Article

Neuron

A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy

Highlights

- A deep neural network optimized for speech and music tasks performed as well as human listeners
- The optimization produced separate music and speech pathways after a shared front end
- The network made human-like error patterns and predicted auditory cortical responses
- Network predictions suggest hierarchical organization in human auditory cortex

Authors

Alexander J.E. Kell, Daniel L.K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, Josh H. McDermott

Correspondence

alexkell@mit.edu (A.J.E.K.), jhm@mit.edu (J.H.M.)

In Brief

Kell et al. show that a deep neural network optimized to recognize speech and music replicated human auditory behavior and predicted cortical fMRI responses. Different network layers best predict primary and non-primary voxels, revealing hierarchical organization in human auditory cortex.

Kell et al., 2018, Neuron *98*, 1–15
May 2, 2018
https://doi.org/10.1016/j.neuron.2018.03.044

CellPress

Please cite this article in press as: Kell et al., A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy, Neuron (2018), https://doi.org/10.1016/j.neuron.2018.03.044

Neuron
Article

A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy

Alexander J.E. Kell,[1,2,6,7,*] Daniel L.K. Yamins,[3,4,6] Erica N. Shook,[1,2] Sam V. Norman-Haignere,[1] and Josh H. McDermott[1,2,5,*]
[1]Department of Brain and Cognitive Science, MIT, Cambridge, MA, USA
[2]Center for Brains, Minds, and Machines, MIT, Cambridge, MA, USA
[3]Departments of Psychology and Computer Science, Stanford University, Stanford, CA, USA
[4]Stanford Neurosciences Institute, Stanford, CA, USA
[5]Program in Speech and Hearing Biosciences and Technology, Harvard University, Cambridge, MA, USA
[6]These authors contributed equally
[7]Lead Contact
*Correspondence: alexkell@mit.edu (A.J.E.K.), jhm@mit.edu (J.H.M.)
https://doi.org/10.1016/j.neuron.2018.03.044

SUMMARY

A core goal of auditory neuroscience is to build quantitative models that predict cortical responses to natural sounds. Reasoning that a complete model of auditory cortex must solve ecologically relevant tasks, we optimized hierarchical neural networks for speech and music recognition. The best-performing network contained separate music and speech pathways following early shared processing, potentially replicating human cortical organization. The network performed both tasks as well as humans and exhibited human-like errors despite not being optimized to do so, suggesting common constraints on network and human performance. The network predicted fMRI voxel responses substantially better than traditional spectrotemporal filter models throughout auditory cortex. It also provided a quantitative signature of cortical representational hierarchy—primary and non-primary responses were best predicted by intermediate and late network layers, respectively. The results suggest that task optimization provides a powerful set of tools for modeling sensory systems.

INTRODUCTION

Human listeners extract a remarkable array of information about the world from sound. These abilities are enabled by neuronal processing that transforms the sound waveform entering the ear into cortical representations thought to render behaviorally important sound properties explicit. Although much is known about the peripheral processing of sound, auditory cortex is less understood, particularly in computational terms. There is growing consensus that frequency and modulation tuning explain aspects of primary auditory cortical responses (Depireux et al., 2001; Humphries et al., 2010; Miller et al., 2002; Santoro et al., 2014), but the organization of the rest of auditory cortex into regions and pathways remains unresolved, particularly in humans (Norman-Haignere et al., 2015; Rauschecker and Scott, 2009; Recanzone and Cohen, 2010).

Our understanding of auditory cortex is limited in part by the lack of quantitative models of how neural circuitry transforms sound waveforms into representations that enable behavior. Existing models of auditory processing are mostly limited to one or two stages, typically based on linear filtering of spectrogram-like input (Carlson et al., 2012; Chi et al., 2005; Dau et al., 1997; McDermott and Simoncelli, 2011; Młynarski and McDermott, 2018). Such models explain aspects of auditory perception (McDermott et al., 2013; Patil et al., 2012) and cortical responses (Norman-Haignere et al., 2015; Santoro et al., 2014; Schönwiesner and Zatorre, 2009), but they are clearly incomplete. Neural responses are known to be nonlinear functions of the spectrogram (Christianson et al., 2008; David et al., 2009), and state-of-the-art machine hearing systems are highly nonlinear (Hershey et al., 2017), suggesting that auditory recognition requires invariances that cannot be obtained from the linear operations typically employed in auditory models.

In this paper, we develop a multi-stage computational model that performs real-world auditory tasks. The underlying hypothesis was that everyday recognition tasks may impose strong constraints on the auditory system, such that a model optimized to perform such tasks might converge to brain-like representational transformations. We optimized a deep neural network to map sound waveforms to behaviorally meaningful categories (words or musical genres), leveraging recent advances in what has become known as deep learning (LeCun et al., 2015). Although some aspects of such networks deviate substantially from biological systems, they are currently the only known model class that attains human-level performance on many real-world classification tasks. Following early hopes that such models would yield biological insights (Lehky and Sejnowski, 1988; Zipser and Andersen, 1988), contemporary deep neural networks have been shown to replicate key aspects of visual system

organization (Eickenberg et al., 2017; Güçlü and van Gerven, 2015; Yamins and DiCarlo, 2016). However, their utility for other brain systems remains unclear.

To evaluate the network, we compared its task performance with that of human listeners across a variety of conditions. The network recognized word and musical genres as well as human listeners did, and its error patterns resembled those of humans despite not being optimized to do so. We then used the network's features to predict fMRI voxel responses throughout auditory cortex, finding it to be substantially more predictive than the commonly used spectrotemporal filter model (Chi et al., 2005).

Motivated by these results, we used the network to address an unresolved question in auditory neuroscience: the extent to which auditory cortical computation is hierarchical—consisting of a sequence of stages, potentially corresponding to cortical regions (Okada et al., 2010; Rauschecker and Scott, 2009; Wessinger et al., 2001). In non-human animals, cytoarchitectonic and tracer studies are consistent with a tripartite hierarchical organization (Kaas and Hackett, 2000), and various sources of physiological evidence have been interpreted as supporting hierarchical organization (Atencio et al., 2012; Camalier et al., 2012; Chechik et al., 2006; Rauschecker et al., 1995; Recanzone and Cohen, 2010). However, the extent to which such findings generalize to humans is unclear, in part because of the unique importance of speech and music to human hearing. In humans, hierarchy is most commonly proposed for speech processing, where speech-specific responses only emerge outside of primary areas, suggestive of multiple processing stages (Chang et al., 2010; de Heer et al., 2017; Evans and Davis, 2015; Liebenthal et al., 2005; Norman-Haignere et al., 2015; Obleser et al., 2010; Overath et al., 2015; Peelle et al., 2010; Uppenkamp et al., 2006). Yet it remains unclear whether such regional differences reflect sequential stages of processing. Indeed, some have argued against hierarchy, instead proposing an anatomically distributed organization (Formisano et al., 2008; Staeren et al., 2009).

Our neural network model is intrinsically hierarchical, with the output of one stage forming the input to the next, and thus it provided a means of operationalizing and evaluating the complexity of responses in different parts of auditory cortex. This approach has proven fruitful in the visual system, where the presence of hierarchy is well established—different network layers best predict responses at different stages of the visual cortical hierarchy (Cichy et al., 2016; Güçlü and van Gerven, 2015; Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014). We used a similar approach to probe the relative complexity of responses in different parts of auditory cortex, where the large-scale organization is less settled. We find that intermediate model layers best explain primary auditory cortical responses, while deeper layers best explain voxels in non-primary areas. These results provide quantitative evidence of a computational hierarchy in human auditory cortex.

RESULTS

Network Tasks

To build our neural network model, we used two tasks that were behaviorally relevant and for which we could obtain large sets of labeled data: word recognition and musical genre identification (Figure 1A). The word task required identifying which of 587 words was positioned at the midpoint of a 2 s excerpt of speech; the genre task required identifying which of 41 musical genres a 2 s music clip belonged to (see Tables S1 and S2 for all words and genres). Speech and music training examples were drawn from large, labeled corpora (Bertin-Mahieux et al., 2011; Garofolo and Consortium, 1993; Paul and Baker, 1992) and were superimposed on different types of real-world background "noise" to make the task more challenging and realistic (see STAR Methods for details). Whereas tasks similar to our word recognition task are arguably ecologically important to humans, the genre task was selected primarily because contemporary methods for training deep neural networks require large, labeled datasets, and genre tags, unlike other musical descriptors, are presently available for millions of music clips. The input to the network was a "cochleagram," a time-frequency decomposition of the sound signal that mimics aspects of cochlear signal processing. The network parameters were optimized to map the cochleagram to class labels for each of the two tasks.

Network Architecture Optimization

The network consisted of a series of layers instantiating several standard operations: convolution with linear filters, pointwise nonlinearities, normalization, and pooling. Neural network training is most often associated with the optimization of network filter weights, but networks are also defined by architectural hyperparameters that can substantially affect performance (Pinto et al., 2009; Yamins et al., 2014; Zoph et al., 2017). These include the number of layers, number of units per layer, operations within each layer, filter sizes, and type of pooling operations. Good task performance can often be achieved using architectures that performed well on a related task (Razavian et al., 2014; Zoph et al., 2017). However, because the two tasks we used were relatively novel for convolutional networks, and because we wanted a single network to perform both tasks, we optimized across architectural hyperparameters in addition to the network filter weights. We selected the model architecture via a two-stage procedure, first searching for architectures that performed well on either task in isolation and then searching over ways of combining architectures into a single network that performed both tasks.

In the first stage, we generated nearly two hundred candidate architectures (Figure 1B; STAR Methods). For each architecture, filter weights were optimized via stochastic gradient descent for either the word or genre task alone. Millions of labeled training examples were generated by superimposing exemplars of each word or genre with background noise excerpts at various signal-to-noise ratios (SNRs). After training, performance for each architecture was assessed with left-out stimuli. The same architecture performed best on both tasks. This architecture had twelve layers of processing: five convolutional, three pooling, two normalization, and two fully connected layers (see STAR Methods for details).

In the second stage, we sought a single model that achieved good performance on both the word and genre tasks. A priori, it seemed plausible that speech and music (and potentially other) tasks could be performed using shared initial stages of acoustic

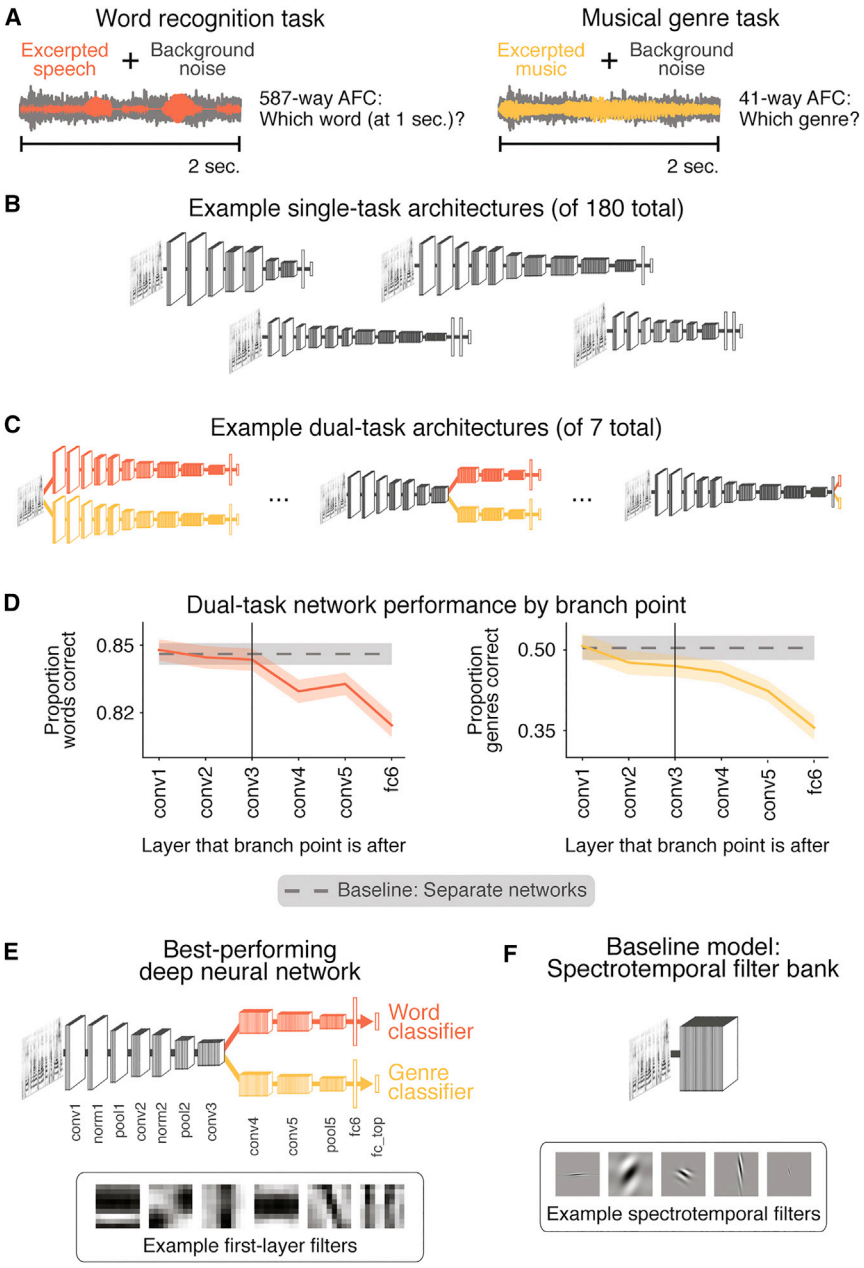


Figure 1. Deep Neural Network Training: Tasks and Architecture Search

(A) Tasks used for model optimization. The network received a 2 s clip of excerpted speech or music mixed with background noise (e.g., a recording of a city street). The network classified either which of 587 words occurred in the middle of the clip or from which of 41 genres the musical excerpt was drawn.

(B) Example candidate single-task architectures. Architectures varied in the number of layers, size of kernels, etc.

(C) Example candidate dual-task architectures, generated by merging the best performing single-task architectures into a single branched network that performed both tasks. Left: architecture with no shared layers (i.e., two separate networks). Middle: an architecture with a few shared layers. Right: an architecture with nearly all shared layers.

(D) Task performance as a function of the branch point position within the network. We considered branch points at each convolutional or fully connected layer, because these are the only layers whose parameters are altered during task training (because they contain the filter weights optimized by backpropagation). Error bars plot SEM, bootstrapped over stimuli and classes. We sought to share as many layers of processing as possible without producing a performance decrement and thus selected the architecture that branched after the third convolutional layer (conv3; vertical black line).

(E) The model architecture that resulted from the task optimization procedure. "conv" denotes convolutional layers (always followed by rectification); "norm" denotes normalization layers; "pool" denotes pooling layers; "fc" denotes fully connected layers. Bottom: example first-layer filters.

(F) Schematic of a commonly used model of auditory cortex, consisting of a single stage of linear spectrotemporal filters on top of a model of the cochlea (a "cochleagram"). Bottom: example spectrotemporal filters.

analysis, after which they might require segregated domain-specific processing. We therefore created "branched" versions of the architecture found in the first stage of architectural optimization (Figure 1C), sharing some number of initial layers of processing before branching into two task-specific processing streams. Because task training did not alter the operations in pooling and normalization layers, the candidate branch points only preceded each of the seven convolutional or fully connected layers. We optimized the filter weights in each of these seven networks for both tasks jointly, using stochastic gradient descent. We then evaluated task performance.

The architecture with fully separate pathways performed better than the architecture with shared processing up until the classification layers. This result is perhaps unsurprising given that the fully separate architecture has nearly twice as many parameters. However, architectures sharing a few early layers performed nearly identically to the fully separate architecture (Figure 1D). On grounds of parsimony, we selected the architecture that shared as much early processing as possible without significantly impairing task performance relative to the fully separate model (determined by bootstrap; STAR Methods). The selected architecture (Figure 1E) shared a total of seven layers, after which the network branched into two sets of five task-specific layers, with each branch culminating in output layers whose responses could be interpreted as probability distributions over classes for each task (i.e., words or genres). The results of this optimization suggest that some degree of speech- and music-specific processing is useful to achieve good task performance, but that shared early processing could be beneficial given a resource

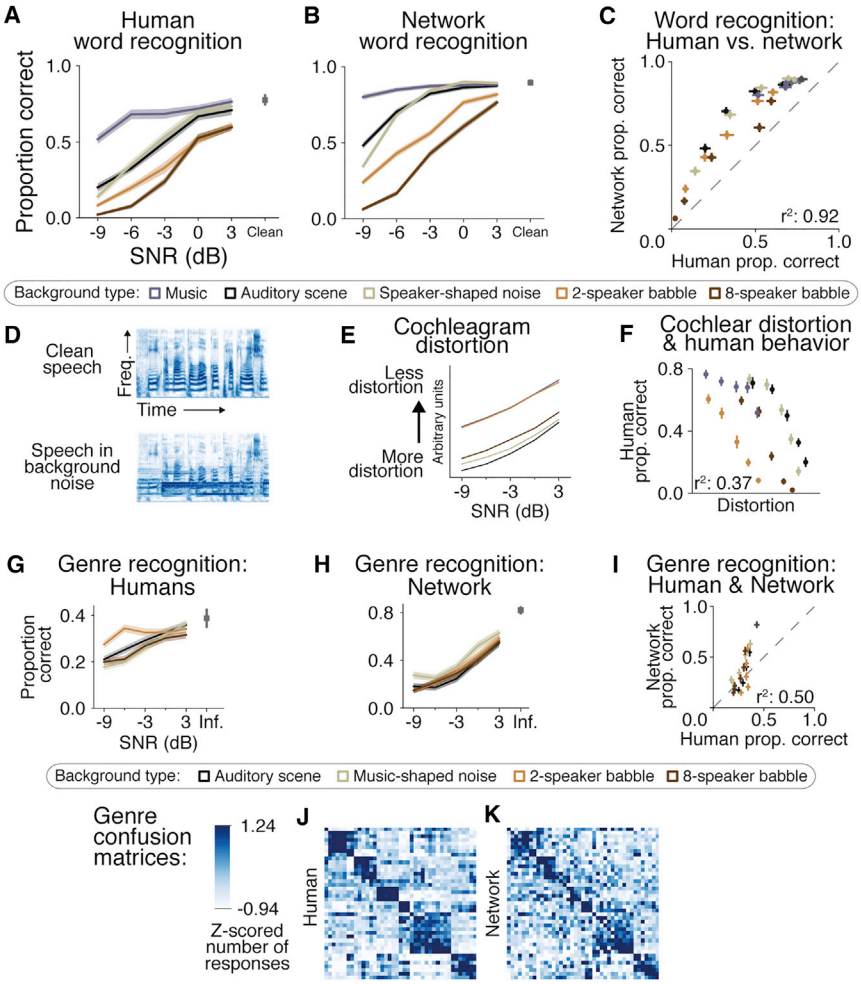


Figure 2. Comparison of Human and Network Behavior: Word and Genre Recognition Tasks

(A) Human performance on word recognition task (n = 18). y axis plots proportion of words correctly identified; x axis plots signal-to-noise ratio (SNR) of the speech signal. Each line plots performance for a particular type of background noise. Gray point on right plots performance without background noise (i.e., infinite SNR). Error bars plot within-subject SEM.

(B) Network performance on the word recognition task (using the same stimuli and task as for human listeners). Same plotting conventions as (A) were used. Error bars plot SEM, bootstrapped over stimuli and words.

(C) Scatterplot comparing word recognition performance of human listeners and the network for each background type at each SNR (i.e., identical data as A and B). Dashed line indicates unity.

(D) Example cochleagrams of a speech signal with and without background noise. Distortion was computed as the mean absolute difference between these two cochleagrams.

(E) Cochleagram distortion by condition. y axis is oriented to facilitate comparison with (A) and (B).

(F) Scatterplot of human performance and cochleagram distortion (i.e., data from A and E). The rank correlation between distortion and human performance is substantially lower than that between human performance and network performance (C). See Figure S1A for results when restricting distortion measurements to time-frequency bins with substantial speech signal power.

(G) Human performance on the genre classification task (n = 111). Same plotting conventions as (A) were used.

(H) Network performance on the same stimuli and task. Same plotting conventions as (B) were used.

(I) Scatterplot comparing genre classification performance of human listeners and the network for each background type at each SNR. Identical data as (G) and (H) are shown. Dashed line indicates unity.

(J) Genre confusion matrix for human listeners. Each column indicates the correct genre, and each row indicates the selected genre. Saturation denotes frequency with which human listeners selected a genre label for exemplars of a particular genre (*Z* scored within columns).

(K) Network confusion matrix. Same plotting conventions as (J) were used.

limitation (e.g., the number of neurons). The resulting network architecture is consistent with recent evidence for segregated speech and music pathways in non-primary auditory cortex (Angulo-Perkins et al., 2014; Leaver and Rauschecker, 2010; Norman-Haignere et al., 2015; Tierney et al., 2013).

Comparison of Model and Human Behavior

A complete model of the auditory system should replicate human auditory behavior. We thus began by measuring human performance for the word and genre tasks, comparing both absolute performance and the pattern of errors to that of the network. For the word classification task, listeners typed what they heard using an interface that auto-completed the 587 words. For the genre classification task, listeners selected the five most likely genres for the 2 s excerpt they heard. A "top 5" task was used to ensure that the task was reasonably well defined given overlap between different genres (e.g., "New Age" versus "Ambient"; see Figure S6B for overlap matrix). In both cases the speech and music excerpts were presented in different types of background noise at a range of SNRs. We measured network performance on the same stimuli and tasks.

Word Recognition Behavioral Comparison

Human listeners' word recognition performance improved with SNR, as expected, but some types of background noise impaired performance more than others (Figure 2A). The network exhibited similar absolute performance levels and similar dependence of performance on background noise ($r^2 = 0.92$, $p < 10^{-13}$; Figures 2B and 2C). The pattern of performance was not explained by simple measures of distortion in the cochlear representation of speech ($r^2 = 0.37$, lower than that with network performance, $p < 10^{-4}$; Figures 2D–2F). Figure S1A shows similar results when restricting the distortion measure to only those cochleagram bins with substantial speech energy; Figure S1B shows the results of distortion measured in network layers.

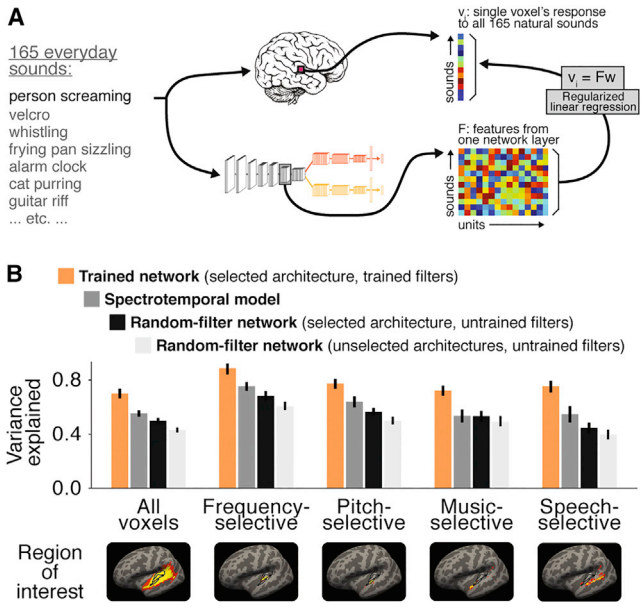


Figure 3. Predicting fMRI Responses to Natural Sounds: Comparison of Network with Baseline Models

(A) Schematic of method for predicting fMRI responses from neural network. We measured responses to 165 natural sounds in an fMRI experiment (Norman-Haignere et al., 2015), estimating each voxel's average response to each sound (top). We then presented the same 165 natural sounds to the trained network and extracted each model unit's time-averaged response to each of these sounds (schematized for one layer in bottom row). We modeled each voxel as a linear combination of model units from a given layer, estimating the linear transform with half the sounds and measuring the prediction quality by correlating the empirical and predicted response to the left-out sounds. We performed this procedure for each network layer and many random splits of the sounds.

(B) Variance explained in functionally localized regions of interest (ROIs). ROIs are shown below graph, with heatmaps of voxel counts across subjects. Black outlines show three anatomically defined sub-divisions of primary auditory cortex (TE 1.1, 1.0, and 1.2), taken from probabilistic maps (Morosan et al., 2001). Orange bar denotes the trained network, dark gray denotes the spectrotemporal filter model, black denotes a network with the identical architecture but with random, untrained filters, and light gray denotes the results from many random-filter networks with different architectures (bar plots median across architectures). Error bars are within-subject SEM.

Genre Recognition Behavioral Comparison

The network also approximately replicated human performance levels on the musical genre task (Figures 2G–2I). Because the number of genres was modest, we compared confusion matrices (this was impractical for the word task because the number of words made the confusion matrix prohibitively large). Confusion matrices for humans and the network were significantly correlated (Figures 2J and 2K; Spearman $r^2 = 0.25$, $p < 10^{-104}$).

Taken together, the behavioral analyses suggest that the performance-optimized neural network replicates non-trivial patterns in human speech and music perceptual behavior, despite not being explicitly optimized to do so.

Predicting Cortical Responses from Network Features

We next examined potential correspondence between representations in the network and auditory cortex, measuring how well different network layers could predict fMRI voxel responses to a broad sample of 165 natural sounds (Table S3). Some of these sounds were speech and music, but most (113 of 165) were not. We measured the response of model units to each sound and predicted each voxel's response from the time-averaged model unit responses from each layer using regularized linear regression (Figure 3A). Time averaging (over the duration of the sound) was used because the blood-oxygen-level-dependent (BOLD) signal is sluggish relative to the 2 s stimulus duration.

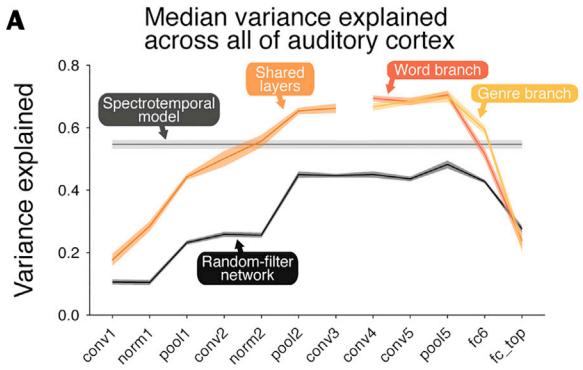
Responses from model units were linearly combined to produce a "synthetic voxel" that best approximated the measured voxel response. This general procedure has become standard for evaluating encoding models of brain responses (Güçlü and van Gerven, 2015; Klindt et al., 2017; Naselaris et al., 2011; Santoro et al., 2014; Yamins et al., 2014), the rationale being that the linear transformation discovered by regression aligns brain and model response spaces as best possible without (nonlinearly) distorting the representations. We measured the amount of BOLD variance predicted in left-out sounds, correcting for both the reliability of the measured voxel response and the reliability of the predicted voxel response (Schoppe et al., 2016; Spearman, 1904). We compared the variance explained by the network features, the spectrotemporal filters in the standard cortical model, and the features from a neural network with untrained (i.e., random) filter weights.

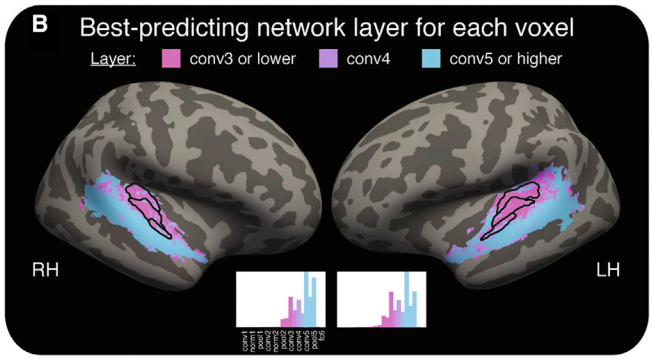
Voxelwise Variance Explained across Auditory Cortex

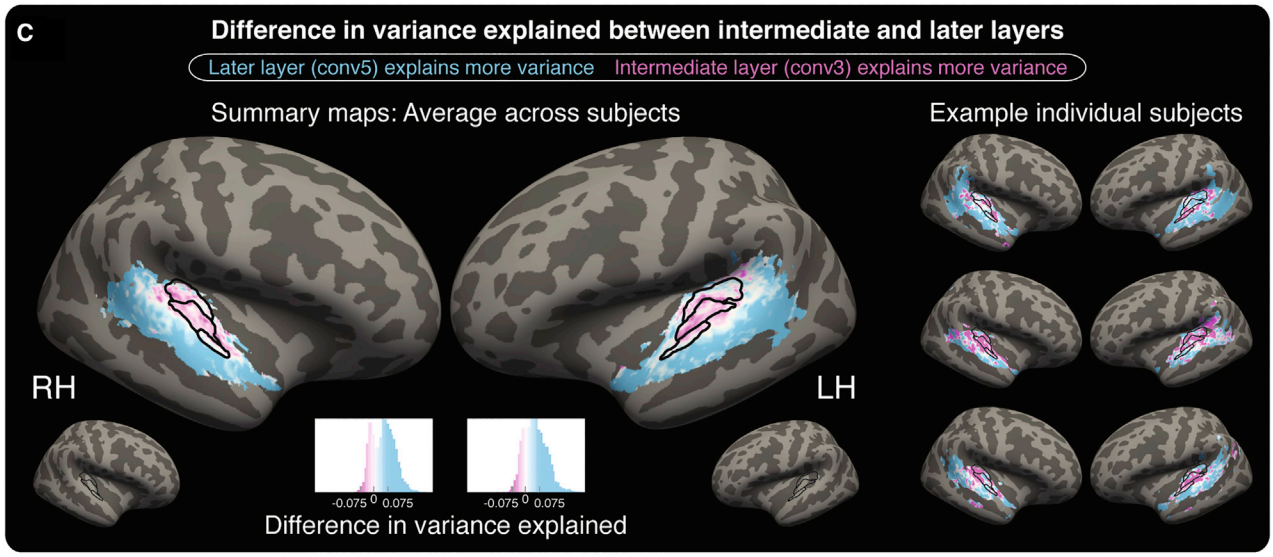
To get an overall sense of the quality of the model predictions, we first computed the median variance explained across all auditory cortical voxels. The network model explained substantially more variance than the spectrotemporal model (69.9% versus 55.1%; $p < 0.001$, paired t test; Figure 3B). We used a relatively large parameterization of the spectrotemporal model that yielded at least as many features as any neural network layer. To ensure a fair comparison, we verified that our parameterization of the spectrotemporal model saturated the amount of variance explainable by such a model for these data (Figure S2C).

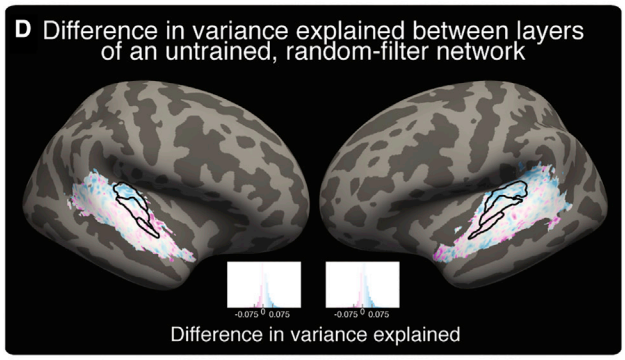
A priori, it seemed possible that the improved voxel predictions could arise merely from the intrinsic hierarchical organization of a deep neural network, irrespective of the particular architecture and filters produced by task optimization. For instance, a feedforward network with convolution and pooling produces a range of receptive field sizes across its layers, some of which might match cortical receptive field sizes. To control for this possibility, we measured predictions from a network with the selected model architecture but with random, untrained weights. As shown in Figure 3B, the random network features explained a substantial fraction of voxel response variance, consistent with previous findings that random nonlinear functions are useful for regression (Lukoševičius and Jaeger, 2009). However, the random network predicted voxel responses substantially worse than the trained network (paired t test on variance explained across all voxels, $p < 0.001$) and the spectrotemporal model (paired t test, $p < 0.05$). Alternative network architectures (also with random weights) yielded even worse response predictions (summarized by the median across architectures, paired t test $p < 0.05$). Overall, these results indicate that task optimization across both architectures and filters was important for achieving good cortical response predictions.

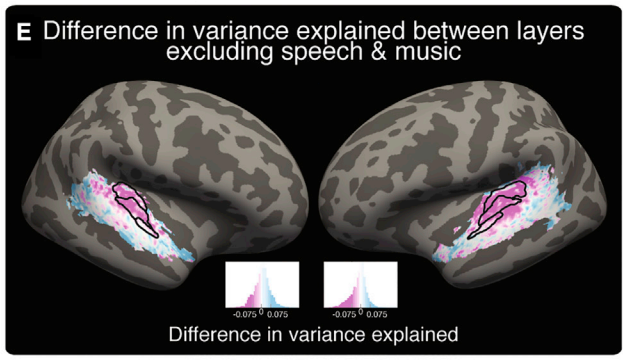
Please cite this article in press as: Kell et al., A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy, Neuron (2018), https://doi.org/10.1016/j.neuron.2018.03.044

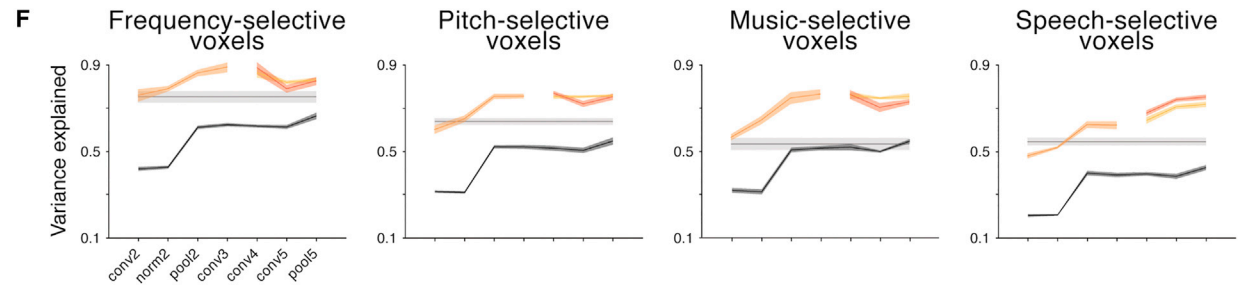












(legend on next page)

Voxelwise Variance Explained in Regions of Interest

To evaluate the quality of voxel response predictions in different parts of auditory cortex, we examined the variance explained within four functionally defined regions of interest (ROIs), localized in each participant with independent data: frequency-, pitch-, music-, and speech-selective voxels (Figure 3B; see STAR Methods for voxel selection details). In all cases, we took the top 5% of all reliable voxels when ranked according to the ROI criterion, excluding any voxels that were included in multiple ROI definitions (Figure S3A). The network model's predictions were better than the standard spectrotemporal filter model and the random filter network in each ROI (paired t tests, $p < 0.001$). Figure S3B shows that results were robust to varying the ROI criterion threshold. Taken together, these results show that our network explains responses to natural sounds substantially better than the spectrotemporal filter model throughout auditory cortex.

Assessing Hierarchical Organization

Because the network transforms its acoustic input via a feedforward cascade, responses of later network layers result from more nonlinear operations than those of earlier layers. We sought to leverage this property to assess the potential hierarchical position of different portions of auditory cortex, comparing voxel responses with responses from different network layers.

Before examining potential differences between different parts of auditory cortex, we first examined the ability of each network layer to predict voxels across all of auditory cortex. The median variance explained increased across layers up until the final two layers, after which it decreased (Figure 4A). Moreover, all but the earliest and latest layers of the trained network surpassed the predictions of the spectrotemporal model. See Figure S2B for significance of individual voxel predictions and Figures S4A and S4B for a map of variance explained across the auditory cortex.

In addition, nearly every layer of the trained network explained more variance than the corresponding layer of the random filter network (paired t tests, all $p < 0.01$, except for final layer), though the dependence on layer for the random filter network was nonetheless coarsely similar to that of the trained network (Pearson's $r = 0.88$, $p < 0.001$). Here, we show the results for one random filter network, but results were similar with different samples of random weights, shown in Figure S2D. These findings are consistent with the idea that the "receptive field" sizes of particular layers (determined by the network architecture) may be well matched to that found in auditory cortical regions but also suggest that task optimization is critical to produce model features that replicate cortical tuning properties. The poor predictions by the final layers of the network could be due to a mismatch in the scale of the resulting features, which pool information across the full duration of the input. The poor predictions could also reflect the fact that final network layers lead up to perceptual decisions. The neurons underlying such decisions might be present in auditory cortex but spatially organized so as to be inaccessible with conventional fMRI, or could reside outside of auditory cortex altogether (e.g., in frontal or parietal areas).

Given that the very early and very late layers were poor predictors of auditory cortical responses, in subsequent analyses we restricted attention to the layers in between.

Hierarchical Organization: Maps

To examine the relationship between the network stages and potential stages of auditory cortex, we determined the best-predicting layer for each reliably sound-responsive voxel in auditory cortex. As shown in Figure 4B, intermediate layers best predicted the voxels in what is classically considered primary ("core") auditory cortex (denoted by the black outlines). Beyond the core, in either anterior, lateral, or posterior directions, the deeper layers provided the best predictions. The results are consistent with the idea that primary and non-primary cortex are situated at distinct hierarchical stages of processing.

Although intuitive, selecting the best-predicting layer for each voxel is categorical and thus does not convey the magnitude of the difference in prediction quality between layers. For a continuous measure of hierarchical position, we additionally measured the difference in variance explained by intermediate and deep layers of the trained network for each voxel. Here we show layers conv5 and conv3, but the general pattern is robust to the particular choice of pairs of layers (Figure S5A). The summary map in Figure 4C averages across individual-subject results, but a similar pattern was evident in individual subjects (Figure 4C, right; see Figure S5B for all eight individuals). The results are again consistent with hierarchical organization. Notably, the analogous map generated from the random-weights network (Figure 4D) does not exhibit signs of hierarchical structure. This result suggests that the differences evident across cortical

Figure 4. Using the Network to Probe Hierarchical Organization in Human Auditory Cortex

(A) Median variance explained across all voxels in auditory cortex for each network layer. Orange denotes results for trained network (break in curve corresponds to branch point). Black denotes results for identical architecture but with random, untrained filter weights. Gray line plots the variance explained by the spectrotemporal filter model for comparison. Error bars are within-subject SEM.

(B) Map of best-predicting layer for each voxel (median across subjects). Inset on right shows color scale and histograms for each hemisphere.

(C) Map of the difference in voxel response variance explained by later and intermediate network layers for the trained network (conv5 versus conv3). Maps on left show mean across subjects; right is three example individual subjects. Below is histogram/color bar of all values for each hemisphere. Black outlines show three anatomically defined sub-divisions of primary auditory cortex (TE 1.1, 1.0, and 1.2), taken from probabilistic maps (Morosan et al., 2001).

(D) Map of the difference in voxel variance explained by higher and lower network layers for the untrained, random-filter network (conv5 versus conv3). Conventions, including color scale, are same as (C).

(E) Map of the difference in variance explained by later and intermediate layers (conv5 versus conv3), excluding speech and music stimuli from the regressions. Analogous to (C), with same conventions and color scale as (C) and (D).

(F) Layerwise variance explained in functionally localized regions of interest for the trained network and random-filter network. Same plotting conventions as (A) were used, except that, for clarity, only intermediate layers (where predictions exceeded those of the spectrotemporal model) are plotted. Error bars are within-subject SEM.

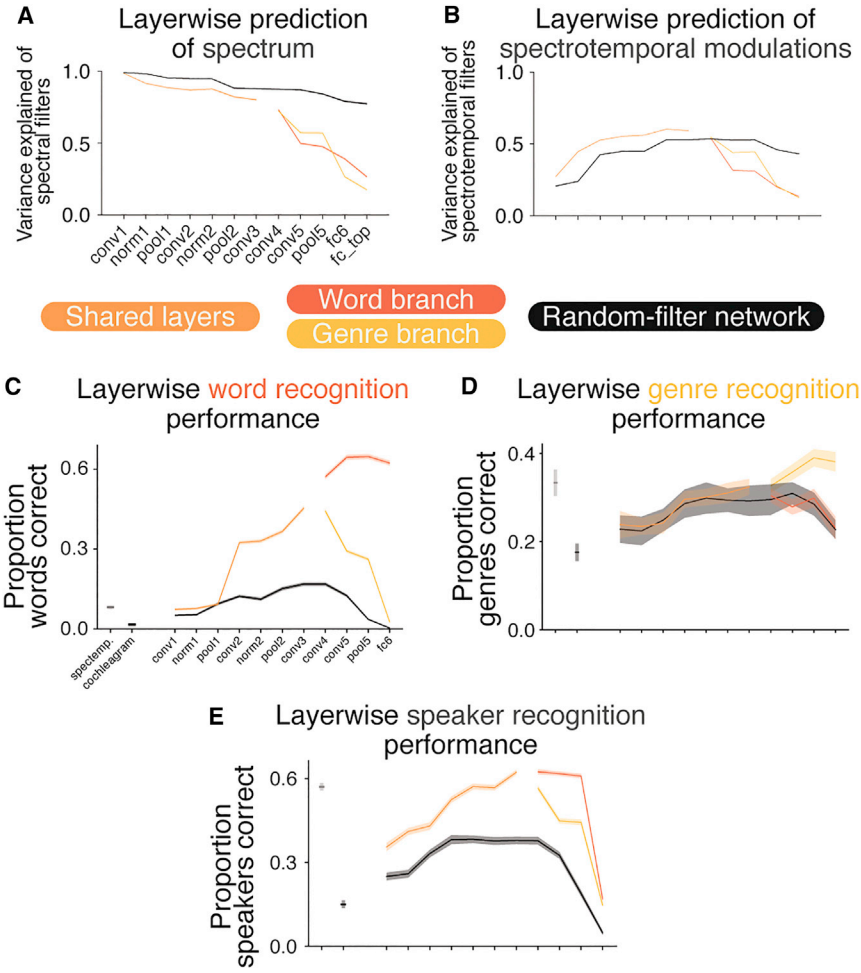


Figure 5. Analysis of Network Representations

(A) Layerwise predictions of the frequency spectrum (as measured by the time-averaged output of the cochlear model that provides the input to the network) from network features. Here and elsewhere, error bars plot one SEM, obtained by bootstrap, and in many cases are so small as to not be clearly visible.

(B) Layerwise predictions of spectrotemporal modulation power (measured from the baseline model shown in Figure 1F) from network features.

(C) Layerwise classification of spoken words using network features.

(D) Layerwise classification of musical genre using network features.

(E) Layerwise classification of speaker identity using network features.

regions reflect tuning properties learned by the network model in the service of auditory task performance.

Given that our network was trained on speech and music tasks, it is natural to wonder whether our evidence for hierarchical cortical organization is simply a reflection of the music and speech selectivity previously reported in non-primary auditory cortex (Angulo-Perkins et al., 2014; Leaver and Rauschecker, 2010; Norman-Haignere et al., 2015; Overath et al., 2015; Tierney et al., 2013). For instance, later layer responses might better serve to detect the presence or absence of speech or music, which could help predict the response of speech- or music-selective voxels. However, even when music and speech stimuli were omitted from the set of 165 sounds (leaving 113 stimuli), the general pattern of results persisted—earlier layers best predicted voxels in the "core," while non-primary voxels were best predicted by later layers (Figure 4E). The hierarchical structure revealed by our network model thus appears to reflect the complexity of cortical responses to everyday sounds more generally.

Hierarchical Organization: Regions of interest

To further assess hierarchical structure, we examined network layer predictions for voxels within the four functionally defined cortical ROIs from Figure 3B. For each network layer, we predicted the responses of individual voxels to all 165 natural sounds and summarized the predictions with the median explained variance across voxels in each ROI. For comparison, we again measured the variance explained by the spectrotemporal filter model and by each layer of the same architecture with random weights. The four ROIs produced distinct layerwise predictivity curves (Figure 4F). Frequency-selective (tonotopic) voxels were best explained by intermediate layers of the network. By contrast, pitch- and music-selective voxels were roughly equally well predicted by intermediate and deep layers. However, the increase in variance explained by the network compared to the spectrotemporal model was significantly larger for music voxels compared to pitch voxels, producing a model-by-region interaction for later network layers ($p < 0.05$ for each layer after pool2). Speech-selective voxels were best explained by deep layers of the network, with a large advantage for the trained network over the spectrotemporal model. Finally, the word and genre branches best predicted speech- and music-selective voxels, respectively, as expected (see Figure S3E for similar results with networks trained on either task individually). Results were again robust to the threshold used to define ROIs (Figure S3C).

The results differed substantially for the random-weight network, whose predictions were consistently lower than those of the trained network (black lines in Figure 4F; paired t test all $p < 0.005$) and were never better than those of the spectrotemporal filter model. Consistent with the map results of Figure 4D, the dependence on network layer did not differ across ROIs (no interaction, $p = 0.99$, unlike the trained network, $p < 0.05$). Indeed, there was little variation in the variance explained by layers beyond pool2 (the slight increase from conv5 to pool5 for the network shown in the figure is inconsistent across multiple random filter initializations; Figure S2D; see also Figure 5 for analogous results with predictions of acoustic features).

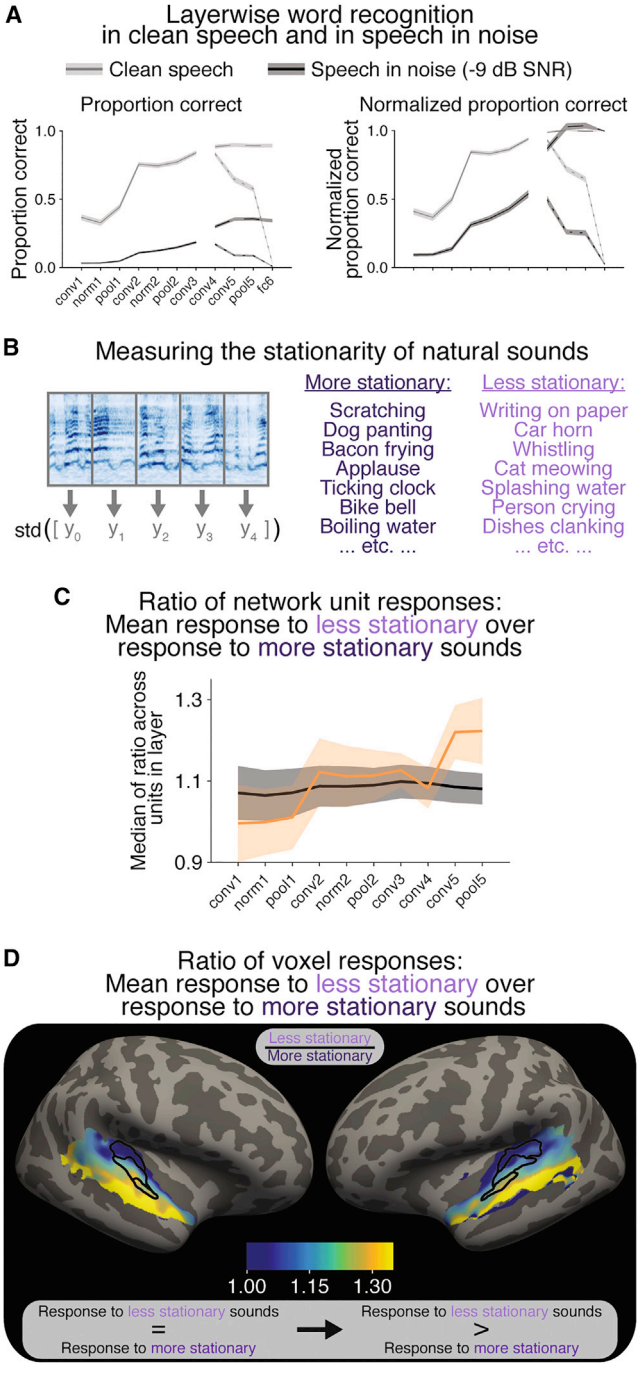


Figure 6. Analysis of Network and Brain Responses to Sound Stationarity

(A) Layerwise classification of spoken words with and without noise. Left graph shows raw performance; to aid comparison of the shape of these curves across layers, right graph shows performance normalized by classification performance of the network's output layer. Here and elsewhere, error bars plot one SEM, obtained by bootstrap.

(B) Schematic of stationarity measure, based on the variability of texture statistics measured in short time windows.

(C) Ratio of mean network unit responses to most and least stationary sounds selected from a set of natural sounds. Error bars are larger than in other plots due to the modest size of the sound sets.

(D) Ratio of voxel responses to the same sets of most and least stationary sounds.

These results suggest that task optimization was critical to instantiating differentiated features across layers that were well matched to auditory cortical tuning.

Overall, the ROI analyses provide further support for hierarchical organization. The results are consistent with the idea that tonotopic voxels (best explained by intermediate layers of the network) are situated early in a cortical hierarchy. By contrast, pitch-, music-, and speech-selective voxels appear to be situated later, best explained by later layers that instantiate more complex functions of the acoustic input.

Analysis of Network Representations

The network and cortex both appear to be organized hierarchically, but what representational transformations do they instantiate? To explore this question, we took advantage of the ability to interrogate the model representations at will.

Predicting Acoustic Features

We first examined the network's representation of standard acoustic features: those captured by a model of the cochlea (which provides the network's input) and the baseline spectrotemporal modulation model (Figure 1F). We examined whether these two types of information were explicit in each layer's representation (i.e., linearly decodable), using a similar procedure to that which we employed to predict voxel responses. We fit a linear mapping from the network features to the acoustic features using one subset of natural sounds and measured the quality of the resulting predictions with another subset. The ability to extract spectral information decreased from early to late layers of the network (Figure 5A), whereas the ability to extract spectrotemporal modulations peaked in intermediate layers (Figure 5B). Later layers predicted both sets of features worse than earlier layers, and this decrease was accentuated in the trained network compared to an untrained, random-filter network, indicating that it is not simply due to the network architecture.

Real-World Task Performance

We next tested the extent to which the network features in each layer could be used to perform tasks: the word and genre tasks that the network was trained on and, to examine generalization, a speaker identification task that played no role in training. We examined performance for each layer by fixing all the weights in the network and optimizing a linear (softmax) classifier that took the output of a given layer as its input. Unlike predictions of standard acoustic features, word and genre classification improved from early to late network layers, differing substantially between the two task-specific branches (Figures 5C and 5D). With the exception of the very last layer of the network, this pattern also held for the speaker task (Figure 5E). The speaker classification performance suggests that the learned network representations generalize at least somewhat to other tasks. Taken together, these analyses indicate that task-related information implicit in the cochlear representation is gradually transformed to become more explicit (see Figure S6 for complementary results using representational similarity analysis).

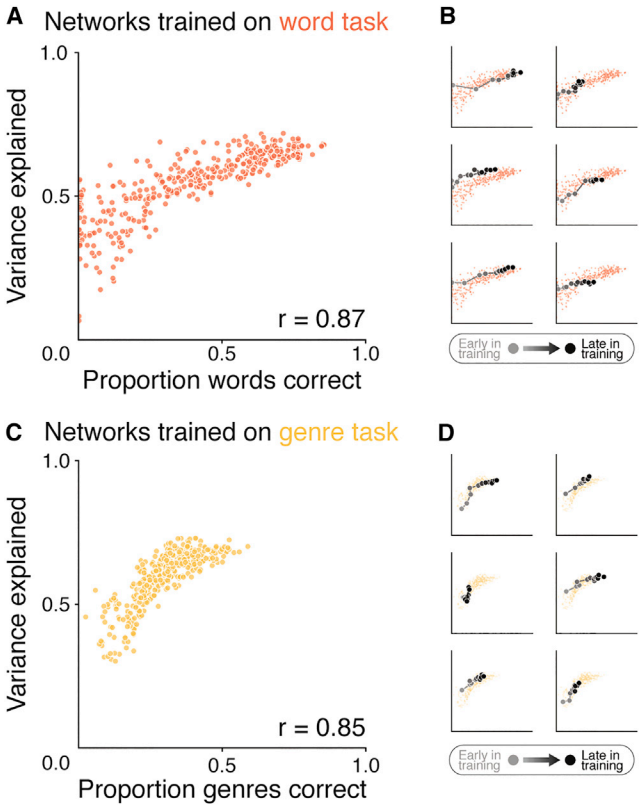


Figure 7. Network Task Performance Correlates with Cortical Predictivity

(A) Scatterplot of the median variance explained by a network across all voxels in auditory cortex versus the proportion of words correctly identified by that network, for networks trained on the spoken word recognition task. Each point is a network (a certain architecture with a certain amount of training). The networks were only optimized for task performance (i.e., the x axis).

(B) Same data as in (A), but with a single network architecture's performance and predictivity evaluated over training epochs plotted in gray/black. Earlier training points are gray; later training points are black. Each panel highlights a different architecture.

(C) Analogous to (A), but for networks trained on the musical genre recognition task. Conventions are as in (A).

(D) Analogous to (B), but for networks trained on the genre task. Conventions are same as (B).

Sensitivity to Noise-like Sounds: Testing a Model Prediction

To further characterize the network's representational transformations, we examined how background noise affected task performance in each layer. We took the results from the word classifiers used in Figure 5C and split them up according to the amount of background noise. Figure 6A shows that in the absence of noise, spoken words could be classified relatively well from intermediate network layers (e.g., conv2 or norm2). By contrast, classification of speech in noise did not approach asymptotic performance levels until later network layers (e.g., conv4 or conv5), indicating that representations in later layers are more noise-robust (see Figure S7A for effect of intermediate SNRs).

The noise robustness of the later layers motivated us to examine their sensitivity to noise-like stimuli. We measured the response in each network layer to two sets of natural sounds: one with relatively stable statistical properties (i.e., stationary and thus noise-like), and one with less stable statistics. We quantified stationarity by dividing a sound's cochleagram into temporal bins, extracting perceptually relevant sound statistics in each bin, and taking the standard deviation over time (Figure 6B; see STAR Methods for details). The stimuli were subsets of those from the fMRI experiment, but with speech and music excluded to ensure that any observed effect was not simply due to selectivity for these sound classes. We then measured each network layer's mean response to the two sets of natural sounds. As shown in Figure 6C, deeper layers of the trained network exhibited a larger response to non-stationary compared to stationary sounds. This effect was not observed for the untrained network, suggesting it was not simply due to the extent of spectral and temporal integration in later layers.

To examine whether a similar effect differentiated primary from non-primary cortex, we compared voxel responses to the same sets of sounds used in the network analysis (taken from the fMRI dataset used throughout this paper). As shown in Figure 6D, responses to stationary and non-stationary sounds were comparable in and around primary auditory cortex but diverged in non-primary areas, with higher responses to non-stationary sounds (see Figure S7C for maps in individuals). Because speech and music were excluded from the analysis, this result is not simply a reflection of speech and music selectivity in non-primary auditory cortex. The result is suggestive of a suppression of noise-like sounds in later stages of the auditory hierarchy and may in part explain the general ability of the network model to predict cortical responses to natural sounds.

Relationship between Cortical Predictions and Task Performance

Given that untrained networks predicted voxel responses substantially worse than a task-optimized network (Figure 3B), it is natural to wonder how a network's task performance relates to its ability to predict cortical responses. Previous work found that networks with better performance on a real-world visual object recognition task better predict cortical responses in the ventral visual stream (Yamins et al., 2014). We explored whether a similar relationship might hold in the auditory system.

We examined task performance and cortical prediction quality for 57 different architectures at 14 different time points during task training (yielding a total of 798 different neural networks). Each network was trained for either word recognition or genre classification (with a single, unbranched processing stream), and we measured how well each network performed the task for which it was trained (with stimuli not used for training). Additionally, we measured the median variance explained by each network across all voxels in auditory cortex. For each voxel, we selected each network's best-predicting layer and evaluated that layer's variance explained in left-out data. The performance of a network on a task strongly correlated with the variance it explained in auditory cortical responses (Figure 7A, word task, Spearman r = 0.87; Figure 7C for genre task, Spearman r = 0.85; both $p < 10^{-100}$; example trajectories of individual architectures over the course of training are shown in Figures 7B and

7D). These findings suggest that task-based optimization of deep neural networks can help yield more predictive models of sensory systems.

DISCUSSION

We developed a quantitative model of auditory computation by optimizing a deep neural network on real-world speech and music tasks. The optimization yielded a model architecture with separate music and speech pathways following a shared front end, potentially replicating one aspect of human cortical organization. The model performed both tasks as well as human listeners, exhibited patterns of behavioral errors like those of humans, and predicted fMRI responses throughout auditory cortex substantially better than a standard cortical model in common use. Task optimization was both necessary and sufficient to achieve the best existing predictions of auditory cortex: necessary in that hierarchical model structure without task optimization yielded poor predictions, and sufficient in that the model was only optimized to perform the tasks and was not otherwise constrained to match human behavior or brain responses. The resulting model provided evidence for hierarchical organization within auditory cortex—intermediate model layers best predicted primary auditory cortex, while deeper layers best predicted non-primary responses. The differentiation between primary and non-primary cortex appears to not simply be a function of pooling information over larger regions of time and/or frequency, in that a network with the same architecture but random weights did not produce the same result. Analyses of the network suggested a hypothesis about cortical processing, which we then tested, finding that non-primary auditory cortex is less responsive to noise-like natural sounds. Despite being trained on speech and music tasks, the key scientific findings from the model were robust to the exclusion of cortical responses to speech and music from the analysis, indicating that the apparent hierarchical structure is somewhat general and is not simply a reflection of neural selectivity for speech and music. Taken together, these results suggest that real-world tasks may substantially constrain both neural processing and behavior, and that task optimization may provide a powerful approach to developing models of neural systems.

A New Model of Auditory Cortex

Our modeling methodology is distinct from traditional approaches rooted in physiological observations or signal processing principles, and the resulting model differs from its predecessors in many respects. Our model is substantially deeper than previous models, with twelve layers of computation following peripheral auditory processing—others have had one or two stages at most (Carlson et al., 2012; Chi et al., 2005; Dau et al., 1997; McDermott and Simoncelli, 2011; Młynarski and McDermott, 2018). This increase in depth aids the compact expression of successively richer sets of functions of the audio input (Montufar et al., 2014). These rich functions are useful for good real-world task performance, and they also enable improved prediction of cortical responses.

We also introduced a method for learning multiple task-specialized pathways, which produced speech- and music-specific processing streams following several shared stages. This architecture presumably reflects partially overlapping demands of speech and music processing and is consistent with recent evidence for functional segregation in non-primary auditory cortex (Angulo-Perkins et al., 2014; Leaver and Rauschecker, 2010; Norman-Haignere et al., 2015; Tierney et al., 2013). Applying this approach with additional tasks (environmental sound recognition, sound localization, etc.) could help reveal the computational relationship between tasks and generate additional hypotheses about functional segregation and pathway organization in the brain.

Compared to more traditional hand-engineered models (Chi et al., 2005; Dau et al., 1997; McDermott and Simoncelli, 2011), one disadvantage of our model is that individual units are less readily understood. However, we emphasize that the deep neural networks evaluated here have approximately the same number of free parameters for predicting voxel data as do more traditional hand-engineered models, such as the spectrotemporal filter model employed in our work. All the nonlinear neural network parameters were determined by task optimization alone; the only parameters fitted to voxel data were those of the linear map from model units to voxels. Furthermore, given that we evaluated prediction accuracy with left-out stimuli, a larger number of parameters in and of itself will not, in general, lead to better predictions. The deep neural network's training task can be considered a normative constraint—i.e., a hypothesis about the phylogenetic and/or ontogenetic pressures that shape human listeners' behavior and auditory cortical representations.

Deep Neural Networks as a Model of Human Perceptual Judgments

Perhaps the most significant departure from previous auditory models is that our model performs real-world tasks on par with human listeners. Achieving human performance on everyday perceptual tasks was unheard of until just a few years ago but is now attainable in an increasing number of domains due to the efficacy of deep learning. Real-world task performance increases the plausibility of a model, as any complete model of the auditory system should perform the tasks that humans perform. It also enables model evaluation via human model comparisons of behaviors that matter for everyday listening. To our knowledge, our model is the first to provide a detailed match to patterns of human performance across conditions on real-world auditory tasks. We note that the comparison of network performance and human behavior involved zero free parameters—we simply measured the performance of the model and of human listeners in different conditions.

How should we interpret the similarity of performance? One possibility is that both the human and the network are approaching performance limits inherent to the tasks, and thus any model reaching human-level task performance would also exhibit human-like error patterns. Alternatively, the human network similarity could reflect algorithmic similarity, such that there could exist models with human-level performance but with different error patterns. Additional model classes (currently unavailable) will be necessary to disambiguate these alternatives.

Independent of the interpretation, the behavioral similarity between humans and the network lends strength to the goal-driven modeling enterprise.

Improved Cortical Response Predictions

The network predicted cortical responses better than the standard spectrotemporal filter model in both primary and non-primary auditory cortex (Figure 3B). The improved predictions in primary regions suggest that even primary sensory cortex may be better understood by considering how task demands shape representations. Given prior evidence of primary cortical neurons' tuning to spectrotemporal modulations (Santoro et al., 2014; Schönwiesner and Zatorre, 2009), the improved predictions from the network could reflect relatively simple nonlinear functions of spectrotemporal filter responses, such as normalization or pooling. These simple operations are absent from the standard spectrotemporal auditory model but present in our network model.

Evidence for Hierarchical Organization of Auditory Cortex

Investigators have long been intrigued by the idea of hierarchical structure in auditory cortex (Boemio et al., 2005; Okada et al., 2010; Rauschecker and Scott, 2009; Rauschecker et al., 1995; Recanzone and Cohen, 2010). Anatomical data in non-human primates suggest a division into three stages—core, belt, and parabelt (Kaas and Hackett, 2000), which differ in tuning properties (Rauschecker et al., 1995; Recanzone and Cohen, 2010) and response latencies (Camalier et al., 2012). But even in non-human animals the divisions and functions of processing stages remain debated. Moreover, it is not obvious how auditory cortical organization in non-human animals may apply to humans, in part because speech and music are both uniquely human and central to human hearing.

In human auditory cortex, hierarchy is most often considered for speech processing. Speech-selective responses emerge only in non-primary areas (de Heer et al., 2017; Evans and Davis, 2015; Mesgarani et al., 2014; Okada et al., 2010; Overath et al., 2015) and cannot be accounted for by modulation features standardly used to model primary areas (Norman-Haignere et al., 2015; Overath et al., 2015). However, large swaths of non-primary human auditory cortex are not speech selective (Norman-Haignere et al., 2015; Overath et al., 2015), leaving the generality of any potential multi-stage organization an open question. The extent and nature of hierarchical organization in humans has thus remained unsettled (Formisano et al., 2008; Leaver and Rauschecker, 2010; Staeren et al., 2009; Wessinger et al., 2001). Part of the difficulty is that it is not obvious how to assess hierarchy without knowledge of fine-grained connectivity between regions, which is not presently available for human auditory cortex (Cammoun et al., 2015).

We propose an alternative method for evaluating hierarchy, using a hierarchical model to operationalize the "complexity" of neuronal tuning. We compare brain and model responses to the same natural sounds, using the similarity between the two at each stage of the model as a measure of neuronal response complexity. This model-based approach provides an advance over previous, more subjective notions of response complexity (Rauschecker et al., 1995), though it requires a rich experimental dataset to distinguish response properties at different model stages. Our analyses thus far do not yield compelling evidence that human auditory cortex exhibits the tripartite hierarchical organization commonly proposed in other animals (i.e., core, belt, and parabelt), but such organization may become evident when our methodology is applied to additional neural datasets, or with a more refined model.

Similarities and Differences with the Visual System

Task-optimized artificial neural networks have recently proven to be powerful models of visual cortex, in part because they recapitulate aspects of the hierarchical structure of the ventral visual stream. Because the visual cortical regions and pathways are well established, this similarity was less a novel scientific finding than an effective way of validating the task-driven modeling approach (Cichy et al., 2016; Güçlü and van Gerven, 2015; Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014). By contrast, less is known about the organization of the auditory cortex, and our model provides novel evidence for hierarchical cortical processing.

One difference between our results and analogous work in the visual system is that primary visual cortical responses have been found to be best predicted by early layers of a deep network (Cadena et al., 2017; Güçlü and van Gerven, 2015; Khaligh-Razavi and Kriegeskorte, 2014), whereas we found primary auditory cortical responses to be best predicted by intermediate network layers. This difference is consistent with the idea that primary auditory cortex may be situated later in a computational hierarchy than primary visual cortex, potentially due to the existence of more subcortical nuclei in the auditory system (King and Nelken, 2009). Further investigating this question, for instance, by predicting subcortical responses with early network layers, is a promising future direction.

Limitations and Future Directions

Although the network better accounts for human behavior and cortical responses compared to previous models, there are many limitations that motivate future work. First, the match to human behavior is imperfect. The genre task exhibited the largest discrepancies, perhaps because it is not critically important for humans and/or may be significantly influenced by cultural experience (it was chosen primarily because it is the only music-related task for which a suitably large database of labeled samples is readily available). Training networks on additional music-related tasks, or tasks not specific to speech or music, could yield a more complete model of human behavior.

Second, the trained network does not explain all of the reliable response variance in the BOLD signal. Some of the remaining variance may reflect the importance of additional tasks, limitations of the regression procedure for mapping between the network model and the brain (Klindt et al., 2017), representations not easily learned in discriminative (e.g., classification-trained) models, or the presence of computations not easily implemented in a feedforward architecture. Finer-grained brain data (e.g., higher-field MRI, electrocorticography, or single-unit physiology) could reveal additional limitations, likely including the absence of realistic temporal dynamics and the lack of a role for behavioral goals or cortical state (e.g., arousal). Such phenomena may

require extending the network architecture to include recurrent dynamics, feedback connections, and/or attention.

Finally, although we propose a model of auditory cortex shaped by task demands, the learning procedure we employ to satisfy those demands likely deviates substantially from biological learning. Humans almost surely do not require millions of labeled examples to learn to recognize words, and instead presumably use some mixture of supervised, unsupervised, and reinforcement learning. The similarity we observe between our model and the human auditory system therefore suggests that systems can arrive at similar final states via different learning algorithms. Future developments in semi- and self-supervised learning will hopefully enable models of human behavior and cortical responses to be learned via more biologically and ecologically realistic procedures.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Word & genre recognition psychophyiscs
  - fMRI cortical responses to natural sounds
- METHOD DETAILS
  - Tasks for network training
  - Definition of constituent operations of a convolutional neural network (CNN)
  - CNN architectural optimization and filter weight training
  - Psychophysics comparing human listeners and the network
  - fMRI data analysis: Preprocessing and voxel selection
  - Using neural network layers as voxelwise encoding models to predict cortical responses
  - Examining representations in the network
  - Analyzing network responses to more and less stationary sounds
  - Examining relationship between network task performance and auditory cortical variance explained
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Branch point selection
  - Psychophysics statistics
  - ROI analysis statistics
  - Network analysis statistics
  - Significance of individual voxel predictions
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and three tables and can be found with this article online at https://doi.org/10.1016/j.neuron.2018.03.044.

ACKNOWLEDGMENTS

The authors thank Nancy Kanwisher for sharing fMRI data, Steve Shannon for MR support, Kevin Woods for analysis of speakers for the representational similarity analysis, Ariel Herbert-Voss for help collecting behavioral data, Atsushi Takahashi for assistance with MR acquisition protocol design, and Satra Ghosh and the OpenMind team for computing resources. The authors also thank Michael Cohen, Elias Issa, Rebecca Saxe, Pedro Tsidivis, and members of the McDermott lab for comments on the manuscript. This work was supported by an NVIDIA Corporation hardware donation, NIH grant NEI-R01 EY014970 to J. DiCarlo (supporting D.L.K.Y.), a DOE Computational Science Graduate Fellowship (DE-FG02-97ER25308) to A.J.E.K., a McDonnell Scholar Award to J.H.M., and NSF grant BCS-1634050 to J.H.M.

AUTHOR CONTRIBUTIONS

Conceptualization, A.J.E.K., D.L.K.Y., and J.H.M.; Methodology, A.J.E.K. and D.L.K.Y.; Investigation, A.J.E.K., E.N.S., and S.V.N.-H.; Software, A.J.E.K. and D.L.K.Y.; Visualization, A.J.E.K. and E.N.S.; Writing – Original Draft, A.J.E.K. and J.H.M.; Writing – Review & Editing, A.J.E.K., D.L.K.Y., E.N.S., S.V.N.-H., and J.H.M.; Supervision and Funding Acquisition, J.H.M.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 18, 2017
Revised: December 22, 2017
Accepted: March 23, 2018
Published: April 19, 2018

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., and Kudlur, M. 2016. TensorFlow: a system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), pp. 265-283.

Angulo-Perkins, A., Aubé, W., Peretz, I., Barrios, F.A., Armony, J.L., and Concha, L. (2014). Music listening engages specific cortical regions within the temporal lobes: Differences between musicians and non-musicians. Cortex *59*, 126–137.

Atencio, C.A., Sharpee, T.O., and Schreiner, C.E. (2012). Receptive field dimensionality increases from the auditory midbrain to cortex. J. Neurophysiol. *107*, 2594–2603.

Bertin-Mahieux, T., Ellis, D.P., Whitman, B., and Lamere, P. (2011). The Million Song Dataset. In International Society for Music Information Retrieval, pp. 591–596.

Boemio, A., Fromm, S., Braun, A., and Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nat. Neurosci. *8*, 389–395.

Brainard, D.H. (1997). The psychophysics toolbox. Spat. Vis. *10*, 433–436.

Cadena, S.A., Denfield, G.H., Walker, E.Y., Gatys, L.A., Tolias, A.S., Bethge, M., and Ecker, A.S. (2017). Deep convolutional models improve predictions of macaque V1 responses to natural images. bioRxiv.

Camalier, C.R., D'Angelo, W.R., Sterbing-D'Angelo, S.J., de la Mothe, L.A., and Hackett, T.A. (2012). Neural latencies across auditory cortex of macaque support a dorsal stream supramodal timing advantage in primates. Proc. Natl. Acad. Sci. USA *109*, 18168–18173.

Cammoun, L., Thiran, J.P., Griffa, A., Meuli, R., Hagmann, P., and Clarke, S. (2015). Intrahemispheric cortico-cortical connections of the human auditory cortex. Brain Struct. Funct. *220*, 3537–3553.

Carlson, N.L., Ming, V.L., and Deweese, M.R. (2012). Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. PLoS Comput. Biol. *8*, e1002594.

Chang, E.F., Rieger, J.W., Johnson, K., Berger, M.S., Barbaro, N.M., and Knight, R.T. (2010). Categorical speech representation in human superior temporal gyrus. Nat. Neurosci. *13*, 1428–1432.

Chechik, G., Anderson, M.J., Bar-Yosef, O., Young, E.D., Tishby, N., and Nelken, I. (2006). Reduction of information redundancy in the ascending auditory pathway. Neuron *51*, 359–368.

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

Christianson, G.B., Sahani, M., and Linden, J.F. (2008). The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. J. Neurosci. *28*, 446–455.

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci. Rep. *6*, 27755.

Dau, T., Kollmeier, B., and Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. J. Acoust. Soc. Am. *102*, 2892–2905.

David, S.V., Mesgarani, N., Fritz, J.B., and Shamma, S.A. (2009). Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. J. Neurosci. *29*, 3374–3386.

de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., and Theunissen, F.E. (2017). The hierarchical cortical organization of human speech processing. J. Neurosci. *37*, 6539–6557.

Depireux, D.A., Simon, J.Z., Klein, D.J., and Shamma, S.A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J. Neurophysiol. *85*, 1220–1234.

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

Evans, S., and Davis, M.H. (2015). Hierarchical organization of auditory and motor representations in speech perception: Evidence from searchlight similarity analysis. Cereb. Cortex *25*, 4772–4788.

Fischl, B. (2012). FreeSurfer. Neuroimage *62*, 774–781.

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. Science *322*, 970–973.

Garofolo, J.S., and Consortium, L.D. (1993). TIMIT: Acoustic-Phonetic Continuous Speech Corpus (Linguistic Data Consortium).

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

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

Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al. (2017). CNN architectures for large-scale audio classification. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131–135.

Humphries, C., Liebenthal, E., and Binder, J.R. (2010). Tonotopic organization of human auditory cortex. Neuroimage *50*, 1202–1211.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., and Gallant, J.L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature *532*, 453–458.

Jones, E., Oliphant, T.E., and Peterson, P. (2001). SciPy: Open source scientific tools for Python.

Kaas, J.H., and Hackett, T.A. (2000). Subdivisions of auditory cortex and processing streams in primates. Proc. Natl. Acad. Sci. USA *97*, 11793–11799.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. (2008). TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In Acoustics, Speech and Signal Processing (IEEE Internal Conference), pp. 3933–39355.

Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. *10*, e1003915.

King, A.J., and Nelken, I. (2009). Unraveling the principles of auditory cortical processing: can we learn from the visual system? Nat. Neurosci. *12*, 698–701.

Klindt, D., Ecker, A.S., Euler, T., and Bethge, M. (2017). Neural system identification for large populations separating "what" and "where". arXiv, arXiv:1711.02653, https://arxiv.org/abs/1711.02653.

Leaver, A.M., and Rauschecker, J.P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J. Neurosci. *30*, 7604–7612.

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

Lehky, S.R., and Sejnowski, T.J. (1988). Network model of shape-from-shading: neural function arises from both receptive and projective fields. Nature *333*, 452–454.

Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T., and Medler, D.A. (2005). Neural substrates of phonemic perception. Cereb. Cortex *15*, 1621–1631.

Lukoševičius, M., and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. Comput. Sci. Rev. *3*, 127–149.

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

McDermott, J.H., Schemitsch, M., and Simoncelli, E.P. (2013). Summary statistics in auditory perception. Nat. Neurosci. *16*, 493–498.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E.F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science *343*, 1006–1010.

Miller, L.M., Escabí, M.A., Read, H.L., and Schreiner, C.E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J. Neurophysiol. *87*, 516–527.

Młynarski, W., and McDermott, J.H. (2018). Learning midlevel auditory codes from natural sound statistics. Neural Comput. *30*, 631–669.

Montufar, G., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. In Advances in Neural Information Processing Systems 27 (NIPS 2014), Z. Gharamani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, eds. (Curran Associates), pp. 2924–2932.

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., and Zilles, K. (2001). Human primary auditory cortex: Cytoarchitectonic subdivisions and mapping into a spatial reference system. Neuroimage *13*, 684–701.

Mustafa, K., and Bruce, I.C. (2006). Robust formant tracking for continuous speech with speaker variability. IEEE Trans. Audio Speech Lang. Process. *14*, 435–444.

Naselaris, T., Kay, K.N., Nishimoto, S., and Gallant, J.L. (2011). Encoding and decoding in fMRI. Neuroimage *56*, 400–410.

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

Norman-Haignere, S., Kanwisher, N.G., and McDermott, J.H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron *88*, 1281–1296.

Obleser, J., Leaver, A.M., Vanmeter, J., and Rauschecker, J.P. (2010). Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. Front. Psychol. *1*, 232.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.H., Saberi, K., Serences, J.T., and Hickok, G. (2010). Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cereb. Cortex *20*, 2486–2495.

Oliphant, T.E. (2006). Guide to NumPy (Brigham Young University).

Overath, T., McDermott, J.H., Zarate, J.M., and Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat. Neurosci. *18*, 903–911.

Patil, K., Pressnitzer, D., Shamma, S., and Elhilali, M. (2012). Music in our ears: the biological bases of musical timbre perception. PLoS Comput. Biol. *8*, e1002759.

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

Peelle, J.E., Johnsrude, I.S., and Davis, M.H. (2010). Hierarchical processing for speech in human auditory cortex and beyond. Front. Hum. Neurosci. *4*, 51.

Pinto, N., Doukhan, D., DiCarlo, J.J., and Cox, D.D. (2009). A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS Comput. Biol. *5*, e1000579.

Rauschecker, J.P., and Scott, S.K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci. *12*, 718–724.

Rauschecker, J.P., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. Science *268*, 111–114.

Razavian, A.S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. In IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 512–519.

Recanzone, G.H., and Cohen, Y.E. (2010). Serial and parallel processing in the primate auditory cortex revisited. Behav. Brain Res. *206*, 1–7.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., and Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput. Biol. *10*. Published online January 2, 2014. https://doi.org/10.1371/journal.pcbi.1003412.

Schönwiesner, M., and Zatorre, R.J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc. Natl. Acad. Sci. USA *106*, 14611–14616.

Schoppe, O., Harper, N.S., Willmore, B.D.B., King, A.J., and Schnupp, J.W.H. (2016). Measuring the performance of neural models. Front. Comput. Neurosci. *10*. Published online February 10, 2016. https://doi.org/10.3389/fncom.2016.00010.

Spearman, C. (1904). The proof and measurement of association between two things. Am. J. Psychol. *15*, 72–101.

Spearman, C. (1910). Correlation calculated from faulty data. Br. J. Psychol. *3*, 271–295.

Staeren, N., Renvall, H., De Martino, F., Goebel, R., and Formisano, E. (2009). Sound categories are represented as distributed patterns in the human auditory cortex. Curr. Biol. *19*, 498–502.

Tierney, A., Dick, F., Deutsch, D., and Sereno, M. (2013). Speech versus song: multiple pitch-sensitive areas revealed by a naturally occurring musical illusion. Cereb. Cortex *23*, 249–254.

Uppenkamp, S., Johnsrude, I.S., Norris, D., Marslen-Wilson, W., and Patterson, R.D. (2006). Locating the initial stages of speech-sound processing in human temporal cortex. Neuroimage *31*, 1284–1296.

Wessinger, C.M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., and Rauschecker, J.P. (2001). Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. J. Cogn. Neurosci. *13*, 1–7.

Woods, K.J.P., Siegel, M.H., Traer, J., and McDermott, J.H. (2017). Headphone screening to facilitate web-based auditory experiments. Atten. Percept. Psychophys. *79*, 2064–2072.

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

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

Zipser, D., and Andersen, R.A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. Nature *331*, 679–684.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q.V. (2017). Learning transferable architectures for scalable image recognition. arxiv, arXiv:1707.07012, https://arxiv.org/abs/1707.07012.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Numpy	Oliphant, 2006	http://www.numpy.org/
Scipy	Jones et al., 2001	https://www.scipy.org/
TensorFlow	Abadi, et al., 2016	https://www.tensorflow.org/
FreeSurfer	Fischl, 2012	http://www.freesurfer.net
Psychophysics Toolbox	Brainard, 1997	http://psychtoolbox.org/
NSL MATLAB Toolbox	Chi et al., 2005	http://www.isr.umd.edu/Labs/NSL/Software.htm
Other		
Wall Street Journal Speech Corpus	Paul and Baker, 1992	https://catalog.ldc.upenn.edu/ldc93s6a
TIMIT Speech Corpus	Garofolo and Consortium, 1993	https://catalog.ldc.upenn.edu/ldc93s1
Million Song Dataset	Bertin-Mahieux et al., 2011	https://labrosa.ee.columbia.edu/millionsong/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Alex Kell (alexkell@mit.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Word & genre recognition psychophyiscs

For the word recognition psychophysics, eighteen subjects participated (12 female, mean age: 23 years, range: 18-33 years). For the genre recognition psychophysics, we performed experiments both in-lab (n = 31; 20 female; mean age: 26 years, range: 19-43) and on Amazon's Mechanical Turk (n = 80; 26 female; mean age: 34 years, range: 19-67). Mechanical Turk was used to supplement the in-lab data once it became clear that it would be useful to estimate a confusion matrix (which required a relatively large number of participants). All subjects had self-reported normal hearing, and provided informed consent. The Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects approved experiments.

fMRI cortical responses to natural sounds

The fMRI data analyzed here is a subset of the data in Norman-Haignere et al. (2015), only including the subjects who completed three scanning sessions. Eight participants (four female, mean age: 22 years, range: 19-25; all right-handed; one participant was author SNH) completed three scanning sessions (each ~2 hours). Subjects were non-musicians (no formal training in the five years preceding the scan), native English speakers, and had self-reported normal hearing. Two other subjects only completed two scans and were excluded from these analyses, and three additional subjects were excluded due to excessive head motion or inconsistent task performance. The decision to exclude these five subjects was made before analyzing any of their fMRI data. All participants provided informed consent, and the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects approved experiments.

METHOD DETAILS

Tasks for network training

The training datasets consisted of more than two and a half million labeled exemplars for each task, where each exemplar was a clip of speech or music excerpted from a large, labeled corpus and embedded in background noise (Figure 1A).

Word task

The word task was a 587-way classification task. We counted the occurrence of all words in the TIMIT (Garofolo and Consortium, 1993) and WSJ (Paul and Baker, 1992) corpora, and included all words at least four characters long that were uttered between 25-100 times in the TIMIT corpus or 75-200 times in the WSJ corpus. The lower limit was intended to ensure a sufficient number of examples per word; the upper limit was intended to help create training sets with some degree balance in the number of examples of each word. These criteria yielded 587 unique words (see Table S1 for list of all words). We then excerpted two-second clips from

these corpora in which one of the 587 target words occurred during the halfway point of the clip (i.e., the word overlapped the one-second mark; note that it did not have to be centered at one second and in general was not). We generated more than two and a half million total excerpts. To make the classification task both more realistic and difficult, we superimposed these speech excerpts on one of three different kinds of background noise: (1) "auditory scenes," (2) music, or (3) two-speaker speech "babble." Auditory scenes were real-world recordings of everyday acoustic environments (e.g., a bus station, an office, a restaurant, a supermarket), and were drawn from the corpus used for the 2013 IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events. Music exemplars were drawn from an in-house corpus of monophonic and polyphonic instrumental music recordings (solo guitar, orchestral ensemble, big band, movie soundtracks, etc.). Speech babble was generated by mixing together a pair of speakers randomly drawn from a large corpus of public domain audiobook recordings (https://librivox.org/). These background clips were added to the speech excerpts at signal-to-noise ratios selected to roughly match recognition difficulty across background type for human listeners (based on pilot psychophysical experiments). To increase heterogeneity within the training stimuli, the exact signal-to-noise ratio (SNR) for each exemplar was drawn randomly from a Gaussian with a standard deviation of 2 dB SNR and a mean of −3 dB (for auditory scenes and speech babble) or −6 dB (for music).

Genre task

The genre task was a 41-way classification task. The Million Song Dataset (Bertin-Mahieux et al., 2011) consists of nearly a million tracks of various types. More than 300,000 of these tracks have user-generated "tags" from the MusicBrainz open-source music encyclopedia (https://musicbrainz.org/). Tags are at the artist level, not the track level (e.g., all of Marvin Gaye's tracks have the same tags), and these 300,000 tracks are by ~9,000 unique artists. There were 2,321 unique tags for these artists, but many of them occurred with low frequency. We first culled all tags that were linked to at least ten different artists, yielding 277 tags, and then screened out tags that did not obviously correspond to a genre (e.g., "American," "London," "Male Vocalist," "Ninja Tune"), leaving a total of 73 tags. More than 180,000 tracks by more than 4,400 different artists had one of these 73 tags. Many of these tags were synonymous (e.g., "hip hop," "hip-hop," and "hiphop") and others could plausibly be considered of a similar genre (e.g., "heavy metal," "thrash metal," "black metal"). To group these tags into genre classes, we computed a co-occurrence matrix of how often an artist was shared between two tags, normalized this 73x73 matrix by the base rates of each tag (so each element was the proportion of overlap), and hierarchically clustered this normalized co-occurrence matrix, grouping together tags that overlapped substantially. This procedure yielded 41 genre clusters (see Table S2 for list of genres and the tags associated with each genre).

To generate training exemplars for the genre task, we randomly excerpted approximately two and a half million two-second clips from these 180,000 tracks. As with the word task, to make the classification task more realistic and difficult, we embedded these two-second excerpts in one of four different background noises: (1) auditory scenes, (2) two-speaker speech babble, (3) eight-speaker speech babble, or (4) music-shaped noise. Auditory scenes and two-speaker babble were generated in the same way as they were for the word recognition task. Eight-speaker babble was generated analogously to the two-speaker babble, but included eight distinct speakers. Music-shaped noise consisted of a two-second clip of noise that was matched to the average spectrum of its corresponding two-second clip of music. SNRs for background noise in the genre task were also selected based on pilot behavioral results in humans to yield roughly equal performance across noise types. The genre training clips were presented at substantially higher SNR relative to the word recognition task due to the difficulty of genre recognition for human listeners. The mean SNR for each of the four background types was 12 dB, and the exact SNR for each training example was drawn randomly from a Gaussian with a standard deviation of 2 dB. For both the genre and the word task, all waveforms were downsampled to 16 kHz.

Stimulus preprocessing for networks: Cochleagrams

The input to the network was a cochleagram of each training exemplar. A cochleagram is a time-frequency decomposition of a sound that mimics aspects of cochlear processing – it is similar to a spectrogram, but with a frequency resolution like that thought to be present in the cochlea, and with a compressive nonlinearity applied to the amplitude in each time-frequency bin.

Cochleagrams were generated similarly to those in previous work (McDermott et al., 2013; McDermott and Simoncelli, 2011). Each two-second waveform was passed through a bank of 203 bandpass filters. Filters were zero-phase with frequency response equal to the positive portion of a single period of a cosine function. The center frequencies ranged from 30 Hz to 7860 Hz. The filters were evenly spaced on an Equivalent Rectangular Bandwidth $(ERB)_N$ scale, approximately replicating the frequency-dependence of bandwidths believed to characterize the human cochlea (Glasberg and Moore, 1990). Filters were designed to perfectly tile the spectrum – the summed squared response across all frequencies was flat – and to achieve this tiling, four low-pass and four high-pass filters were included. Adjacent filters overlapped in frequency by 87.5%. With the lowpass and highpass filters on the ends of the spectrum, there were in total 211 filters.

The envelope of each filter subband was computed as the magnitude of the analytic signal (via the Hilbert transform). To simulate basilar membrane compression, these envelopes were raised to the power of 0.3. Compressed envelopes were downsampled to 200 Hz, yielding a cochleagram representation that was 211 by 400 (frequency x time). Finally, these cochleagrams were resampled (with an antialiasing filter) to 256 × 256 to input to the networks. Cochleagram generation was done in Python, making heavy use of the numpy and scipy libraries (Jones et al., 2001; Oliphant, 2006).

Definition of constituent operations of a convolutional neural network (CNN)

The 256 × 256 cochleagrams were passed into convolutional neural networks (CNNs), which are a feedforward cascade of linear and nonlinear operations – the output of each layer is passed as the input to the subsequent layer. In our CNNs there were four different kinds of layers: (1) a *convolutional layer*, (2) a *normalization layer*, (3) a *pooling layer*, and (4) a *fully connected layer*. Below we define the operations of each type of layer.

Convolutional layer

Each *convolutional layer* consisted of a bank of linear filters and a pointwise nonlinearity. The input to a *convolutional layer* is a three-dimensional array of shape (n_{in}, n_{in}, $d_{channels\_in}$). Where n_{in} is the spectral and temporal dimensionality of the input to that layer. In the case of the first layer, n_{in} is the number of time or frequency bins of the preprocessed cochleagram (i.e., 256). $d_{channels\_in}$ is the number of channels from the previous layer. In the case of the first convolutional layer, $d_{channels\_in}$ = 1, because the cochleagram representation had a single value for each time-frequency bin.

The *convolutional layer* is defined by the following five parameters:

1. k_s: The size of the convolutional kernels
2. n_k: The number of different kernels
3. l_s: The length of the stride of the convolution
4. K: The kernel weights for each of the n_k kernels; this is an array of dimensions (k_s, k_s, $d_{channels\_in}$, n_k).
5. b: The bias vector, of length n_k

For any input array X of shape (n_{in}, n_{in}, $d_{channels\_in}$), the output of a *convolutional layer* is an array Y of shape (n_{in}/l_s, n_{in}/l_s, n_k):

$$Y(i,j,k) = \text{relu}\left(b[k] + \frac{1}{k_s^2}\sum K[:,:,:,k] \odot N_{ks}(X, l_s \cdot i, l_s \cdot j)\right),$$

where i and j range in (1, …, n_{in}/l_s), $\odot$ denotes the pointwise array multiplication, N_{ks}(X,q,r) selects the square neighborhood across adjacent time and frequency bins of size k_s by k_s, centered at location q, r in X, returning an array of shape (k_s, k_s, $d_{channels\_in}$). The sum is over all elements of this 3d array. The convolution is done with "same" mode, meaning that the edges are padded with zeros to produce an output that would be the same dimensionality of the input if the stride were set to 1. Finally, *relu* denotes the rectified linear operator, which is a pointwise nonlinearity applied to every element of the output:

$$\text{relu}(x) = \max(0, x).$$

Normalization layer

A *normalization layer* implements divisive normalization of a unit by its neighboring filters at the identical time-frequency bin. It is defined by three parameters: α, β, and $n_{adjacent}$, which had values of 0.001, 0.75, 5, respectively. It operates on an array of X of shape (n_{in}, n_{in}, $d_{channels\_in}$) and returns an array Y of the identical shape:

$$Y(i,j,k) = \frac{X[i,j,k]}{\left(1 + \alpha\sum_{l\in A} X[i,j,l]^2\right)^{\beta}},$$

where A is the set of the neighboring input channels whose responses are included in the normalization procedure and k indexes filter kernels. A is defined as [k – ($n_{adjacent}$ - 1)/2, k – ($n_{adjacent}$ - 1)/2 + 1, …, k, …, k + ($n_{adjacent}$ - 1)/2 – 1, k + ($n_{adjacent}$ - 1)/2], which in our case of $n_{adjacent}$ = 5 results in A = [k-2,k-1,k,k+1,k+2]. This procedure thus normalizes a filter's in a time-frequency bin by other filter responses for different filters at that same time-frequency bin. We handled boundaries by zero-padding.

Pooling layer

A *pooling layer* downsamples its input by aggregating values across nearby time and frequency bins:

1. p_s, the size of the pooling kernel
2. p_o, the pooling order (in our case: 1, 2, or infinity)
3. l_s, the length of the stride

A *pooling layer* operates on an array X of shape (n_{in}, n_{in}, $d_{channels\_in}$) and returns an array Y of shape of (n_{in}/l_s, n_{in}/l_s, $d_{channels\_in}$), via the following function:

$$Y(i,j,k) = \left(\frac{1}{p_s^2}\sum N_{ps}(X^{p_o}, l_s \cdot i, l_s \cdot j)[:,:,k]\right)^{\frac{1}{p_o}},$$

where N_{ps} is the local square neighborhood in time and frequency. The sum is over all elements in the resulting (p_s, p_s) matrix. p_o = 1 yields pooling via the mean, p_o = 2 yields pooling via the root mean square ("L2 pooling"), and $p_o = \infty$ yields max pooling.

Fully connected layer

Unlike the previously described layers, a *fully connected layer* does not have a notion of localized frequency or time. It takes an input vector X and returns a vector Y of length of (n_{out}):

$$Y(i) = \text{relu}\left(\frac{1}{n_T}\sum_{j=1}^{n_T} w_{ij}X[j]\right),$$

where j iterates the number of input elements in the preceding layer, w_{ij} is a weight (a real-valued number), and the point-wise nonlinearity (*relu*) is as defined before. If the preceding layer is another *fully connected layer* the input is a vector of length (n_T). If the preceding layer is a *convolutional layer*, a *normalization layer* or a *pooling layer*, then the input is an array A of shape (n_{in}, n_{in}, $d_{channels\_in}$) unwrapped into a single vector X of length $n_T = n_{in}^2 \times d_{channels\_in}$.

Dropout during training

During training, we instantiated a dropout layer after each fully connected layer that was not the final classification layer. During each batch of training data, a randomly drawn fifty percent of the fully connected layer connections were eliminated (i.e., their connection weight was set to zero). Dropout is common in neural network training and can be seen as a form of model averaging. For evaluation, we followed the conventional procedure by removing this dropout layer and multiplying the output of each fully connected unit by the probability of dropout during training (i.e., 0.5).

Softmax classifier

The final layer in the network is a *fully connected layer* where n_{out} is the number of target classes (587 or 41, respectively in the case of the word or genre task) and the *relu* operator is replaced with a *softmax*, an operation that receives a vector as its input and whose element-wise output is defined as:

$$y(i) = \frac{\exp\left(\sum_{j=1}^{n_T} w_{ij}x_j\right)}{\sum_{k=1}^{n_{out}} \exp\left(\sum_{l=1}^{n_T} w_{kl}x_l\right)}.$$

Each element of the output of the softmax is nonnegative and together they sum to one, and can thus be interpreted as a probability distribution over the classes (either words or genres).

Convolving in frequency

The convolutional neural networks we used in this paper convolved filter kernels with the cochleagram in both time and frequency. Applying a given kernel in the first layer yields an $n_{spectral}$ by $n_{temporal}$ feature map – i.e., outputs values at $n_{temporal}$ different time bins for each of $n_{spectral}$ different frequencies. Convolving in time is natural for a model of the auditory system because sound waveforms are naturally ordered in a temporal sequence. The naturalness of convolving in frequency may be less obvious. However, convolution in frequency can be conceptualized as having $n_{spectral}$ different model units, each centered at a different center frequency, where each of the $n_{spectral}$ units is constrained to have the same filter weights. Instantiating the same filter at different frequencies might be reasonable given that sounds are to some extent translation invariant in frequency, and is present in standard models of spectrotemporal filtering (Chi et al., 2005). Furthermore, we empirically found in pilot experiments that task performance was better when convolution was applied in frequency as well as time, likely because this substantially reduces the number of parameters to be learned and thus may have acted as an useful form of regularization.

CNN architectural optimization and filter weight training

Filter weight training

During training, the filter weights in each *convolutional layer* and each *fully connected layer* were adjusted to improve classification performance via stochastic gradient descent (SGD). Training was performed on 5.1 million sounds and performance was monitored on a left-out validation set of 400,000 sounds. We used the cross-entropy loss function and a batch size of 256 stimuli. Error on this loss function was backpropagated to update the weights in each *convolutional layer* and *fully connected layer* to decrease the loss function.

Network architecture selection: Overview

We selected the architecture for our network via a two-stage procedure. First, we defined a broad set of architectural hyperparameters and searched for architectures that performed well on either the word or genre task separately (Figure 1B). Then we searched across ways of merging the best single-task architectures into a single network to perform both tasks (Figure 1C). We divided the architecture search into these two stages for practical reasons – simultaneously searching over both base architecture and branch point would have been prohibitively computationally expensive.

Network architecture selection: Family of potential architectures

To optimize the model architecture, we defined a space of potential architectures by creating a distribution over architectural hyperparameters. This hyperparameter space was defined as follows:

- The first six stages of processing consisted of the following layers in the following order: *convolution*, *normalization*, *pooling*, *convolution*, *normalization*, *pooling.* Although this order was identical across all architectures we considered, we varied some of the architectural hyperparameters that define each of these stages, as we describe below.

- After these first six stages, there were a variable number of additional *convolutional layers* (between 0 and 3)
- Following the above, there were one or two stages of *fully connected layers.* During training, a *dropout layer* was added after each of these layers.
- And the final stage was a *fully connected layer* with either 587 or 41 units, whose outputs were passed through a softmax and interpreted as a probability distribution over classes (words or genres)

Additionally, the number of filters in each of these potential *convolutional layers* were fixed (in order of *convolutional layer* number): 96, 256, 512, 1024, and 512. The number of fully connected units (other than the output layer) was set to 4096. The pooling window size was set to three, and normalization was performed over neighborhoods of five filters. To simplify our search over architectures, we also fixed the convolutional filters to be ''square,'' with equal spectral and temporal extent.

The choice of this template architecture was made based on pilot experiments and experience with networks trained on visual object recognition tasks. Within this template architecture, there were a number of degrees of freedom that we searched over:

- Total number of convolutional layers: [2, 3, 4, 5]
- Number of fully connected layers before the softmax layer: [1, 2]
- Convolutional filter kernel sizes: [3, 5, 7, 9, 11, 13]
- Stride of each convolutional filter: [1, 2, 3, 4]
- The pooling order: [1, 2, ∞] (i.e., average, root mean square, or max)

Network architecture selection, first stage: Single-task networks

To search over these architectural hyperparameters parameters, we set a uniform probability distribution over each of these choices, which acted as our prior over architectures. Then we drew 100 sample architectures from this prior and randomly assigned each architecture to be trained on either the word or genre task. We optimized the *convolutional layer* and *fully connected layer* filter weights with SGD for 28 epochs (complete passes over the training data). We then evaluated the performance of each architecture on the task on which it was trained using left-out validation data. We used tree-structured Parzen estimation to update our probability distribution over each hyperparameter, assigning more probability mass to those hyperparameter values that produced better task performance. Specifically, to update the probability distribution for a given architectural hyperparameter (e.g., convolutional filter size), we first separated the 100 original architectures into two groups: the 50 that yielded above median performance and the 50 that yielded below median performance. We then fit one truncated Gaussian ($q(x)$) to the values of that given architectural hyperparameter in the above median group, and separately fit another truncated Gaussian ($r(x)$) to those hyperparameters for the below median performance group. To bias selection of hyperparameter values toward those that yielded higher performance, we updated the distribution over each architecture hyperparameter as the ratio of those two truncated Gaussians (i.e., $q(x) / r(x)$).

We then drew an additional 40 architectures from this new distribution over hyperparameters, initialized the weights for two networks for each of these 40 architectures (one for the word and one for the genre task), and optimized the filter weights in each one of these 80 new networks for their respective task. We trained all 180 networks for a total of 42 epochs (the original 100 and these latter 80). We then evaluated the validation set performance for each of these 180 architectures on the task that it was trained. We found that the same architecture performed best on both the word and the genre task.

It is unclear the extent to which an explicit hyperparameter optimization procedure, as used here, is more useful than random search over architectures. Moreover, there are obvious limitations to the optimization procedure that we used. For instance, our optimization procedure updates the probability distribution over each hyperparameter separately, and thus assumes that performance as a function of architectural hyperparameter is separable across hyperparameters, which is not obviously the case. Note, however, that the architecture that performed best on each task was drawn from the hyperparameter distribution after the update to the prior, potentially suggesting some benefit from the optimization procedure.

Network architecture selection, second stage: Merging optimal architectures by selecting the branch point

We sought a single architecture to perform both word recognition and genre classification, which could share some number of early layers before splitting into two branches, one for word recognition and one for genre classification. Finding this shared network architecture was simplified by the fact that the same architecture produced the best performance for each task separately, such that the architectural decision was reduced to selecting the location of the ''branch point'' (Figure 1C).

Although there were twelve layers of processing for the network, only seven of them were convolutional or fully connected, and thus had weights that would be optimized during SGD. The other five layers were normalization or pooling layers, and the operations of those layers were not altered by task training. Therefore, normalization and pooling layer operations remained the same regardless of whether they were shared between tasks or split into separate branches, and thus it was nonsensical to consider branch points at these layers. We considered branch points ranging from before the first layer (i.e., yielding two fully separate networks), to just prior to the softmax classification layers (i.e., a branch point after layer fc6). This network with a branch point after fc6 had nearly half as many parameters as the entirely separate networks, because of the shared processing. To minimize the total number of network parameters while maximizing performance, we sought the latest branch point for which performance was not significantly lower than performance for the fully separate networks, with significance levels determined via a bootstrap over both classes (i.e., words or genres) and stimuli.

We generated networks with all seven possible branch points (from totally separate networks to branching after layer fc6), and optimized the filter weights from scratch in each of these seven networks for both the word and genre tasks. We found that performance on the validation set was not significantly affected by sharing up to three convolutional layers of processing (Figure 1D).

Network architecture selection: Final CNN architecture

The final architecture consisted of 12 layers of processing that split into separate streams after the conv3 layer, culminating in word classification or genre classification softmax layers (Figure 1E). The layers in the two streams are denoted by _W or _G. The architectural hyperparameters of parallel layers in the two streams are identical except for the final classification layers (which differ in dimensionality due to the different number of classes for each task).

- Input (256x256): Cochleagram: 256 frequency bins x 256 time bins
- conv1 (85x85x96): Convolution of 96 kernels with a kernel size of 9 and a stride of 3
- rnorm1 (85x85x96): Response normalization over 5 adjacent kernels
- pool1 (42x42x96): Max pooling over window size of 3x3 and a stride of 2
- conv2 (22x22x256): Convolution of 256 kernels with a kernel size of 5 and a stride of 2
- rnorm2 (22x22x256): Response normalization over 5 adjacent kernels
- pool2 (11x11x256): Max pooling over a window size of 3x3 and a stride of 2
- conv3 (13x13x512): Convolution of 512 kernels with a kernel size of 3 and a stride of 1
- conv4_W & conv4_G (15x15x1024): Convolution of 1024 kernels with a kernel size of 3 and a stride of 1
- conv5_W & conv5_G (17x17x512): Convolution of 512 kernels with a kernel size of 3 and a stride of 1
- pool3_W & pool3_G (8x8x512): Mean pooling over a window size of 3 and a stride of 2
- fc1_W & fc1_G (4096): A fully connected layer
- fctop_W & fctop_G (587 or 41): A fully connected layer whose outputs are passed through a softmax function and interpreted as a probability distribution over either words (n = 587) or genres (n = 41).

During training, there was a dropout layer after fc1_W and fc1_G.

Although most of our analyses are based on this final network, we also report voxel prediction results for the two totally separate networks (i.e., that shared no layers), to examine the effects of training on only one of the two tasks (Figure S3E).

Control model: A spectrotemporal modulation filter bank

To compare the neural network to an existing model of auditory cortex, we also performed analyses with a standard spectrotemporal filter model (Chi et al., 2005). The model consists of a bank of linear filters tuned to spectrotemporal modulations at different acoustic frequencies, spectral scales, and temporal rates (see Figure 1F for example filters). We used the NSL MATLAB Toolbox (http://www.isr.umd.edu/Labs/NSL/Software.htm) implementation of the spectrotemporal model, with 63 audio frequencies (ranging between 20 Hz and 8 kHz), 17 temporal rates (logarithmically spaced between 0.5 Hz and 128 Hz), and 13 spectral scales (logarithmically spaced between 0.125 and 8 cycles/octave), yielding a spectrotemporal modulation filter bank with 29,736 filters – 1071 temporal modulation filters (17 rates by 63 frequencies), 819 spectral modulation filters (13 rates by 63 frequencies), and 27,846 spectrotemporal filters (17 temporal rates by 13 spectral rates by 63 frequencies, by two orientations corresponding to upward and downward frequency modulations). The filters were applied to a cochleagram. The cochleagrams that were used as input to the CNN were resampled in frequency to match the log spacing of the spectrotemporal model. To evaluate the model response we passed a sound through the model, extracted the magnitude (absolute value) of each filter's response at each time step and averaged these magnitudes over time, to get 29,736 measures for each sound. To ensure that the comparison with the spectrotemporal filter model was fair, we verified that this parameterization of the spectrotemporal filter model saturated the amount of variance that a spectrotemporal filter model could explain in the voxel data (see Figure S2C, and methods below for more details).

Psychophysics comparing human listeners and the network

As an initial test of the plausibility of the trained neural network as a model of human auditory cortex, we compared the performance characteristics of the model with that of human listeners on the two tasks we used to train the network.

Human psychophysics: Word recognition

We measured word recognition performance in twenty-six different conditions – five different background types at five different SNRs along with a clean condition without any added background noise. The five background types were: (1) music, (2) two-speaker speech babble, (3) eight-speaker speech babble, (4) speaker-shaped noise, and (5) auditory scenes. The five SNRs were: −9, −6, −3, 0, and +3 dB SNR. Speaker-shaped noise (included to maximize ''energetic'' masking of the target speech by the background) was generated for each clip by estimating the average amplitude spectrum for the clip's speaker from that speaker's exemplars in the corpus, and synthesizing a noise sample with the same spectrum (by replacing the Fourier amplitudes of a white noise signal with the speaker's average amplitude spectrum). All other background types were generated with the same procedure used for the network training stimuli.

Each subject completed a two-hour in-lab experimental session. Sounds were played via the sound card on a MacMini at a sampling rate of 16 kHz, via a Behringer HA400 amplifier. The Psychtoolbox for MATLAB (Brainard, 1997) was used to present the stimuli. Subjects heard the sounds via Sennheiser HD280 headphones (circumaural) in a soundproof booth (Industrial Acoustics). All stimuli were presented at 70 dB SPL.

Each trial consisted of a two-second clip generated according to the network training data generation procedure described above. Participants were instructed to report the word that occurred during the middle of the clip (i.e., during the one-second mark). The network had a closed-set recognition task (587 possible words), and human subjects performed an analogous 587 alternative forced choice (AFC) task. To facilitate performance with such a large number of classes, we familiarized participants with the 587 words before their behavioral session by allowing them to look over the list, and we programmed an interface that only allowed responses from the 587-word dictionary. Participants typed responses but could enter in a part of a word, hit the spacebar key, and see all words in the dictionary that had the typed characters. Trials were grouped into runs of 78 trials between which subjects could take a break. Subjects performed up to seven runs. Across our 18 subjects, the average number of trials per subject was 451 (range: 312-546).

Human psychophysics: Genre classification

For the genre recognition task, we measured human genre classification performance in twenty-one different conditions – four background types at five different SNRs and a "clean" (background-less) condition. The four background types were: (1) auditory scenes, (2) clip-shaped noise, (3) two-speaker babble, and (4) eight-speaker babble. All backgrounds were generated as they were for the network training stimuli. We measured performance at the following SNRs: −9, −6, −3, 0, +3 dB.

Listeners heard a two-second clip randomly excerpted from a track from the Million Song Dataset embedded in background noise. Subjects had to report the genre they thought the clip belonged to. Because of the difficulty of the task (due to the brief stimulus and overlap between genres), subjects performed a "top 5" task in which they selected the five genres the clip was most likely to belong to, in order of confidence (the most likely genre first, then the second-most likely genre, and so on). All forty-one genres were presented in a numbered list on the screen during the response period and participants entered the five numbers corresponding to their five top choices. To familiarize participants with the genres, we played the in-lab subjects three examples per genre just prior to the experimental session. For the Turk subjects, we sought to insure their attention by having 41 familiarization trials (one for each genre) before the experimental session. Turk subjects heard three examples for each genre in succession and performed a two-way AFC on the genre, deciding between the true genre and an alternative (selected by hand to be substantially different; e.g., for "hip hop" the distractor was "opera"). Feedback was provided during this familiarization phase only, not during experimental trials.

In lab, stimuli were presented at 70 dB SPL. It was impractical to control the level for our Turk subjects. To help ensure that Turk subjects were wearing headphones, we required that the Turk subjects pass a headphone check task previously demonstrated to screen out listeners not wearing headphones or earphones (Woods et al., 2017). Turk subjects listened to three pure tones in a row and had to report which tone was quietest. Two of the tones were in phase across the two stereo channels and a third was in antiphase. One of the in-phase tones was 5 dB less intense than the other two. However, if subjects were listening over speakers rather than headphones the antiphase tone would be expected to be quietest due to in-air phase cancellation. All Turk subjects performed six of these loudness discrimination trials, and subjects who failed to get five of six trials correct were excluded from participating in the Turk experiment. We ran 300 subjects on Turk, but 146 (49%) failed the headphone check. Of the remaining 154 subjects, we considered data only from those who completed a minimum of 84 trials (i.e., four trials per background noise type and SNR pair), leaving a total of 80 Turk subjects.

Figure 2A plots the mean performance of all subjects for each condition in the word recognition task. Figure 2G plots mean performance of all subjects for each condition in the genre recognition task (a trial was counted as correct if the correct genre was include in any of the five guesses from the subject).

Neural network psychophysics: Word and genre classification

To evaluate model behavior on these two psychophysical tasks, we presented the same set of stimuli to the model that we presented to human listeners, extracting the model unit responses from the relevant top layer (FCTop_W for the word stimuli and FCTop_G for the genre stimuli). For each word stimulus, we took the argmax across the 587 model unit responses and recorded whether the unit with the highest response corresponded to the target word or not (Figure 2B; scatter comparing human and network word performance in Figure 2C). For genre stimuli, we noted which units had the five highest responses (5 highest probabilities of a genre conditioned on the input). If the correct genre was included in the five largest softmax values, then that trial was counted as correct when computing performance for Figures 2H and 2I.

Word psychophysics control analysis: Cochleagram distortion analyses

To better understand the similarity in the pattern of word recognition performance between the humans and the network across conditions, we explored whether performance could be explained by the extent to which different background noises distort the speech signal. We did not perform distortion analyses for the genre task because performance did not vary substantially across background types, and was thus determined primarily by SNR as measured in the waveform domain. We measured distortion for each stimulus by generating cochleagrams of each target speech signal with and without the assigned background noise, and then computed the absolute difference between corresponding time-frequency bins in the two cochleagrams (Figure 2D). We took the mean of this absolute difference over both time and frequency, aggregating this distortion metric by taking its mean across all stimuli for each background condition. Using the root mean square difference instead of the mean absolute difference yielded nearly identical results

(Pearson's r between MAE and RMS is 0.98, $p < 10^{-17}$). Figure 2E shows this measure of mean absolute distortion and Figure 2F shows the scatter of this distortion measure versus human behavior for each condition.

A limitation of the above distortion analysis is that it measures distortion in all time-frequency bins, but it is possible that distortion that overlaps temporally and spectrally with the speech signal may be particularly detrimental to word recognition abilities. To control for this possibility, we conducted a modified distortion analysis, where for each stimulus we only measured distortion in cochleagram bins that had speech signal within a certain number of decibels (dB) from peak signal power (Figure S1A). We varied our selection criterion from 50 dB down to 10 dB down in increments of 10 dB.

Word psychophysics control analysis: Network response distortion analyses

Given that cochleagram distortion did not predict human performance particularly well, we examined whether there would be a closer relationship with human performance for an analogous measure of distortion computed in different layers of the trained network model. We presented target speech signals with and without background noise to the network and recorded the model unit responses in each of the layers. We then computed the mean absolute difference between these unit responses for each stimulus and each layer, aggregating over background type. Figure S1B shows scatters of human performance versus distortion for each layer of the network.

Genre psychophysics comparisons: Confusion matrices and control models

Given that performance across background types did not vary substantially for the genre task, we instead compared error patterns for both the network and the humans. This comparison was feasible for the genre task because there were only 41 classes, yielding a total of just under 1700 bins (41 × 41). We did not perform this analysis for the word task because it would have required orders of magnitude more behavioral data – the full confusion matrix would have more than 340,000 bins (587 × 587).

We generated two 41-by-41 confusion matrices, one for the human behavioral data and one for the network (Figures 2J and 2K, respectively). Each column of the matrix contained the responses for a genre. The elements of each column contained the proportion of trials on which a genre tag was in the top five guesses for the genre. Each column was z-scored to control for response biases (which might be expected to be different for the network and human listeners). To compare confusion matrices, we unwrapped each into a vector and measured the Spearman correlation between the resulting vectors.

fMRI data analysis: Preprocessing and voxel selection

Natural sound stimuli

The stimuli were a set of 165 two-second sounds selected to span the sorts of sounds that listeners most frequently encounter in day-to-day life (Norman-Haignere et al., 2015). All sounds were recognizable – i.e., classified correctly at least 80% of the time in a ten-way alternative forced choice task run on Amazon Mechanical Turk, with 55-60 participants per sound. Turk participants also assigned each of these 165 sounds to one of eleven categories (instrumental music, music with vocals, English speech, foreign speech, non-speech vocal sound, animal vocalization, human non-vocal sound, animal non-vocal sound, nature sound, mechanical sound, or environmental sound; 30-33 participants per sound). Category assignments were highly reliable (split-half kappa of 0.93). In analyses that we describe later in this methods section where we excluded speech and music stimuli from the fMRI dataset, we relied upon these judgments of third-party listeners to determine whether a stimulus contained speech or music (i.e., we excluded the stimuli the Turk subjects classified as instrumental music, music with vocals, English speech, or foreign speech). See Table S3 for names of all stimuli and category assignments. To download all 165 sounds, see the McDermott lab website: http://mcdermottlab.mit.edu/downloads.html.

Sounds were presented using a block design. Each block included five presentations of the identical two-second sound clip. After each two-second sound, a single fMRI volume was collected ("sparse scanning"), such that sounds were not presented simultaneously with the scanner noise. Each acquisition lasted one second and stimuli were presented during a 2.4 s interval (200 ms of silence before and after each sound to minimize forward/backward masking by scanner noise). Each block lasted 17 s (five repetitions of a 3.4 s TR). This design was selected based on pilot results showing that it gave more reliable responses than an event-related design given the same amount of overall scan time. Blocks were grouped into eleven runs, each with fifteen stimulus blocks and four blocks of silence. Silence blocks were the same duration as the stimulus blocks and were spaced randomly throughout the run. Silence blocks were included to enable estimation of the baseline response.

To encourage subjects to attend equally to each sound, subjects performed a sound intensity discrimination task. In each block, one of the five sounds was 7 dB lower than the other four (the quieter sound was never the first sound). Subjects were instructed to press a button when they heard the quieter sound. Sounds were presented through MR-compatible earphones (Sensimetrics S14) at 75 dB SPL (68 dB SPL for the quieter sounds).

fMRI data acquisition

MR data were collected on a 3T Siemens Trio scanner with a 32-channel head coil at the Athinoula A. Martinos Imaging Center of the McGovern Institute for Brain Research at MIT. Each functional volume consisted of fifteen slices oriented parallel to the superior temporal plane, covering the portion of the temporal lobe superior to and including the superior temporal sulcus. Repetition time (TR) was 3.4 s (although acquisition time was only 1 s), echo time (TE) was 30 ms, and flip angle was 90 degrees. For each run, the five initial volumes were discarded to allow homogenization of the magnetic field. In-plane resolution was 2.1 × 2.1 mm (96 × 96 matrix), and slice thickness was 4 mm with a 10% gap, yielding a voxel size of 2.1 × 2.1 × 4.4 mm. iPAT was used to minimize acquisition time. T1-weighted anatomical images were collected in each subject (1mm isotropic voxels) for alignment and surface reconstruction.

fMRI data preprocessing

Functional volumes were preprocessed using FSL and in-house MATLAB scripts. Volumes were corrected for motion and slice time. Volumes were skull-stripped, and voxel time courses were linearly detrended. Each run was aligned to the anatomical volume using FLIRT and BBRegister. These preprocessed functional volumes were then resampled to vertices on the reconstructed cortical surface computed via FreeSurfer, and were smoothed on the surface with a 3mm FWHM 2D Gaussian kernel to improve SNR. All analyses were done in this surface space, but for ease of discussion we refer to vertices as ''voxels'' throughout this paper. For each of the three scan sessions, we estimated the mean response of each voxel (in the surface space) to each stimulus block by averaging the response of the second through the fifth acquisitions after the onset of each block (the first acquisition was excluded to account for the hemodynamic lag). Pilot analyses showed similar response estimates from a more traditional GLM. These signal-averaged responses were converted to percent signal change (PSC) by subtracting and dividing by each voxel's response to the blocks of silence. These PSC values were then downsampled from the surface space to a 2mm isotropic grid on the FreeSurfer-flattened cortical sheet. For summary maps, we registered each subject's surface to Freesurfer's fsaverage template.

Voxel selection

For individual subject analyses, we used the same voxel selection criterion as Norman-Haignere et al. (2015), selecting voxels with a consistent response to sounds from a large anatomical constraint region encompassing the superior temporal and posterior parietal cortex. Specifically, we used two criteria: (1) a significant response to sounds compared with silence ($p < 0.001$); and (2) a reliable response to the pattern of 165 sounds across scans. The reliability measure was as follows:

$$r = 1 - \frac{\|\mathbf{v}_{12} - \mathrm{proj}_{\mathbf{v}3}\mathbf{v}_{12}\|_2}{\|\mathbf{v}_{12}\|_2},$$

$$\mathrm{proj}_{\mathbf{v}_3}\mathbf{v}_{12} = \left(\frac{\mathbf{v}_3^T}{\|\mathbf{v}_3\|_2}\mathbf{v}_{12}\right)\mathbf{v}_3$$

where $\mathbf{v}_{12}$ is the response of a single voxel to the 165 sounds averaged across the first two scans (a vector), and $\mathbf{v}_3$ is that same voxel's response measured in the third. The numerator in the second term in the first equation is the magnitude of the residual left in $\mathbf{v}_{12}$ after projecting out the response shared with $\mathbf{v}_3$. This ''residual magnitude'' is divided by its maximum possible value (the magnitude of $\mathbf{v}_{12}$). The measure is bounded between 0 and 1, but differs from a correlation in assigning high values to voxels with a consistent response to the sound set, even if the response does not vary substantially across sounds. We found that using a more traditional correlation-based reliability measure excluded many voxels in primary auditory cortex because some of them exhibit only modest response variation across natural sounds. We included voxels with a value of this modified reliability measure of 0.3 or higher, which when combined with the sound responsive t test yielded a total of 7694 voxels across the eight subjects (mean number of voxels per subject: 961.75; range: 637-1221).

For summary maps, which aggregated across individuals, voxels were included if at least four subjects had a reliability of at least 0.3.

Region of interest (ROI) selection: Overview

We localized four regions of interest (ROIs) in each participant, consisting of voxels selective for (1) frequency (i.e., tonotopy), (2) pitch, (3) speech, and (4) music. In each case we ran a ''localizer'' statistical test and selected the top 5% most significant individual voxels in each subject and hemisphere (including all voxels identified by the sound-responsive and reliability criteria described above). We excluded voxels that were identified in this way by more than one localizer (see Figure S3A for amount of overlap between ROIs before this exclusion criterion was applied). Key results were robust to varying the ROI selection criterion to 2% or 10% (Figures S3B and S3C). The frequency, pitch, and speech localizers required acquiring additional imaging data, and were collected either during extra time during the natural sound stimuli scan sessions or on additional sessions on different days. Scanning acquisition parameters were identical to those used to acquire the natural sounds data. Throughout this paper we refer to voxels chosen by these criteria as ''selective,'' for ease and consistency. Heatmaps displaying ROI voxel counts across subject are at the bottom of Figure 3B. Red indicates at least one subject had a voxel and yellow indicates two or more, except for the ''all'' ROI in which red indicates one subject and yellow indicates four or more.

ROI selection: Frequency-selective voxels

To identify frequency-selective voxels, we measured responses to pure tones in six different frequency ranges (center frequencies: 200, 400, 800, 1600, 3200, 6400 Hz) (Humphries et al., 2010; Norman-Haignere et al., 2013). For each voxel, we ran a one-way ANOVA on its response to each of these six frequency ranges and selected voxels that were significantly modulated by pure tones (top 5% of all selected voxels in each subject). Although there was no spatial contiguity constraint built into our selection method, in practice most selected voxels were contiguous and centered around Heschl's gyrus (Figure 3B).

ROI selection: Pitch-selective voxels

To identify pitch-selective voxels, we measured responses to harmonic tones and spectrally-matched noise (Norman-Haignere et al., 2013). For each voxel we ran a one-tailed t test evaluating the response to tones was greater than that to noise. We selected the top 5% of individual voxels in each subject that had the lowest p values for this contrast.

ROI selection: Speech-selective voxels

To identify speech-selective voxels, we measured responses to German speech and to temporally scrambled ("quilted") speech stimuli generated from the same German source recordings (Overath et al., 2015). We used foreign speech to identify responses to speech acoustical structure, independent of linguistic structure. Note that two of the subjects had studied German in school and for one of these subjects we used Russian utterances instead of German. The other subject was tested with German because the Russian stimuli were not available at the time of the scan. For each voxel we ran a one-tailed t test evaluating whether responses were higher to intact speech than to statistically matched quilts. We selected the top 5% of all selected voxels in each subject.

ROI selection: Music-selective voxels

To identify music-selective voxels, we used the music component derived by Norman-Haignere et al. (2015). We inferred the "voxel weights" for each voxel to all six of the components from its response to the 165 sounds:

$$\mathbf{w} = \mathbf{C}'\mathbf{v},$$

where **w** contains the inferred voxel weights (a vector of length 6), **C'** is the Moore-Penrose pseudoinverse of the "response components" (a 6 by 165 matrix), and **v** is the measured response of a given voxel (vector of length 165). We assessed the significance of each voxel's music component weight via a permutation test. During each iteration, we shuffled all the component elements, re-computed this new matrix's pseudoinverse, and re-computed each voxel's weights via the matrix multiply above. We performed this procedure 10,000 times, and fit a Gaussian to each voxel's null distribution of music weights. We then calculated the likelihood of the empirically observed voxel weight from this null distribution, and took the top 5% of voxels with the lowest likelihood under this null distribution.

Using neural network layers as voxelwise encoding models to predict cortical responses

Feature extraction from CNN

We first generated cochleagrams for each of the 165 sounds from the fMRI experiment and passed them through the CNN, recording each model unit's response in each layer to each sound (schematic in Figure 3A). We performed this procedure for all sounds and layers.

Because the neural network's input (a cochleagram) had a temporal dimension, the extracted responses from all layers (except for the fully connected layers) had a temporal dimension. This temporal component was critical to the computation of the features throughout the layers of the deep network, as temporal information at each layer is integrated by model filters at the succeeding layer to compute a signal that depends on both time and frequency. However, the hemodynamic signal to which we were comparing the model blurs the temporal variation of the cortical response, thus a fair comparison of the model to the fMRI data involved predicting each voxel's time-averaged response to each sound from time-averaged model responses. We therefore averaged the model responses over the temporal dimension after extraction. As a result of the *relu* and *softmax* operators, all responses in all layers were nonnegative, such that averaging preserved the mean response amplitude. For each layer that had a temporal dimension (i.e., each *convolutional layer*, *normalization layer*, and *pooling layer*), we first extracted a three-dimensional array of responses of shape ($n_{spectral}$, $n_{temporal}$, $n_{kernels}$). We then reshaped this 3d array into a matrix where each row corresponded to a kernel at a given frequency and each column corresponded to a time point. We finally took the average of each row over columns, yielding for each sound a vector whose length was the product of $n_{spectral}$ and $n_{kernels}$.

As a result of this procedure, each layer produced the following number of regressors for each sound: conv1 (8160), rnorm1 (8160), pool1 (4032), conv2 (5632), rnorm2 (5632), pool2 (2816), conv3 (6656), conv4_W (15360), conv4_G (15360), conv5_W (8704), conv5_G (8704), pool5_W (4096), and pool5_G (4096). Each *fully connected* layer did not have a temporal dimension, and so we simply extracted each model unit's response yielding the following number of features: fc1_W (4096), fc1_G (4096), fctop_W (587), and fctop_G (41). We used each of these 17 feature sets to predict the fMRI responses to the natural sound stimuli.

Voxelwise modeling: Regularized linear regression and cross validation

We modeled each voxel's time-averaged response as a linear combination of a layer's time-averaged unit responses. We first generated 10 randomly selected train/test splits of the 165 sound stimuli into 83 training sounds and 82 testing sounds. For each split, we estimated a linear map from model units to voxels on the 83 training stimuli and evaluated the quality of the prediction using the remaining 82 testing sounds (described below in greater detail). For each voxel-layer pair, we took the median across the 10 splits.

The linear map was estimated using regularized linear regression. Given that the number of regressors (i.e., time-averaged model units) typically exceeded the number of sounds used for estimation (83), regularization was critical. We used L2-regularized ("ridge") regression, which can be seen as placing a zero-mean Gaussian prior on the regression coefficients. Introducing the L2-penalty on the weights results in a closed-form solution to the regression problem, which is similar to the ordinary least-squares regression normal equation:

$$\mathbf{w} = \left(\mathbf{X}^T\mathbf{X} + n\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y},$$

where **w** is a d-length column vector (the number of regressors – i.e., the number of time-averaged units for the given layer), **y** is an n-length column vector containing the voxel's mean response to each sound (length 83), **X** is a matrix of regressors (n stimuli by d regressors), n is the number of stimuli used for estimation (83), and **I** is the identity matrix (d by d). We demeaned each column of the

regressor matrix (i.e., each model unit's response to each sound), but we did not normalize the columns to have unit norm. By not constraining the norm of each column to be one, we implemented ridge regression with a non-isotropic prior on each unit's learned coefficient. Under such a prior, units with larger norm were expected *a priori* to contribute more to the voxel predictions. In pilot experiments, we found that this procedure led to more accurate and stable predictions in left-out data, compared with a procedure where the columns of the regressor matrices were normalized (i.e., with an isotropic prior).

Performing ridge regression requires selecting a regularization parameter (denoted above by λ) that trades off between the fit to the (training) data and the penalty for weights with high coefficients. To select this regularization parameter, we used leave-one-out cross validation within the set of 83 training sounds. Specifically, for each of 50 logarithmically-spaced regularization parameter values (10^0, 10^{-1}, …, 10^{-49}, 10^{-49}), we measured the squared error in the resulting prediction of the left out sound using regression weights derived from the other sounds in the training split. We computed the average of this error (across the 83 training sounds) for each of the 50 potential regularization parameter values. We then selected the regularization parameter that minimized this mean squared error. Finally, with the regularization parameter selected, we used all 83 training sounds to estimate a single linear mapping from a layer's features to a given voxel's response. We then used this linear mapping to predict the response to the left-out 82 test sounds, and evaluated the Pearson correlation of the predicted voxel response with the observed voxel response. We squared this correlation coefficient to yield a measure of variance explained. We found that the selected regularization parameter values never fell on the boundaries of the search grid, suggesting that the range of the search grid was appropriate. We emphasize that the 82 test sounds on which predictions were ultimately evaluated were not incorporated into the procedure for selecting the regularization parameter nor for estimating the linear mapping from layer features to a voxel's response – i.e., the procedure was fully cross-validated.

Selecting regularization coefficients independently for each voxel-layer regression was computationally expensive, but seemed important for our scientific goals given that the optimal regularization parameter could vary across voxel-layer pairs. For instance, differences in the extent to which the singular value spectrum of the feature matrix is uniform or peaked (which influences the extent to which the $\mathbf{X^TX} + \mathbf{n\lambda I}$ matrix in the normal equation above is well-conditioned) can lead to differences in the optimal amount of regularization. Measurement noise, which varies across voxels (as seen in the variation in test-retest reliability across voxels in Figure S2A) can also influence the degree of optimal regularization. By allowing different feature sets (layers) to have different regularization parameters we are enabling each feature set to make the best possible predictions, which is appealing given that the prediction quality is the critical dependent variable that we compare across voxels and layers. Varying the regularization parameter across feature sets while predicting the same voxel response will alter the statistics of the regression coefficients across feature sets, and thus would complicate the analysis and interpretation of regression coefficients. However, we are not analyzing the regression coefficients in this work.

Voxelwise modeling: Correcting for reliability of the measured voxel response

The use of explained variance as a metric for model evaluation is inevitably limited by measurement noise. To correct for the effects of measurement noise we computed the reliability of both the measured voxel response and the predicted voxel response. Correcting for the reliability of the measured response is important to make comparisons across different voxels, because, as seen in Figure S2A, the reliability of the BOLD response varies across voxels. This variation can occur for a variety of reasons (e.g., distance from the head coil elements). Not correcting for the reliability of the measured response will downwardly bias the estimates of variance explained and will do so differentially across voxels. This differential downward bias could lead to incorrect inferences about how well a given set of model features explains the response of voxels in different parts of auditory cortex.

Our data were relatively reliable for an fMRI experiment: the median test-rest reliability across all included voxels (i.e., correlation coefficient between pairs of single scans) was 0.33; the median estimated test-retest reliability of the average response across all three scans (estimated via the Spearman-Brown correction) was 0.59 (top of Figure S2A). Moreover, our estimates of the reliability were themselves reliable (bottom of Figure S2A): the correlation of voxel reliability measured in different pairs of scans was r = 0.88. Individual voxel predictions were nearly all significant, as well, even under the relatively stringent Bonferroni correction for multiple comparisons (Figure S2B).

Voxelwise modeling: Correcting for reliability of the predicted voxel response

Measurement noise corrupts the test data to which model predictions are compared (which we accounted for by correcting for the reliability of the measured voxel response, as described above), but noise is also present in the training data and thus also inevitably corrupts the estimates of the regression weights mapping from model features to a given voxel. This second influence of measurement noise is often overlooked, but can be addressed by correcting for the reliability of the predicted response. Doing so is important for two reasons. First, as with the reliability of the measured voxel response, not correcting for the predicted voxel response can yield incorrect inferences about how well a model explains different voxels. Second, the reliability of the predicted response for a given voxel can vary across feature sets, and failing to account for these differences can lead to incorrect inferences about which set of features best explains that voxel's response.

Differences in the reliability of the predicted response across layers are due in part to differences in the singular value spectra of features from the different layers. A flatter, more uniform distribution of singular values will lead to more reliable predictions, whereas a more peaked distribution will lead to less reliable predictions. More peaked singular value spectra will inflate the contribution of noise to the regression weights due to the matrix inversion in the regularized least-squares normal equation above. Adding regularization "flattens" the singular value spectrum and thus reduces the contribution of noise.

It was thus in practice important to correct for the reliability of the predicted voxel response. By correcting for both the reliability of the measured voxel response and the reliability of the predicted response, the ceiling of our measured r-squared values was 1 for all voxels and all layers, enabling comparisons of voxel predictions across all voxels and all neural network layers.

Voxelwise modeling: Corrected measure of variance explained

To correct for the reliability, we employ the correction for attenuation (Spearman, 1904). It is a standard technique in many fields, and is becoming more common in neural data analysis. The correction estimates the correlation between two variables independent of measurement noise (here the measured voxel response and the model prediction of that response). The result is an unbiased estimator of the correlation coefficient that would be observed from noiseless data.

Our corrected measure of variance explained was the following:

$$r^{2*}_{\mathbf{v},\widehat{\mathbf{v}}} = \frac{\mathrm{r}(\mathbf{v}_{123}, \widehat{\mathbf{v}}_{123})^2}{r'_{\mathbf{v}} r'_{\widehat{\mathbf{v}}}},$$

where $\mathbf{v}_{123}$ is the voxel response to the 82 left-out sounds averaged over the three scans, $\widehat{\mathbf{v}}_{123}$ is the predicted response to the 82 left-out sounds (with regression weights learned from the other 83 sounds), r is a function that computes the correlation coefficient, $r'_{\mathbf{v}}$ is the estimated reliability of that voxel's response to the 83 sounds and $r'_{\widehat{\mathbf{v}}}$ is the estimated reliability of that predicted voxel's response. $r'_{\mathbf{v}}$ is the median of the correlation between all 3 pairs of scans (scan 0 with scan 1; scan 1 with scan 2; and scan 0 with scan 2), which is then Spearman-Brown corrected to account for the increased reliability that would be expected from tripling the amount of data (Spearman, 1910). Figure S2A shows the histograms of the pairwise correlation and the Spearman-Brown corrected correlation (i.e., the estimate of the reliability of the average data across three scans). $r'_{\widehat{\mathbf{v}}}$ is analogously computed by taking the median of the correlations for all pairs of predicted responses and Spearman-Brown correcting this measure. Note that for very noisy voxels, this division by the estimated reliability can be unstable and can cause for corrected variance explained measures that exceed one. To ameliorate this problem, we limited both the reliability of the prediction and the reliability of the voxel response to be greater than some value k (Huth et al., 2016). For $k = 1$, the denominator would be constrained to always equal one and thus the "corrected" variance explained measured would be identical to uncorrected value. For $k = 0$, the corrected estimated variance explained measure is unaffected by the value k. This k-correction can be seen through the lens of a bias-variance tradeoff: this correction reduces the amount of variance in the estimate of variance explained across different splits of stimuli, but does it at the expense of a downward bias of those variance explained metrics (by inflating the reliability measure for unreliable voxels). We used a k of 0.128, which is the $p < 0.05$ significance threshold for the correlation of two 165-dimensional Gaussian variables (i.e., with the same length as our 165-dimensional voxel response vectors).

Voxelwise modeling: Summary

We repeated this procedure for each layer and voxel ten times, once each for 10 random train/test splits, and took the median explained variance across the ten splits for a given layer-voxel pair. We performed this procedure for all 17 layers and all 7694 selected individual voxels. For comparison, we performed an identical procedure with the layers of a random-filter network with the same base architecture as our main network (Figures 3B, 4A, 4D, 4F, S2D, and S3B), single-task networks (Figure S2E), and for the time-averaged magnitudes for a bank of spectrotemporal filters (Figures 3B, 4A, 4B, S2C, and S3B–S3E). Additionally, we performed the regression for 78 randomly chosen random-filter networks with architectures that were not selected, and we summarized these predictions in ROIs by taking the median across all networks (Figures 3B and S3B).

When computing mean r^2 values throughout this paper, we averaged the values after Fisher transforming the correlation values. That is, we took the square root of each of the r^2 values we sought to average, performed the Fisher r-to-z transform, took the average across these values, performed the inverse of the Fisher z-to-r transform, and then squared the result. Averaging z-transformed values is appealing because the sampling distribution of correlation coefficients is skewed, and averaging in z-space reduces the bias of the estimate of the true mean.

All regression and analysis code was written in Python, making heavy use of the numpy and scipy libraries (Jones et al., 2001; Oliphant, 2006).

Varying the parameterization of the spectrotemporal filter bank

For voxelwise predictions, our baseline model was a spectrotemporal filter bank (Chi et al., 2005). To give as generous of comparison as possible, we sought to maximize the ability of the spectrotemporal model to explain cortical responses. We generated thirty-six different parameterizations of the spectrotemporal model; each with 63 different acoustic frequencies, and we varied across six different sets of spectral scales and six different temporal rates, all of which were logarithmically spaced.

Temporal rates: 3 rates (0.5, 4, 32 Hz), 5 rates (0.5, 2, 8, 32, 128 Hz), 9 rates (0.5, 1, 2, 4, 8, 16, 32, 64, 128 Hz), 13 rates (0.5, 0.79, 1.26, 2, 3.17, 5.04, 8, 12.70, 20.16, 32, 50.80, 80.63, 128 Hz), 17 rates (0.5, 0.71, 1, 1.41, 2, 2.83, 4, 5.66, 8, 11.31, 16, 22.63, 32, 45.25, 64, 90.51, 128 Hz), 23 rates (0.5, 0.66, 0.87, 1.15, 1.52, 2, 2.64, 3.48, 4.59, 6.06, 8, 10.56, 13.93, 18.38, 24.25, 32, 42.22, 55.72, 73.52, 97.01, 128 Hz).

Spectral scales: 2 scales (0.25, 2 cycles/octave), 4 scales (0.125, 0.5, 2, 8 cycles/octave), 7 scales (0.125, 0.25, 0.5, 1, 2, 4, 8 cycles/octave), 10 scales (0.125, 0.20, 0.32, 0.5, 0.79, 1.26, 2, 3.17, 5.04, 8 cycles/octave), 13 scales (0.125, 0.18, 0.25, 0.35, 0.5, 0.71, 1.0, 1.41, 2.0, 2.83, 4.0, 5.66, 8 cycles/octaves), 16 scales (0.125, 0.16, 0.22, 0.29, 0.38, 0.5, 0.66, 0.87, 1.15, 1.52, 2, 2.64, 3.48, 4.59, 6.06, 8 cycles/octave).

We examined the voxel prediction abilities of models with all 36 conjunctions (Figure S2C). We computed the variance explained for each of the resulting 36 different spectrotemporal models for each voxel from each subject, computed the median for each subject, and took the mean across subjects. These parameterizations varied in the number of parameters, from around 1,000 to over 40,000, and we examined the relationship between number of features in the spectrotemporal filter model and the cortical variance explained. We found that the variance explained by the model reached an asymptote by 30,000 features and thus for our control model we used the parameterization with 17 temporal rates, 13 spectral scales, and 63 audio frequencies.

Examining the effect of the random seed for the random-filter network

Throughout the paper we compare the trained network to a single untrained random-filter network. To examine the effect of the particular filter weights generated from a given randomization seed, we examined the predictions of each layer from ten other random-filter networks (i.e., with ten different seeds), as seen in Figure S2D.

ROI analyses

To estimate the voxel response variance explained by the network within functionally-defined regions of interests (ROIs), we selected the most predictive layer for the ROI in each individual subject using a leave-one-subject-out procedure. For a given subject and a given ROI, we computed the median variance explained by each network layer across the voxels in each of the other seven subject's ROI. We then took the average of this value across those subjects, yielding the mean variance explained by each layer for that ROI from the other subjects. We selected the layer that was most predictive in these seven subjects and measured the variance explained in the left-out subject. This cross validation across subjects avoided issues of non-independence. We iterated over subjects and ROIs and report the mean across subjects in Figures 3B and S3B.

To examine the variance explained by each layer across all of auditory cortex (Figure 4A), we took the median variance explained by each layer of the trained network for each subject across all selected voxels for that subject and then computed the mean of this measure across subjects. We performed the identical procedure for the random-filter networks and the spectrotemporal filter model. Similarly, to examine the median variance explained by each layer within the voxels in each ROI (Figures 4F, S3C, and S3D), we took the median variance explained by each layer over voxels in each subject's ROI and computed the mean. Black lines in Figure 4F shows analogous plots for the random-filter CNN. Additionally, to examine the effects of optimizing a network for either task alone, we performed an identical procedure with networks trained either for word or genre recognition alone (i.e., the baseline model in Figure 1D, with a branch point before any processing), and plot the results in Figure S3E.

Summary maps

For summary maps, we predicted responses in individuals and then aggregated results across subjects (with either the mean or median), after they were aligned in a common coordinate system (i.e., the fsaverage surface from FreeSurfer).

Summary maps: Variance explained by best-predicting layer

To explore how well the network predicted responses across all of the brain that we measured, we examined the variance explained by the best-predicting layer for each voxel, without employing an inclusion criterion. For the summary map of the variance explained by the best layer in the network (Figure S4A), we first computed the r^2 value for each session separately for each subject, correcting for the reliability of the voxel's response and of the layer's prediction of that response as estimated from the three pairs of individual scans. To measure the variance explained by the best-predicting layer of the network for each voxel, we compute the average variance explained across two scans, selected which layer had the highest r^2 for this pair of scans, and measured the r^2 for that layer in the left-out scan. We took the median of this left-out r^2 value across all three scans, and then took the mean over subjects. For comparison, we performed the same procedure on the "raw" variance explained measures (i.e., uncorrected for reliability, but Spearman-Brown corrected to account for using one-third the amount of data that we used for other maps); the result is shown in Figure S4B.

Summary maps: Difference between intermediate and higher layer

To compare the variance explained by different layers of the network, we selected an intermediate and higher layer and computed the difference in the r^2 value for each voxel for each individual, subtracting the value of the intermediate layer from that for the higher layer, and then averaging this value across subjects. Figure 4C shows this difference map for the conv5 and conv3 layers. Three example individual subjects (sub0, sub2, sub5) can be seen on the right of Figure 4C; Figure S5B shows all eight. Figure S5A shows the group-summary maps for other pairs of layers; the general form of the map is consistent across the particular layers used. For conv5 and other layers located after the branch point (that thus had separate layers for the genre and word streams), we took the mean of the r^2 values for the layers in the two streams. Figure 4D shows analogous maps for the same layers of the random-filter network. Figure 4E shows an analogous plot for the trained neural network but for the predictions of the voxel responses when all speech and music stimuli were excluded from the fMRI dataset.

Summary maps: Best-predicting layer

We also examined which layer best predicted each layer's response (an "argmax" analysis), which we summarized by taking the median across subjects for each voxel (Figure 4B). Individual subject maps are show in Figure S4C.

Examining representations in the network

To probe the representations of different layers of the network, we conducted the series of analyses shown in Figures 5, 6, S6, and S7. Although the networks used to generate Figures 1, 2, 3, 4, and 7 were trained using in-house software, these network representational analyses were conducted on a network retrained in TensorFlow (Abadi et al., 2016). TensorFlow, which became available only during the late stages of the project, facilitated these analyses and was anticipated to facilitate dissemination/distribution of

the network. This retrained network was similar to the in-house trained network, but had 1024 units in fc6 rather than 4096 (it was also trained from different random initial conditions). This TensorFlow network is the one that we analyze for Figures 5, 6, S6, and S7, and that will be posted on the McDermott lab website (http://mcdermottlab.mit.edu/).

Layerwise predictions of cochlear and spectrotemporal filters

We tested the ability of units in different network layers to predict responses to simulated cochlear filters and spectrotemporal modulation filters. The cochlear filter responses that we predicted were the time-average of each of the cochleagram channels (n = 256), which approximate the power spectrum of a sound; the spectrotemporal filters were taken from the same parameterization of the spectrotemporal filter model that we used elsewhere in paper (n = 29,736). We used the same regularized regression procedure that we used for the voxel predictions, modeling each cochlear or spectrotemporal filter's response to each of the 165 natural sounds as a linear combination of network units in a given layer (Figures 5A and 5B). We performed the identical procedure with the units from the untrained, random-filter network.

Layerwise performance on word, genre, and speaker tasks: Training stimulus sets

To examine the task relevance of the features in different network layers, we examined the ability of each layer to support performance of the word and genre task that we trained the network to perform. For training stimuli, we used more than 1.5 million examples for each task (i.e., a subset of those that were used for network optimization).

To examine the generality of the network's learned representations, we tested the ability of features in different network layers to perform a speaker identification task on which the network was not trained. We trained and evaluated classifiers used a subset of the speech stimuli that were used as training stimuli for the word task. Because we wanted a large number of examples per speaker, we used the subset from the WSJ corpus where each of 199 speakers had at least seventy-five hundred examples in our training set, leaving us a dataset of more than two million clips.

Layerwise performance on word, genre, and speaker tasks: Classifier optimization

We trained linear (softmax) classifiers for each layer on each of three tasks: word, genre, and speaker classification. We fixed all of the weights in the network and optimized only a softmax classifier that took a given layer's output as its as input. We used a cross-entropy loss function and updated weights with stochastic gradient descent and a batch size of 256. Because optimization hyperparameters can affect classification performance, we tested two learning rates: 10^{-4} and 10^{-5}. These two learning rates were selected based on pilot experiments. The one exception was for layer fc6 of the untrained network, for which a learning rate of 10^{-3} was used (lower learning rates produced much worse performance). For each layer of the trained network and the untrained, random-filter network, as well as for the spectrotemporal filter model and the cochlear model, we trained a classifier with both learning rates for twelve epochs (full passes over the training set). For each feature set, we selected the classifier that had the lowest loss on held-out validation stimuli. If the classifier overfit during the training procedure, we employed early stopping and used the classifier weights from the epoch at which the classifier minimized the loss on the validation stimuli during the training run.

We then evaluated each of these classifiers on a separate, unseen test set. For the word and genre tasks, we used the same stimuli on which we measured human psychophysical performance for the graphs of Figure 2. For the speaker task, we used the subset of the word psychophysical stimuli that included speakers on which the classifier was trained. The layerwise performance for the word, genre, and speaker tasks is shown in Figures 5C–5E, respectively. An additional analysis showing the effect of background noise on the word classifiers is shown in Figures 6A and S7A.

Layerwise representational similarity analysis: Word, genre, and speaker

To supplement the layerwise performance analyses, we employed representational similarity analysis to further explore the representations learned by different network layers (Figure S6). We selected 18 words, randomly selected 50 examples per word, and computed the correlation matrix for each layer's response to each word example (the "representational similarity matrix"). Separately, we computed the Levenshtein edit distance between each pair of words as a measure of word similarity, and then compared each layer's similarity matrix with the resulting word similarity matrix by correlating the two for each layer. We performed the identical procedure with the untrained, random-filter network, the spectrotemporal features, and the cochleagram.

We performed analogous procedures for the genre and speaker tasks. For the genre task, we used all 41 genres (40 stimuli per genre), and estimated intrinsic genre similarity as the proportion of artists shared between genres (based on the human-annotated tags). For the speaker task, we used 20 examples for each of 16 speakers. To estimate similarity between speakers we measured the mean and standard deviation over time of the F0 for each speaker (using STRAIGHT; Kawahara et al. 2008), and the mean frequency of the first formant during "schwa" utterances (using the Mustafa-Bruce formant tracker; Mustafa and Bruce, 2006). We z-scored each of these statistics across speakers and then computed the (Euclidean) distance between speakers in this three-dimensional space, as shown in the speaker distance matrix.

Analyzing network responses to more and less stationary sounds

To test whether responses to stationary sounds might be attenuated in later network layers, we examined the response of all layers to more and less stationary stimuli. We excluded speech and music stimuli from this stationarity analysis to help ensure that any effect we observed was not due to speech or music-selectivity that might be present in later layers of the network or in non-primary auditory cortical regions.

Measuring the stationarity of natural sounds

Stationarity was operationalized as the variability of sound statistics measured in short time windows. For each sound, we computed the cochleagram, divided it into temporal bins, and measured the following statistics in each of these bins: (i) the mean of each frequency channel (capturing the spectrum); (ii) the correlation across different frequency channels; and (iii) the power in a set of temporal modulation filters applied to the envelopes. To capture stationarity across different timescales, we varied the bin size: 50, 100, 200, and 400 ms. As a summary measure of stationarity we computed the standard deviation of each these statistics over time, averaging across statistics and bin width to get a single measure of temporal variability for each sound. A schematic of this procedure can be found in Figure 6B.

Using this measure of stationarity, we selected a set of more noise-like sounds (n = 38; top third according to the stationarity measure) and a set of less noise-like sounds (n = 38; bottom third) from the 113 natural sounds (of the original set of 165) that were neither speech nor music.

Measuring response of network units to more and less stationary stimuli

We measured each unit's response in each layer to each of these 76 sounds (38 more stationary, 38 less stationary). For each unit in a given layer, we took the ratio of the mean response to the less stationary sounds over the mean response to the more stationary sounds. We took the median across all units in a given layer, excluding "dead" units (i.e., units that did not respond at all to either sound set). For the main text figure we took the median over both branches for branched layers; for the supplemental figure we took the median over each branch separately. We performed the identical procedure with the untrained network. The results are shown in Figures 6C and S7B.

Measuring voxel responses to more and less stationary stimuli

Analogously to our analysis of network units, we measured the mean response of each voxel to the 38 less stationary sounds and the 38 more stationary sounds, and took the ratio (less stationary over more stationary). In Figure 6D we show the map of the median of this value over subjects for each voxel. In Figure S7C, we show individual maps.

Examining relationship between network task performance and auditory cortical variance explained

To examine the relationship between how well a network performed the task on which it was trained and how well it predicted auditory cortical responses, we examined task performance and voxelwise variance explained in a subset of the networks that were generated in the process of our first step of architectural hyperparameter search (Figure 7). We examined a random subset of networks rather than all networks because of practical constraints due to the amount of compute time and disk space required to perform the following analysis. We examined fifty-seven different architectures at fourteen different time points during task training (for a total of 798 different networks). Each network was trained for either word recognition or genre classification (respectively, 392 and 406 networks), and we measured how well each network performed the task for which it was trained using approximately 25,000 left-out validation stimuli. On each of these 798 networks, we trained a softmax classifier to convergence (using SGD) while holding the rest of the network parameters constant.

For each of these 798 networks we measured the median voxelwise variance explained, across all selected individual-subject voxels in auditory cortex. For a given voxel and a given network layer, we performed a similar split-half (across sounds) cross validation scheme that we describe above, except that we predicted responses for each scan separately. We determined which layer best predicted each voxel by averaging the variance explained across the predictions of each layer for two sessions, taking the argmax across layers, and then noting the variance explained by that layer in the left-out third session. We repeated this procedure three times, leaving each session out. We took the median of the explained variance for the three left-out splits. For each network, we iterated this procedure over all 7694 selected voxels, and summarized how well that network predicted auditory cortical responses by taking the median across these 7694 voxelwise variance explained measures. We iterated this procedure over all 798 neural networks.

Figure 7A shows the results across the 392 word-trained networks. Figure 7C shows the results for the 406 genre-trained networks. Figures 7B and 7D show the same data, but highlight the trajectories of individual architectures over training time – each subplot highlights the 14 network checkpoints over training for a given architecture.

QUANTIFICATION AND STATISTICAL ANALYSIS

Branch point selection

The branch point for the multitask network was selected with the goal of sharing as many layers as possible without seeing a significant decrement in task performance. Significance was determined by bootstrapping mean performance on the validation data over resamples of the classes (i.e., words or genres) and stimuli (Figure 1D).

Psychophysics statistics

For the plots of human psychophysical performance on the word and genre tasks (Figures 2A and 2G; also in Figures 2C, 2F, and 2I), error bars are within-subject SEM. For corresponding plots of network psychophysical performance (Figures 2B and 2H; also in Figures 2C and 2I), error bars are SEM, bootstrapped over stimuli and classes (either words or genres). For distortion plots (Figures 2E, 2F, S1A, and S1B), error bars are bootstrapped over stimuli. To examine the similarity between pairs of genre confusion matrices

(Figures 2J and 2K), we computed the Spearman correlation coefficient between unwrapped matrices, and evaluated the significance via NHST p values.

ROI analysis statistics

For all ROI (region of interest) analyses (Figures 3B, 4A, 4F, and S3B–S3E), differences were evaluated with either paired t tests or ANOVAs (specified in main text).

Network analysis statistics

For regressions to cochlear and spectrotemporal filter responses (Figures 5A and 5B), error bars are SEM, bootstrapped over filters and train-test splits of sounds. For layerwise classifiers (Figures 5C–5E), error bars are SEM, bootstrapped over class (i.e., words, genres, or speakers) and stimuli.

For the similarity of layerwise representational similarity matrices with the target matrices (Figure S6), error bars are SEM, analytically computed for the correlation coefficient.

For layerwise classification of words, broken down by background noise level (Figures 6A and S7A), error bars are SEM, bootstrapped over stimuli and classes (i.e., words). For the layerwise ratio of response to less v. more stationary sounds (Figures 6C, 6D, and S7B), error bars are SEM, bootstrapped over sounds.

Significance of individual voxel predictions

The statistical significance of individual voxels was not directly relevant to most of the results described in the paper, because the key results involve pooling voxels together either explicitly across ROIs (either functionally-defined or reliability-defined, i.e., all reliable voxels in auditory cortex) or implicitly in coarse-scale patterns evident in the maps. Nevertheless, to assess significance of individual voxel predictions, we compared the variance explained in each voxel to a null model in which the procedure for calculating explained variance was repeated with random response and prediction vectors (Figure S2B).

We compared the observed raw r^2 values (i.e., those obtained before noise correction) with a null distribution obtained by correlating two 82-length vectors of Gaussian noise (because 82 sounds were used to measure r^2 values), taking the median across ten such samples (as we do with ten randomly selected sets of test stimuli), and repeating this procedure ten million times. The likelihood of observing an r^2 value by chance (i.e., when there is no correlation between the underlying variables) can be obtained from its probability under this null distribution. We set a criterion for determining that a single voxel is incorrectly considered significant ($p = 0.05$) and then applied a stringent correction for multiple comparisons (Bonferroni correction) by dividing this threshold by the total number of comparisons that we are making (7694 voxels across the eight subjects).

DATA AND SOFTWARE AVAILABILITY

The Tensorflow network implementation described above, including the trained filter weights, will be made available on the McDermott lab website.

Code and data are available by request to the Lead Contact (alexkell@mit.edu).