual cortex: A metric-space analysis," *Journal of Neurophysiology*, vol. 76, no. 2, pp. 1310–1326, Aug. 1996, Publisher: American Physiological Society.

[127] A. Carleton, R. Accolla, and S. A. Simon, "Coding in the mammalian gustatory system," *Trends in Neurosciences*, vol. 33, no. 7, pp. 326–334, Jul. 2010.

[128] E. L. Mackevicius, M. D. Best, H. P. Saal, and S. J. Bensmaia, "Millisecond precision spike timing shapes tactile perception," *Journal of Neuroscience*, vol. 32, no. 44, pp. 15 309–15 317, Oct. 31, 2012, Publisher: Society for Neuroscience Section: Articles.

[129] M. K. Qin and A. J. Oxenham, "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," *The Journal of the Acoustical Society of America*, vol. 114, no. 1, pp. 446–454, Jul. 3, 2003.

[130] M. C. Liberman, "Auditory-nerve response from cats raised in a low-noise chamber," *The Journal of the Acoustical Society of America*, vol. 63, no. 2, pp. 442–455, Feb. 1978.

[131] P. X. Joris and E. Verschooten, "On the limit of neural phase locking to fine structure in humans," in *Basic Aspects of Hearing*, B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel, Eds., ser. Advances in Experimental Medicine and Biology, New York, NY: Springer, 2013, pp. 101–108.

[132] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *The Journal of the Acoustical Society of America*, vol. 111, no. 5, pp. 2219–2236, May 6, 2002.

[133] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, no. 6713, pp. 747–749, Dec. 1998, Number: 6713 Publisher: Nature Publishing Group.

[134] P. M. Hofman, J. G. A. Van Riswick, and A. J. Van Opstal, "Relearning sound localization with new ears," *Nature Neuroscience*, vol. 1, no. 5, pp. 417–421, Sep. 1998, Number: 5 Publisher: Nature Publishing Group.

[135] A. W. Mills, "On the minimum audible angle," *The Journal of the Acoustical Society of America*, vol. 30, no. 4, pp. 237–246, Apr. 1958.

[136] B. Moore, M. Huss, D. A. Vickers, B. R. Glasberg, and J. Alcántara, "A test for the diagnosis of dead regions in the cochlea," *British Journal of Audiology*, vol. 34, no. 4, pp. 205–224, Aug. 1, 2000, Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/03005364000000131.

[137] D. W. Batteau and H. E. Huxley, "The role of the pinna in human localization," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 168, no. 1011, pp. 158–180, Aug. 1967, Publisher: Royal Society.

[138] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT Press, 1997, 512 pp., Google-Books-ID: wBiEKPhw7r0C.

[139] L. Rayleigh, "On our perception of sound direction," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, Feb. 1907.

[140] E. R. Hafter, R. H. Dye Jr., and R. H. Gilkey, "Lateralization of tonal signals which have neither onsets nor offsets," *The Journal of the Acoustical Society of America*, vol. 65, no. 2, pp. 471–477, Feb. 1, 1979.

[141] G. B. Henning, "Lateralization of low-frequency transients," *Hearing Research*, vol. 9, no. 2, pp. 153–172, Feb. 1, 1983.

[142] M. Klein-Hennig, M. Dietz, V. Hohmann, and S. D. Ewert, "The influence of different segments of the ongoing envelope on sensitivity to interaural time delays," *The Journal of the Acoustical Society of America*, vol. 129, no. 6, pp. 3856–3872, Jun. 14, 2011.

[143] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2188–2200, May 1, 1990.

[144] S. Carlile, P. Leong, and S. Hyams, "The nature and distribution of errors in sound localization by human listeners," *Hearing Research*, vol. 114, no. 1, pp. 179–196, Dec. 1, 1997.

[145] K. C. Wood and J. K. Bizley, "Relative sound localisation abilities in human listeners," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 674–686, Aug. 6, 2015.

[146] A. Brughera, L. Dunai, and W. M. Hartmann, "Human interaural time difference thresholds for sine tones: The high-frequency limit," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2839–2855, May 6, 2013.

[147] J. Zwislocki and R. S. Feldman, "Just noticeable differences in dichotic phase," *The Journal of the Acoustical Society of America*, vol. 28, no. 5, pp. 860–864, Sep. 1956.

[148] L. A. Jeffress, "A place theory of sound localization," *Journal of Comparative and Physiological Psychology*, vol. 41, no. 1, pp. 35–39, 1948, Place: US Publisher: American Psychological Association.

[149] H. S. Colburn and N. I. Durlach, "Models of binaural interaction," *Handbook of perception*, vol. 4, pp. 467–518, 1978.

[150] B. Grothe and D. H. Sanes, "Synaptic inhibition influences the temporal coding properties of medial superior olivary neurons: An in vitro study," *Journal of Neuroscience*, vol. 14, no. 3, pp. 1701–1709, Mar. 1, 1994, Publisher: Society for Neuroscience Section: Articles.

[151] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, Jun. 1, 1995.

[152] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "A precedence effect in sound localization," *The Journal of the Acoustical Society of America*, vol. 21, no. 4, p. 468, 1949.

[153] W. A. Yost and X. Zhong, "Sound source localization identification accuracy: Bandwidth dependencies," *The Journal of the Acoustical Society of America*, vol. 136, no. 5, pp. 2737–2746, Nov. 1, 2014.

[154] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *The Journal of the Acoustical Society of America*, vol. 56, no. 6, pp. 1829–1834, Sep. 16, 1974.

[155] S. Popham, D. Boebinger, D. P. W. Ellis, H. Kawahara, and J. H. McDermott, "Inharmonic speech reveals the role of harmonicity in the cocktail party problem," *Nature Communications*, vol. 9, no. 1, p. 2122, May 29, 2018, Number: 1 Publisher: Nature Publishing Group.

[156] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, *DDSP: Differentiable digital signal processing*, Jan. 14, 2020.

[157] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, Jan. 1, 2000.

[158] A. C. Furman, S. G. Kujawa, and M. C. Liberman, "Noise-induced cochlear neuropathy is selective for fibers with low spontaneous rates," *Journal of Neurophysiology*, vol. 110, no. 3, pp. 577–586, Aug. 2013, Publisher: American Physiological Society.

[159] S. Tabibi, J. Boulet, N. Dillier, and I. C. Bruce, "Phenomenological model of auditory nerve population responses to cochlear implant stimulation," *Journal of Neuroscience Methods*, vol. 358, p. 109 212, Jul. 1, 2021.

[160] T. Golan, P. C. Raju, and N. Kriegeskorte, "Controversial stimuli: Pitting neural networks against each other as models of human cognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 47, pp. 29 330–29 337, Nov. 24, 2020, Publisher: National Academy of Sciences Section: Colloquium Paper.

[161] L. Dai, V. Best, and B. G. Shinn-Cunningham, "Sensorineural hearing loss degrades behavioral and physiological measures of human spatial selective auditory attention," *Proceedings of the National Academy of Sciences*, vol. 115, no. 14, E3286–E3295, Apr. 3, 2018, Publisher: Proceedings of the National Academy of Sciences.

[162] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012.

[163] S. Yadav and M. E. Foster, *GISE-51: A scalable isolated sound events dataset*, Oct. 7, 2021.

[164] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, *FSD50k: An open dataset of human-labeled sound events*, Apr. 23, 2022.

[165] B. Shinn-Cunningham, J. Desloge, and N. Kopco, "Empirical and modeled acoustic transfer functions in a simple room: Effects of distance and direction," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, Oct. 2001, pp. 183–186.

[166] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, and M. D. Plumbley, "A database and challenge for acoustic scene classification and event detection," in *21st European Signal Processing Conference (EUSIPCO 2013)*, ISSN: 2076-1465, Sep. 2013, pp. 1–5.

[167] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[168] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18-a corpus for music separation," 2017.

[169] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *The Journal of the Acoustical Society of America*, vol. 88, no. 4, pp. 1725–1736, Oct. 1, 1990.

[170] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27 403, 1993.

[171] R. Y. Litovsky and S. P. Godar, "Difference in precedence effect between children and adults signifies development of sound localization abilities in complex listening tasks," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 1979–1991, Oct. 18, 2010.

[172] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE CVPR 2018*, 2018, pp. 586–595.

[173] R. Pilarczyk and W. Skarbek, "Multi-objective noisy-based deep feature loss for speech enhancement," in *Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2019*, R. S. Romaniuk and M. Linczuk, Eds., Backup Publisher: International Society for Optics and Photonics, vol. 11176, SPIE, 2019, pp. 858–865.

[174] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," *arXiv preprint arXiv:2006.05694*, 2020.

[175] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Improving perceptual quality by phone-fortified perceptual loss for speech enhancement," *arXiv preprint arXiv:2010.15174*, 2020.

[176] S. Kataria, J. Villalba, and N. Dehak, "Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models," *arXiv preprint arXiv:2010.11860*, 2020.

[177] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv:1806.03185 [cs, eess, stat]*, Jun. 8, 2018.

[178] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[179] O. J. Hénaff and E. P. Simoncelli, "Geodesics of learned representations," *arXiv:1511.06394 [cs]*, Feb. 22, 2016.

[180] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the Workshop on Speech and Natural Language*, Association for Computational Linguistics, 1992, pp. 357–362.

[181] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*, IEEE, 2016, pp. 1128–1132.

[182] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.

[183] K. J. P. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, and Psychophysics*, vol. 79, pp. 2064–2072, 2017.

[184] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[185] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[186] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, 2006, Publisher: IEEE.

[187] I. Ananthabhotla, S. Ewert, and J. A. Paradiso, "Towards a perceptual loss: Using a neural network codec approximation as a loss for generative audio models," in *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France*, ACM, 2019, pp. 1–8.

[188] P. Manocha, A. Finkelstein, Z. Jin, N. J. Bryan, R. Zhang, and G. J. Mysore, "A differentiable perceptual audio metric learned from just noticeable differences," *arXiv preprint arXiv:2001.04460*, 2020.

[189] N. Kanwisher, M. Khosla, and K. Dobs, "Using artificial neural networks to ask 'why' questions of minds and brains," *Trends in Neurosciences*, vol. 46, no. 3, pp. 240–254, Mar. 1, 2023, Publisher: Elsevier.

[190] B. A. Richards, T. P. Lillicrap, P. Beaudoin, *et al.*, "A deep learning framework for neuroscience," *Nature Neuroscience*, vol. 22, no. 11, pp. 1761–1770, Nov. 2019, Number: 11 Publisher: Nature Publishing Group.

[191] E. Margalit, H. Lee, D. Finzi, J. J. DiCarlo, K. Grill-Spector, and D. L. K. Yamins, *A unifying principle for the functional organization of visual cortex*, Pages: 2023.05.18.541361 Section: New Results, May 18, 2023.

[192] T. Brochier, J. Schlittenlacher, I. Roberts, T. Goehring, C. Jiang, D. Vickers, and M. Bance, "From microphone to phoneme: An end-to-end computational neural model for predicting speech perception with cochlear implants," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 11, pp. 3300–3312, Nov. 2022, Conference Name: IEEE Transactions on Biomedical Engineering.

[193] R. Sanchez-Lopez, M. Fereczkowski, T. Neher, S. Santurette, and T. Dau, "Robust data-driven auditory profiling towards precision audiology," *Trends in Hearing*, vol. 24, p. 2 331 216 520 973 539, Jan. 1, 2020, Publisher: SAGE Publications Inc.

[194] F. Drakopoulos, V. Vasilkov, A. Osses Vecchi, T. Wartenberg, and S. Verhulst, "Model-based hearing-enhancement strategies for cochlear synaptopathy pathologies," *Hearing Research*, vol. 424, p. 108 569, Oct. 1, 2022.