PLOS | BIOLOGY

RESEARCH ARTICLE

Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex

Sam V. Norman-Haignere[1,2,3]*, **Josh H. McDermott**[1,4,5]

1 Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **2** Zuckerman Institute of Mind, Brain and Behavior, Columbia University, New York, New York, United States of America, **3** Laboratoire des Sytèmes Perceptifs, Département d'Études Cognitives, ENS, PSL University, CNRS, Paris France, **4** Program in Speech and Hearing Biosciences and Technology, Harvard University, Cambridge, Massachusetts, United States of America, **5** McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

* snormanhaignere@gmail.com

Abstract

A central goal of sensory neuroscience is to construct models that can explain neural responses to natural stimuli. As a consequence, sensory models are often tested by comparing neural responses to natural stimuli with model responses to those stimuli. One challenge is that distinct model features are often correlated across natural stimuli, and thus model features can predict neural responses even if they do not in fact drive them. Here, we propose a simple alternative for testing a sensory model: we synthesize a stimulus that yields the same model response as each of a set of natural stimuli, and test whether the natural and "model-matched" stimuli elicit the same neural responses. We used this approach to test whether a common model of auditory cortex—in which spectrogram-like peripheral input is processed by linear spectrotemporal filters—can explain fMRI responses in humans to natural sounds. Prior studies have that shown that this model has good predictive power throughout auditory cortex, but this finding could reflect feature correlations in natural stimuli. We observed that fMRI responses to natural and model-matched stimuli were nearly equivalent in primary auditory cortex (PAC) but that nonprimary regions, including those selective for music or speech, showed highly divergent responses to the two sound sets. This dissociation between primary and nonprimary regions was less clear from model predictions due to the influence of feature correlations across natural stimuli. Our results provide a signature of hierarchical organization in human auditory cortex, and suggest that nonprimary regions compute higher-order stimulus properties that are not well captured by traditional models. Our methodology enables stronger tests of sensory models and could be broadly applied in other domains.

OPEN ACCESS

Citation: Norman-Haignere SV, McDermott JH (2018) Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. PLoS Biol 16(12): e2005127. https://doi.org/10.1371/journal.pbio.2005127

Academic Editor: Matt Davis, University of Cambridge, United Kingdom of Great Britain and Northern Ireland

Received: December 15, 2017

Accepted: November 8, 2018

Published: December 3, 2018

Data Availability Statement: Our data and stimuli are available here: https://osf.io/73pfv/.

Funding: National Science Foundation (grant number BCS-1634050). Received by JHM. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. NEC Research Award (grant number). Received by JHM. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. McDonnell Scholar Award (grant number). Received by JHM. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. HHMI-funded LSRF Postdoctoral Research Fellowship http://www.lsrf.org/ (grant number). Received by SNH. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: fMRI, functional magnetic resonance imaging; HG, Heschl's gyrus; LH, left hemisphere; NSE, normalized squared error; PAC, primary auditory cortex; RH, right hemisphere; SNR, signal-to-noise ratio.

Author summary

Modeling neural responses to natural stimuli is a core goal of sensory neuroscience. A standard way to test sensory models is to predict responses to natural stimuli. One challenge with this approach is that different features are often correlated across natural stimuli, making their contributions hard to tease apart. We propose an alternative in which we compare neural responses to a natural stimulus and a "model-matched" synthetic stimulus designed to yield the same responses as the natural stimulus. We tested whether a standard model of auditory cortex can explain human cortical responses measured with fMRI. Model-matched and natural stimuli produced nearly equivalent responses in primary auditory cortex, but highly divergent responses in nonprimary regions, including those selective for music or speech. This dissociation was not evident using model predictions because of the influence of feature correlations in natural stimuli. Our results provide a novel signature of hierarchical organization in human auditory cortex, and suggest that nonprimary regions compute higher-order stimulus properties that are not captured by traditional models. The model-matching methodology could be broadly applied in other domains.

Introduction

One definition of understanding a neural system is to be able to build a model that can predict its responses. Responses to natural stimuli are of particular interest, both because natural stimuli are complex and varied and thus provide a strong test of a model, and because sensory systems are presumably adapted to represent features present in natural stimuli [1–3]. The evaluation of models by their ability to predict responses to natural stimuli is now widespread in sensory neuroscience [4–16].

A challenge for this approach is that because natural stimuli are richly structured, the features of a set of natural stimuli in one model (or model stage) are often correlated with the features in other models (or model stages) [17,18]. Model features can thus in principle predict neural responses to a natural stimulus set, even if the neural responses are in fact driven by other features not captured by the model. Related issues have been widely discussed in the receptive field estimation literature [4,19] but have been less noted in cognitive neuroscience [17,18].

A canonical example of this phenomenon occurs in the auditory domain, where there is still considerable uncertainty regarding computational descriptions of cortical processing. Consider a common model of auditory processing, in which a sound waveform is processed by two stages of filters intended to mimic cochlear and cortical filtering, respectively [20] (Fig 1A). The filters in the second model stage are tuned to temporal and spectral modulations in the spectrogram-like representation produced by the cochlea. Such filters and variants thereof are commonly used to account for human perceptual abilities [21–25] and to explain neural responses throughout the auditory pathway [2,7,11,12,26–36]. But in natural stimuli, the responses of these second-stage filters are often correlated with other sound properties, such as semantic categories (Fig 1B) [37], which can confound the interpretation of neural responses. Speech, for instance, has a distinctive temporal modulation rate that corresponds loosely to the rate of syllabic patterning [38], music has distinctive temporal modulations reflective of its beat structure [39], and both speech and music have characteristic spectral modulations due to harmonic frequency structure [20]. However, speech, music, and other natural sounds also have many unique properties that are not captured by spectrotemporal modulation alone [40]. Thus, if a neuron responds more to speech than to other sounds,

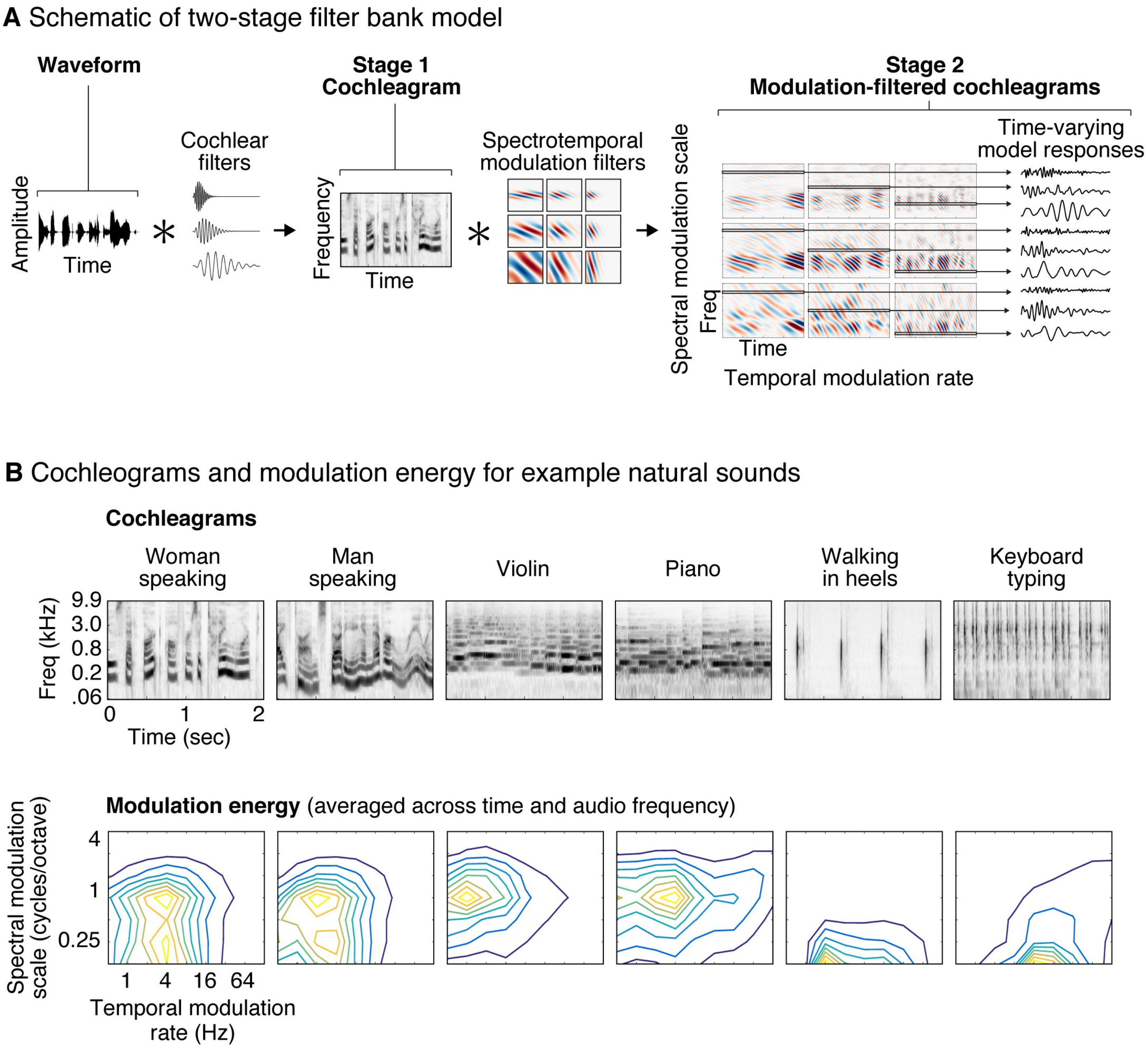
A Schematic of two-stage filter bank model
Waveform
Amplitude
Time
Cochlear filters
Stage 1 Cochleagram
Frequency
Time
Spectrotemporal modulation filters
Stage 2 Modulation-filtered cochleagrams
Time-varying model responses
Spectral modulation scale
Freq
Time
Temporal modulation rate
B Cochleograms and modulation energy for example natural sounds
Cochleagrams
Woman speaking
Man speaking
Violin
Piano
Walking in heels
Keyboard typing
Freq (kHz)
9.9
3.0
0.8
0.2
.06
0
1
2
Time (sec)
Modulation energy (averaged across time and audio frequency)
Spectral modulation scale (cycles/octave)
4
1
0.25
1
4
16
64
Temporal modulation rate (Hz)

Fig 1. Illustration of the auditory model tested in this study. (A) The model consists of two cascaded stages of filtering. In the first stage, a cochleagram is computed by convolving each sound with audio filters tuned to different frequencies, extracting the temporal envelope of the resulting filter responses, and applying a compressive nonlinearity to simulate the effect of cochlear amplification (for simplicity, envelope extraction and compression are not illustrated in the figure). The result is a spectrogram-like structure that represents sound energy as a function of time and frequency. In the second stage, the cochleagram is convolved in time and frequency with filters that are tuned to different rates of temporal and spectral modulation. The output of the second stage can be conceptualized as a set of filtered cochleagrams, each highlighting modulations at a particular temporal rate and spectral scale. Each frequency channel of these filtered cochleagrams represents the time-varying output of a single model feature that is tuned to audio frequency, temporal modulation rate, and spectral modulation scale. (B) Cochleagrams and modulation spectra are shown for six example natural sounds. Modulation spectra plot the energy (variance) of the second-stage filter responses as a function of temporal modulation rate and spectral modulation scale, averaged across time and audio frequency. Different classes of sounds have characteristic modulation spectra.

https://doi.org/10.1371/journal.pbio.2005127.g001

modulation filters may be able to predict the neuron's response, even if the response is driven by another property of speech that is not captured by such filters. This is what we term a "stimulus-driven response correlation," created when different stimulus properties (e.g., spectrotemporal modulations and semantic categories) are correlated within a particular stimulus set, making their contribution to the neural response difficult to tease apart.

Here, we propose a complementary method for evaluating models that circumvents the challenge of stimulus-driven response correlations. The idea is simple: we synthesize a stimulus that yields the same response in a model as a natural stimulus, and then test whether the "model-matched" stimulus elicits the same neural response as the natural stimulus. The synthesized sounds are not influenced by the correlations between different feature sets that may exist in natural stimuli because they are constrained only by the features in the model. As a result, they generally differ in other properties that could potentially be important to the neural response, and often sound markedly different from their natural counterparts. Comparing responses to natural and model-matched sounds thus provides a strong test of the model's explanatory power.

We demonstrate the method by using it to evaluate whether a common filter bank model of auditory cortex can explain human cortical responses to natural sounds measured with fMRI. Many prior fMRI studies of auditory cortex have identified aspects of cortical tuning that are unique to nonprimary regions [16,17,41], such as selectivity for voice [42], speech [43,44], and music [45–47]. At the same time, other studies have demonstrated that the standard filter bank model has relatively good predictive accuracy throughout primary and nonprimary regions [7,12,16], raising the possibility that primary and nonprimary regions encode sound using similar representations. Alternatively, such predictions could in part reflect stimulus-driven correlations. Here, we addressed this question by comparing cortical fMRI responses to natural and model-matched stimuli.

The model-matched stimuli were synthesized to yield the same response as a natural sound in one of several models of varying complexity, ranging from a model of just the cochlea's response to the two-stage spectrotemporal filter bank model shown in Fig 1A [20]. Our results show that tuning for temporal and spectral modulations explains much of the voxel response to natural sounds in human primary auditory cortex (PAC) but much less of the response in nonprimary areas. This functional difference between primary and nonprimary regions was much less evident using conventional model predictions because of the effect of stimulus-driven response correlations. Our findings provide novel evidence for functional differentiation between primary and nonprimary auditory cortex, and suggest that nonprimary regions build higher-order representations that cannot be explained by standard models. Our methodology could provide stronger tests of neural models in any system for which models are used to predict neural responses.

Results

Overview of model-matching method and underlying assumptions

The goal of this paper was to test whether conventional auditory models can explain voxel responses in auditory cortex to natural sounds. The models we consider are described by a set of model features ($m_k(t)$), each of which has a time-varying response to sound determined by the feature's filter (Fig 2A). In general, the response of these features will differ across natural sounds, both in their temporal pattern and their time-averaged properties (S1A Fig). The BOLD signal reflects a time-averaged measure of neural activity, and thus we expect that if a model provides a good description of the underlying neural responses, any two sounds with the same time-averaged model responses should yield the same fMRI response, even if the

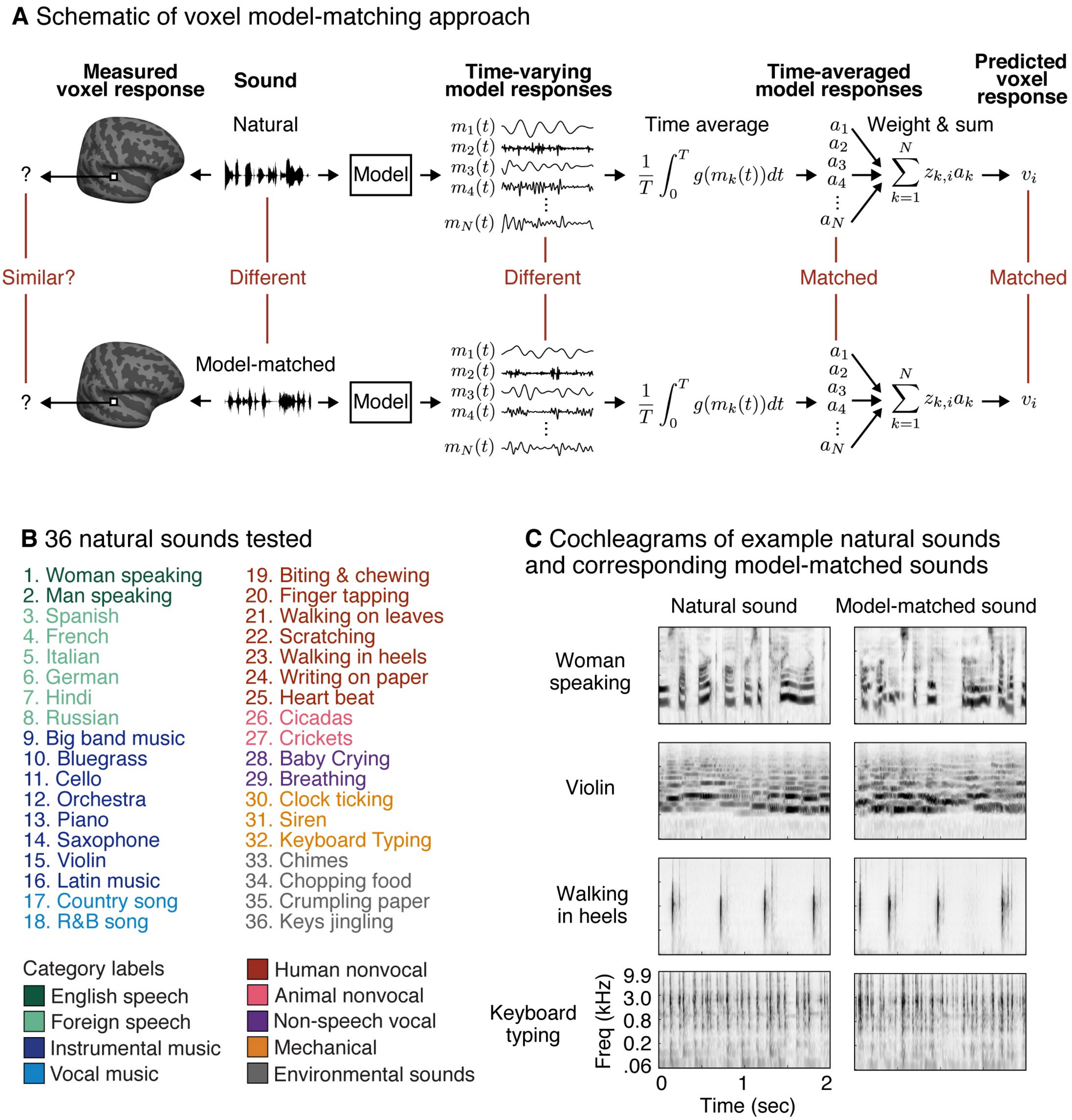

Fig 2. Model-matching methodology and experimental stimuli. (A) The logic of the model-matching procedure, as applied to fMRI. The models we consider are defined by the time-varying response of a set of model features ($m_k(t)$) to a sound (as in the auditory model shown in Fig 1A). Because fMRI is thought to pool activity across neurons and time, we modeled fMRI voxel responses as weighted sums of time-averaged model responses (Eqs 1 and 2, with a_k corresponding to the time-averaged model responses and $z_{k,i}$ to the weight of model feature k in voxel i). Model-matched sounds were designed to produce the same time-averaged response for all of the features in the model (all a_k matched) and thus to yield the same voxel response (for voxels containing neurons that can be approximated by the model features), regardless of how these time-averaged activities are weighted. The temporal response pattern of the model features was otherwise unconstrained. As a consequence, the model-matched sounds were distinct from the natural sounds to which they were matched. (B) Stimuli were derived from a set of 36 natural sounds. The sounds were selected to produce high response variance in auditory cortical voxels, based on the results of a prior study [45]. Font color denotes membership in

one of nine semantic categories (as determined by human listeners [45]). (C) Cochleagrams are shown for four natural and model-matched sounds constrained by the spectrotemporal modulation model shown in Fig 1A.

https://doi.org/10.1371/journal.pbio.2005127.g002

temporal pattern of the response is different. To test this prediction, we iteratively modified a noise stimulus (that was initially unstructured) so as to match the time-averaged model responses (S1B Fig), similar to methods for texture synthesis [40,48–50]. Because the temporal patterns of the model responses are unconstrained, the model-matched sounds differ from the natural sounds to which they were matched.

Formally, we assume that the response of a voxel to a sound can be approximated as the weighted sum of time-averaged neuronal firing rates. Here, we assume the voxel response to be a single number because the sounds we present are short relative to the timescale of the BOLD response. Our goal is to test whether these model feature responses approximate neuronal responses within a voxel, in which case we should be able to approximate the voxel's response (v_i) as a weighted sum of time-averaged model responses (a_k) (Fig 2A):

$$a_k = \frac{1}{T}\int_0^T g(m_k(t))dt \tag{1}$$

$$v_i = \sum_{k=1}^{N} z_{k,i}a_k \tag{2}$$

where g is an (unknown) point-wise function that maps the model responses to a neuronal firing rate (e.g., a rectifying nonlinearity), $z_{k,i}$ is the weight of model feature k in voxel i, and T is the duration of the response to a sound. The most common approach for testing Eqs 1 and 2 is to estimate the weights ($z_{k,i}$) that best predict a given voxel's response to natural sounds (for a particular choice of g) and to assess the cross-validated prediction accuracy of the model using these weights (via explained variance). Here, we instead test the above equations by synthesizing a "model-matched" sound that should yield the same voxel response as a natural sound for all voxels that are accurately described by the model (Fig 2A). We then test the model's validity by assessing whether the voxel responses to the two sounds are similar.

In principle, one could synthesize a separate model-matched sound for each voxel after learning the weights ($z_{k,i}$). However, this approach is impractical given the many thousands of voxels in auditory cortex. Instead, we matched the time-averaged response of all features in the model (i.e., all a_k in Eq 2 are matched; see Fig 2A), which guarantees that all voxel responses that can be explained by the model should be matched, regardless of that voxel's weights. We accomplished this objective by matching the histogram of each feature's response (S1 Fig; see "Model-matching synthesis algorithm" in Materials and methods) [48]. Histogram matching implicitly equates the time-averaged response of the model features for any point-wise transformation (g) since, for any such transformation, the time-averaged response can be approximated via its histogram. It thus obviates the need to choose a particular nonlinearity.

Whether or not a voxel responds similarly to natural and model-matched sounds depends on the response properties of the model features and underlying neurons. If the model features are good approximations to the neurons in a voxel, then the voxel response to natural and model-matched sounds should be similar; if not, they could differ. Here, we consider model features that are tuned to different patterns of temporal and/or spectral modulation [20] in a "cochleagram" (Fig 1A) produced by passing a sound through filters designed to mimic cochlear tuning. Each model feature is associated with a time-frequency filter tuned to a particular temporal rate and/or scale, as well as to a particular audio frequency. The response of each model feature is computed by convolving the spectrotemporal filter with the cochleagram.

Although the response time courses of the models considered here are sufficient to reconstruct the stimulus with high accuracy, the time-averaged properties of the filters, as captured by a histogram, are not. As a consequence, the model-matched sounds differed from the natural sounds they were matched to. Indeed, many of the model-matched stimuli sound unnatural (see http://mcdermottlab.mit.edu/svnh/model-matching/Stimuli_from_Model-Matching_Experiment.html for examples). This observation demonstrates that the time-averaged properties of the model's features, which approximately capture the modulation spectrum (Fig 2A), fail to capture many perceptually salient properties of natural stimuli (e.g., the presence of phonemic structure in speech or melodic contours in music). This additional structure is conveyed by temporal patterns in the feature responses, which are not made explicit by the model but which might be extracted by additional layers of processing not present in modulation filter models. If the neurons in a voxel respond to such higher-order properties (e.g., the presence of a phoneme or melodic contour), we might expect their time-averaged response to differ between natural and model-matched sounds. Thus, by measuring the similarity of voxel responses to natural and model-matched sounds, we can test whether the features of the filter bank model are sufficient to explain their response, or whether other features are needed.

Comparing fMRI responses to natural and model-matched sounds

We measured fMRI responses to a diverse set of 36 natural sounds and their corresponding model-matched sounds (Fig 2B). Each sound was originally 10 seconds in duration, but the sounds were broken up into successively presented 2-second excerpts to accommodate the fMRI scanning procedure (S2 Fig; see "Stimulus presentation and scanning procedure" in Materials and methods). The model-matched sounds were constrained by all of the features from the two-stage filter bank model shown in Fig 1A (see below for results from sounds constrained by simpler models). We first plot the response of two example voxels from a single subject (Fig 3A), which illustrate some of the dominant trends in the data. One voxel was located in the low-frequency area of the "high-low-high" tonotopic gradient thought to span PAC, and which is organized in a roughly V-shaped pattern [51–55]. Another voxel was located outside of tonotopically defined PAC. We note that how best to define PAC is a matter of active debate [54,56–59], and thus we have quantified our results using both tonotopic and anatomical definitions of PAC (described below).

As shown in Fig 3A, the response of the primary voxel to natural and model-matched sounds was similar. By contrast, the nonprimary voxel responded notably less to the model-matched sounds. We quantified the dissimilarity of responses to natural and model-matched sounds by computing the squared error between corresponding pairs of natural and model-matched sounds, normalized by the squared error that would be expected if there was no correspondence between the two sound sets (see "Normalized squared error" in Materials and methods). We quantified response differences using the squared error rather than the correlation because model matching makes no prediction for how responses to natural and model-matched sounds should differ if the model is inaccurate, and, in practice, responses to model-matched sounds were often weaker in nonprimary regions, a phenomena that would not have been captured by correlation. At the end of the results, we quantify how natural and model-matched sounds differ by comparing correlation and squared error metrics.

For these example voxels, the normalized squared error (NSE) was higher for the nonprimary voxel (NSE = 0.729) than the primary voxel (NSE = 0.101), reflecting the fact that the

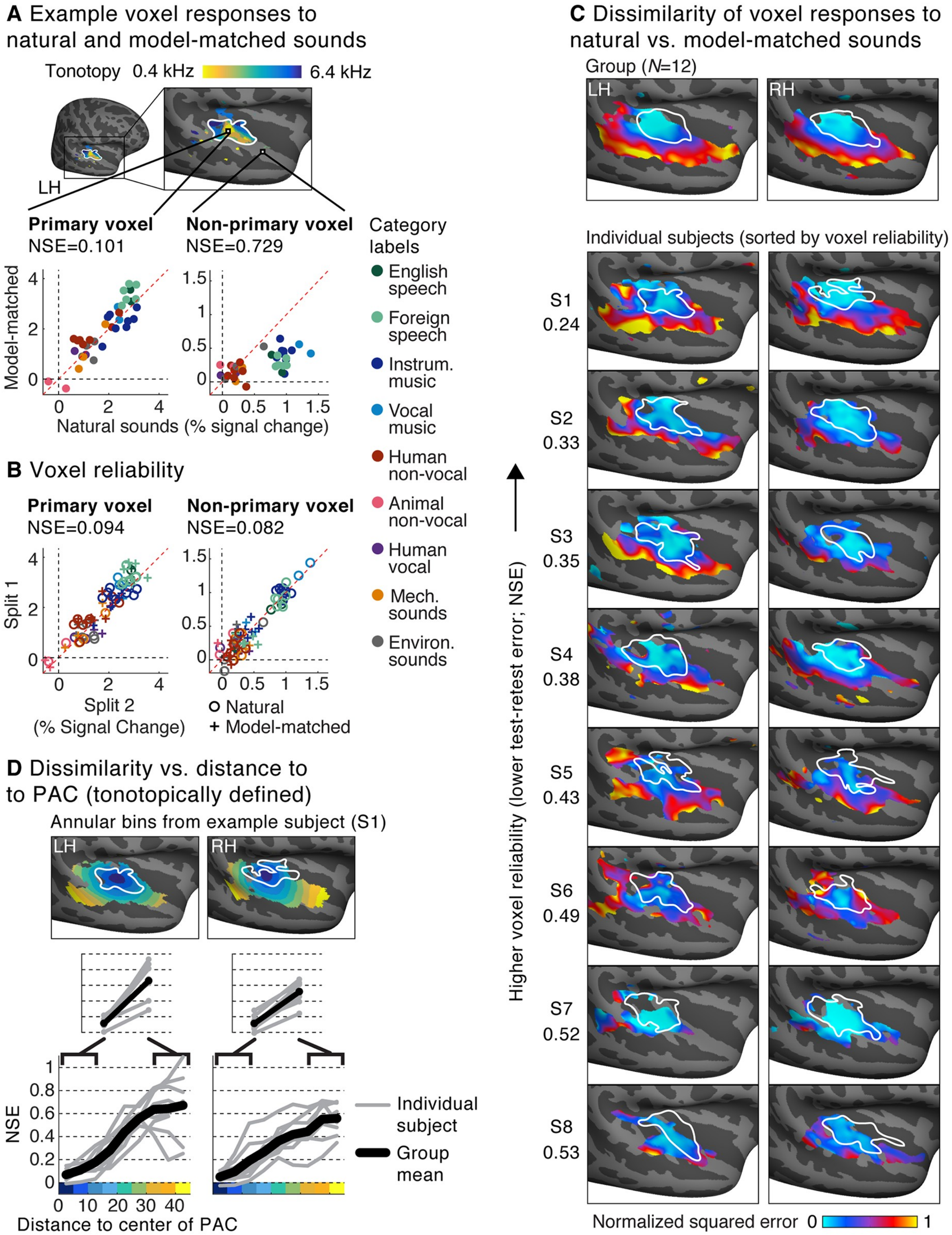

Fig 3. Voxel responses to natural and model-matched sounds. (A) Responses to natural and model-matched sounds from two example voxels from a single subject. One voxel is drawn from the low-frequency region of PAC (defined tonotopically) and one from outside of PAC. A tonotopic map measured in the same subject is shown for anatomical comparison; the map plots the pure tone frequency that produced the highest voxel response. Each dot represents the response to a single pair of natural and model-matched sounds. The primary voxel responded similarly to natural and model-matched sounds, while the nonprimary voxel exhibited a weaker response to model-

matched sounds. We quantified the dissimilarity of voxel responses to natural and model-matched sounds using a normalized squared error metric (NSE) metric (see text for details). (B) Split-half reliability of the responses to natural (circles) and model-matched sounds (crosses) for the two voxels shown in panel A. Both primary and nonprimary voxels exhibited a reliable response (and thus a low NSE between the two measurements). (C) Maps plotting the NSE between each voxel's response to natural and model-matched sounds, corrected for noise in fMRI measurements (see S4 Fig for uncorrected maps). Maps are shown both for voxel responses from eight individual subjects (who were scanned more than the other subjects) and for group responses averaged across 12 subjects in standardized anatomical coordinates (top). The white outline plots the boundaries of PAC, defined tonotopically. Only voxels with a reliable response were included (see text for details). Subjects are sorted by the median test-retest reliability of their voxel responses in auditory cortex, as measured by the NSE (the number to the left of the maps for each subject). (D) A summary figure plotting the dissimilarity of voxel responses to natural and model-matched sounds as a function of distance to the low-frequency region of PAC (see S5 Fig for an anatomically based analysis). This figure was computed from the individual subject maps shown in panel C. Voxels were binned based on their distance to PAC in 5-mm intervals. The bins for one example subject (S1) are plotted. Each gray line represents a single subject (for each bin, the median NSE value across voxels is plotted), and the black line represents the average across subjects. Primary and nonprimary auditory cortex were defined as the average NSE value across the three bins closest and farthest from PAC (inset). In every subject and hemisphere, we observed larger NSE values in nonprimary regions. Note that the left hemisphere has been flipped in all panels to facilitate comparison between the left and right hemispheres. LH, left hemisphere; PAC, primary auditory cortex; RH, right hemisphere.

https://doi.org/10.1371/journal.pbio.2005127.g003

nonprimary voxel showed a more dissimilar response to natural and model-matched sounds. Moreover, most of the error between responses to natural and model-matched sounds in the primary voxel could be attributed to noise in the fMRI measurements, because a similar NSE value was observed between two independent measurements of the voxel's response to natural and model-matched sounds (NSE = 0.094) (Fig 3B). By contrast, in the nonprimary voxel, the test-retest NSE (NSE = 0.082) was much lower than the NSE between responses to natural and model-matched sounds, indicating that the difference in response to natural and model-matched sounds cannot be explained by a lower signal-to-noise ratio (SNR).

We quantified these effects across voxels by plotting the NSE between responses to natural and model-matched sounds for each voxel (Fig 3C). Maps were computed from voxel responses in eight individual subjects who were scanned substantially more than the other subjects (see "Participants" in Materials and methods for details) and from responses that were averaged across all twelve subjects after aligning their brains. Data were collected using two different experiment paradigms that differed in the sounds that were repeated within a scanning session. The results were similar between the two paradigms (S3 Fig), and so we describe them together (see Materials and methods for details; subjects S1, S2, S3, S7, and S8 were scanned in Paradigm I; subjects S4, S5, and S6 were scanned in Paradigm II. Group results are based on data from Paradigm I). In Paradigm I, only responses to natural sounds were repeated, while in Paradigm II, both natural and model-matched sounds were repeated. Only voxels with a reliable response are plotted (test-retest NSE < 0.4; see "Evaluating the noise-corrected NSE with simulated data" in Materials and methods for a justification of this criterion; reliability was calculated using natural sounds for Paradigm I and both natural and model-matched sounds for Paradigm II). Subjects have been ordered by the overall reliability of their data (median test-retest NSE across the superior temporal plane and gyrus, evaluated using natural sounds so that we could apply the same metric to subjects from Paradigms I and II). These maps have been corrected for noise in the fMRI measurements (see "Noise-correcting the NSE" in Materials and methods), but the results were similar without correction (S4 Fig).

Both group and individual subject maps revealed a substantial change across the cortex in the similarity of responses to natural and model-matched sounds. Voxels in PAC showed a similar response to natural and model-matched sounds with noise-corrected NSEs approaching 0, indicating nearly identical responses. Moving away from PAC, NSE values rose substantially, reaching values near 1 in some voxels far from PAC (Fig 3C). This pattern of results suggests that the filter bank model can explain much of the voxel response in primary regions but much less of the response in nonprimary regions, plausibly because nonprimary regions respond to higher-order features not made explicit by the model. This result is suggestive of a hierarchy of feature selectivity in auditory cortex and demonstrates where in the cortex the standard filter bank model fails to explain voxel responses.

We quantified the gradient we observed between primary and nonprimary voxels by binning the NSE of voxels from individual subjects based on their distance to PAC. Similar results were observed for tonotopic (Fig 3D) and anatomical definitions of PAC (S5 Fig; PAC was defined either as the center of the high-low-high gradient or as the center of anatomical region TE1.1 [58], in posteromedial Heschl's gyrus (HG)). To directly compare primary and nonprimary regions, we then averaged NSE values within the three bins nearest and farthest from PAC (Fig 3D, inset). This analysis revealed that responses to natural and model-matched sounds became more dissimilar in nonprimary regions in both the left and right hemisphere of every subject tested, leading to a highly significant difference between primary and nonprimary regions ($p < 0.01$ via sign test for both hemispheres and for both tonotopic and anatomical definitions of PAC). The gradient between primary and nonprimary regions was observed in both scanning paradigms, regardless of smoothing (S3 Fig), and could not be explained by selectivity for intelligible speech (a similar pattern was observed when intelligible speech sounds were excluded from the analysis; see S6 Fig). These results also could not be explained by variations in voxel reliability across brain regions, both because our NSE measures were noise-corrected and because voxel responses were similarly reliable throughout primary and nonprimary regions (S4C Fig). As a consequence of the similar reliability across auditory cortex, the increase in the NSE between natural and model-matched sounds between primary and nonprimary regions was significantly greater than the change in voxel reliability. This was true using both corrected and uncorrected values for the natural versus model-matched NSE, both tonotopic and anatomical definitions of PAC, and with reliability measured using just natural sounds (for Paradigm I) and both natural and model-matched sounds (for Paradigm II) ($p < 0.01$ via sign test in all cases; see S3 Fig for a breakdown by paradigm). Thus, our results demonstrate that the modulation filter bank model is worse at accounting for voxel responses in nonprimary regions.

Comparing responses to sounds matched on subsets of model features

We next used a similar approach to test whether responses in PAC could be explained by simpler models. For example, if neurons in a voxel are tuned primarily to audio frequency, then all sounds with similar spectra should produce similar responses, regardless of their modulation properties. To test such alternative models, we synthesized three new sounds for each natural sound. Each synthetic sound was matched on a different subset of features from the full model (Fig 4A). One sound was synthesized to have the same marginal distribution of cochlear envelopes as a natural sound and, thus, a similar audio spectrum, but its modulation properties were otherwise unconstrained. Another sound was constrained to have the same temporal modulation statistics within each cochlear frequency channel, computed using a bank of modulation filters modulated in time but not frequency. A third sound was synthesized to have matched spectral modulation statistics, computed from a bank of filters modulated in frequency but not time. All of the modulation-matched sounds also had matched cochlear marginal statistics, thus making it possible to test whether adding modulation structure enhanced the similarity of cortical responses to natural and model-matched sounds.

The results of this analysis suggest that all of the model features are necessary to account for voxel responses to natural sounds in PAC (Fig 4B and 4C; S7 Fig). Responses to model-matched sounds constrained just by cochlear statistics differed substantially from responses to natural sounds even in PAC, leading to significantly larger NSE values than those observed for

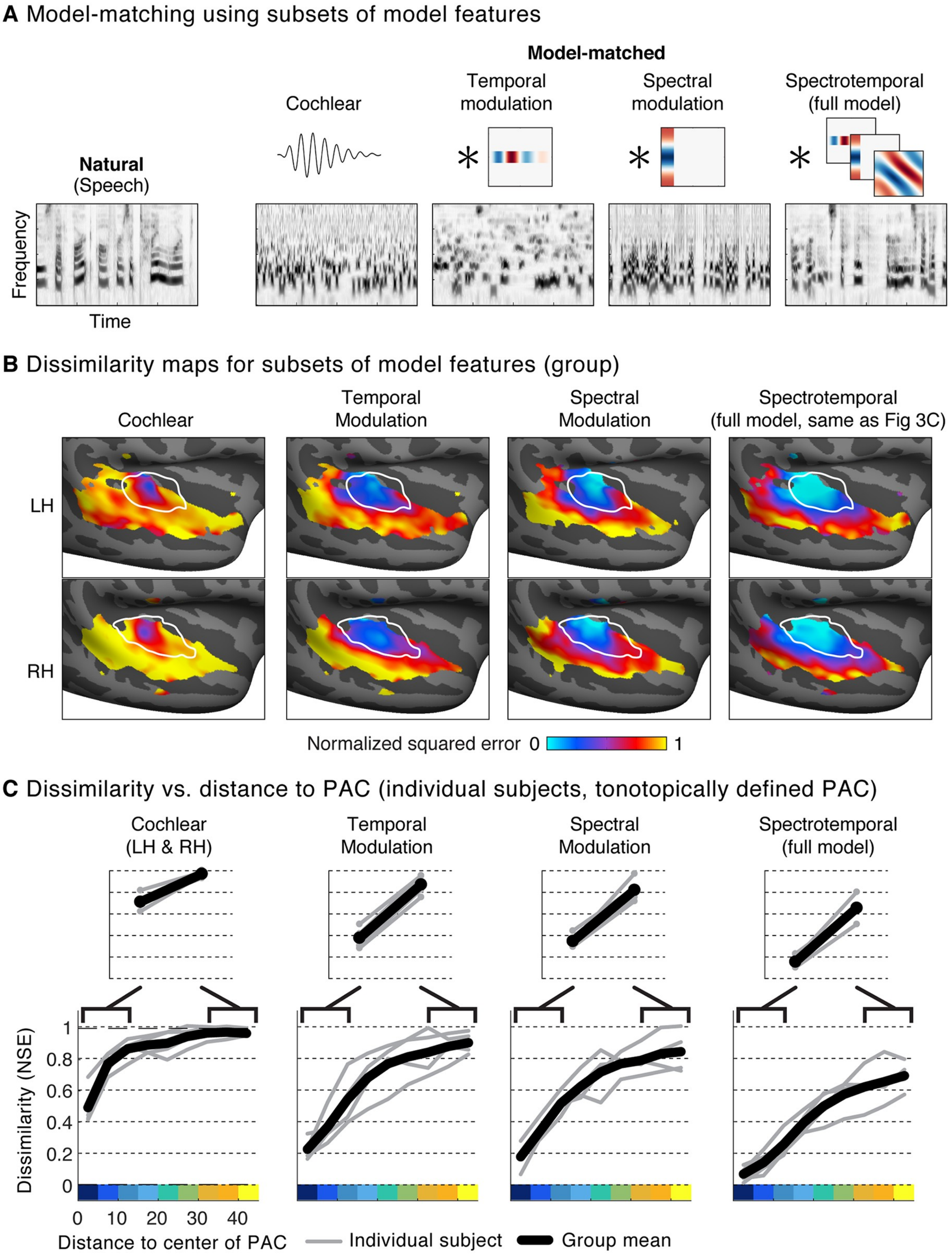

Fig 4. Comparison of responses to model-matched sounds constrained by different models. (A) Cochleagrams for an example natural sound and several corresponding model-matched sounds constrained by subsets of features from the full two-stage model. Cochlear-matched sounds were constrained by time-averaged statistics of the cochleagram representation but not by any responses from the second-stage filters. As a consequence, they had a similar spectrum and overall depth of modulation as the corresponding natural sound, but were otherwise unconstrained. The other three sounds were additionally constrained by the response of second-stage filters, tuned

either to temporal modulation, spectral modulation, or both temporal and spectral modulation (the full model used in Fig 3). Temporal modulation filters were convolved separately in time with each cochlear frequency channel. Spectral modulation filters were convolved in frequency with each time slice of the cochleagram. In this example, the absence of spectral modulation filters causes the frequency channels to become less correlated, while the absence of temporal modulation filters results in a signal with more rapid temporal variations than that present in natural speech. (B) Maps of the NSE between responses to natural and model-matched sounds, constrained by each of the four models. The format is the same as panel 3C. See S7 Fig for maps from individual subjects. (C) Dissimilarity between responses to natural and model-matched sounds versus distance to the low-frequency area of PAC. Format is the same as panel 3D. Results are based on data from the four subjects that participated in Paradigm I, because model-matched sounds constrained by subsets of features were not tested in Paradigm II. LH, left hemisphere; NSE, normalized squared error; PAC, primary auditory cortex; RH, right hemisphere.

https://doi.org/10.1371/journal.pbio.2005127.g004

the full model ($p < 0.001$ in PAC via bootstrapping across subjects; see "Statistics" in Materials and methods). Thus, even though PAC exhibits selectivity for frequency due to tonotopy, this selectivity only accounts for a small fraction of its response to natural sounds. Responses to natural and model-matched sounds in PAC became more similar when the sounds were constrained by either temporal or spectral modulation properties alone (NSE temporal < NSE cochlear: $p < 0.001$ via bootstrapping; NSE spectral < NSE cochlear: $p < 0.001$). However, we only observed NSE values near 0 when sounds were matched in both their temporal and spectral modulation properties (NSE full model < NSE temporal: $p < 0.001$; NSE full model < NSE spectral: $p < 0.001$). These results provide further support for the idea that selectivity for both temporal and spectral modulation is a prominent feature of cortical tuning in PAC [7,32,33]. In nonprimary auditory cortex, we also observed more similar responses when matching sounds on spectrotemporal modulation compared with simpler models (NSE spectrotemporal < NSE cochlear: $p < 0.001$; NSE spectrotemporal < NSE temporal: $p < 0.05$; NSE spectrotemporal < NSE spectral: $p < 0.01$). However, the noise-corrected NSE values were high for all of the models tested, indicating that the modulation model fails to account for a substantial fraction of nonprimary responses.

Predicting responses to natural sounds from model features

Part of the motivation for using model-matched stimuli comes from the more common approach of predicting responses to natural stimuli from the features of a model (e.g., via linear regression). As discussed above, good predictive accuracy is not sufficient to guarantee that the features of a model drive a neural response, due to the potential for correlations between different feature sets across natural stimuli. Model matching provides one way to circumvent this issue, because the synthesized sounds are only constrained by the statistics of the particular model being tested. Here, we test whether our approach yields novel insights compared with simply predicting cortical responses to natural sounds from model features.

We attempted to predict responses to the 36 natural sounds from time-averaged statistics of the same model features used to generate the model-matched sounds (Fig 5A; see S8 Fig for individual subject prediction error maps for the full spectrotemporal model). Specifically, we used ridge regression to predict voxel responses from the amplitude of each model feature's response to each natural sound [7,16], measured as the standard deviation across time (for the cochlear model, we used the mean rather than the standard deviation because the features were the result of an envelope extraction operation, and the mean thus conveyed the amplitude of the filter's response). Because histogram matching approximately matches all time-averaged statistics of a distribution, predictions based on a single time-averaged statistic, such as the standard deviation, provide a conservative estimate of the predictive power of time-averaged statistics. Good predictions in voxels whose responses to model-matched sounds deviated

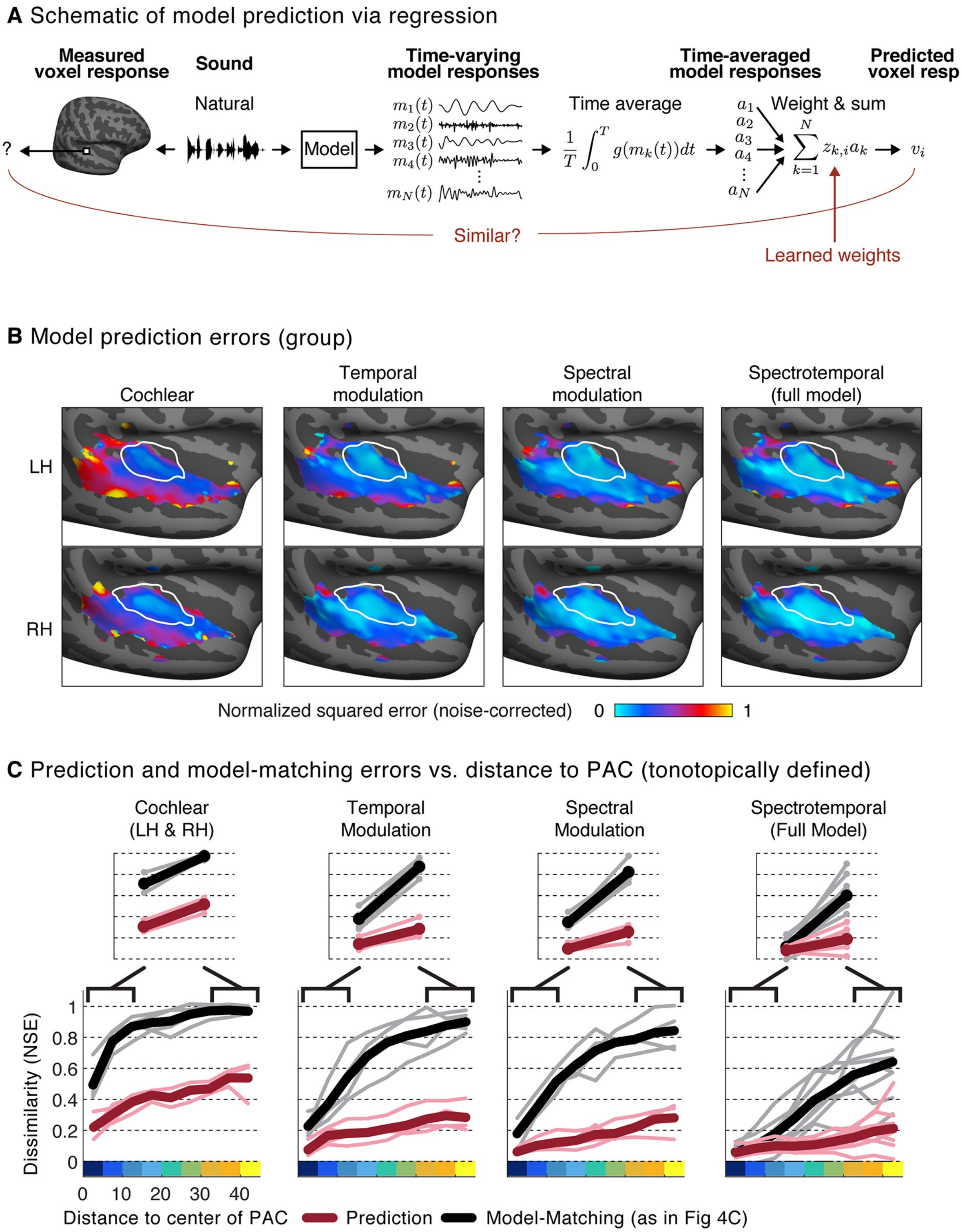

Fig 5. Predicted responses to natural sounds via regression using the same auditory model used to constrain the model-matched sounds. (A) Schematic of regression procedure used to predict neural responses from model features. For each natural sound, we computed the response time course for each feature in the model, as was done for model matching. We then computed a time-averaged measure of each feature's activity (the mean across time for the cochlear features, because they are the result of an envelope operation, and the standard deviation for the modulation features, because they are raw filter outputs) and estimated the weighted combination of these time-averaged

statistics that yielded the best-predicted response (using ridge regression, cross-validated across sounds). (B) Maps showing the prediction error (using the same NSE metric employed in Figs 3 and 4) between measured and predicted responses to natural sounds for the corresponding models shown in Fig 4 (see S8 Fig for maps from individual subjects). (C) Prediction error versus distance to the low-frequency area of PAC (maroon lines: thin lines correspond to individual subjects, thick lines correspond to the group average). For comparison, the corresponding NSE values derived from the model-matching procedure are replotted from Fig 4C (black lines). The analyses are based on individual subject maps. Results for the full model (rightmost plot) are based on data from the same eight subjects shown in Fig 3C. Results for model subsets (cochlear, temporal modulation, and spectral modulation) are based on data from four subjects that were scanned in Paradigm I (sounds constrained by subsets of model features were not tested in Paradigm II). LH, left hemisphere; NSE, normalized squared error; PAC, primary auditory cortex; RH, right hemisphere.

https://doi.org/10.1371/journal.pbio.2005127.g005

from those to natural sounds would thus suggest that prediction-based analyses overestimate the model's explanatory power. We quantified prediction accuracy by measuring the NSE between measured and predicted responses for left-out sounds that were not used to learn the regression weights (see "Model predictions" in Materials and methods).

Overall, we found that voxel responses to natural sounds were substantially more similar to the predicted model responses than to the measured responses to the model-matched stimuli (Fig 5B and 5C), leading to smaller NSEs for model predictions compared with model-matched stimulus responses. This difference was particularly pronounced in nonprimary regions, where we observed relatively good predictions from the full two-stage model despite highly divergent responses to model-matched sounds, leading to a significant interaction between the type of model evaluation (model prediction versus model matching) and region (primary versus nonprimary) ($p < 0.01$ via sign test for both tonotopic and anatomical definitions of PAC; a sign test was used to evaluate whether the change in NSE values between primary and non-primary regions was consistently larger for model matching compared with model prediction). Because the natural and model-matched sounds were matched in the features used for prediction, the divergent responses to the two sound sets imply that the features used for prediction do not in fact drive the response. Thus, good predictions for natural sounds in the presence of divergent model-matched responses must reflect the indirect influence of correlations between the features of the model and the features that actually drive the neuronal response. Model matching thus reveals a novel aspect of functional organization not clearly evident from model predictions by demonstrating the failure of the filter bank model to account for nonprimary responses.

Our prediction analyses were based on responses to a set of 36 natural sounds that was smaller than the sound sets that have been used elsewhere to evaluate model predictions [7,16,45,60]. Because our analyses were cross-validated, small sound sets should reduce prediction accuracy and thus cannot explain our finding that model predictions were better than would be expected given responses to model-matched sounds. Nonetheless, we assessed the robustness of our findings by also predicting responses to a larger set of 165 natural sounds [45]. We observed similar results with this larger sound set, with relatively good prediction accuracy for the full spectrotemporal model throughout primary and nonprimary auditory cortex (S9 Fig).

Another way to assess the utility of the model-matching approach is to train a model to predict natural sounds, and then test its predictive accuracy on model-matched sounds (and vice versa). In practice, this approach yielded similar results to directly comparing responses to natural and model-matched sounds: good cross-predictions in PAC but poor cross-predictions in nonprimary auditory cortex (S10 Fig). This observation is expected given that (a) the model predictions for natural sounds were good throughout auditory cortex and (b) responses to natural and model-matched sounds diverged in nonprimary regions, but it provides a consistency check of the two types of analyses.

Voxel decomposition of responses to natural and model-matched sounds

All of our analyses described thus far were performed on individual voxels, summarized with maps plotting the NSE between each voxel's response to natural and model-matched sounds. However, these error maps do not reveal in what respect the responses to natural and model-matched sounds differ, and, because of the large number of voxels, it is not feasible to simply plot all of their responses. We previously found that voxel responses to natural sounds can be approximated as a weighted sum of a small number of canonical response patterns (components) [45] (Fig 6A). Specifically, six components explained over 80% of the noise-corrected response variance to a diverse set of 165 natural sounds across thousands of voxels. We thus used these six components to summarize the responses to natural and model-matched sounds described here. This analysis was possible because many of the subjects from this experiment also participated in our prior study. As a consequence, we were able to learn a set of voxel weights that reconstructed the component response patterns from our prior study and then apply these same weights to the voxel responses from this experiment (see "Voxel decomposition" in Materials and methods).

We found that all six components exhibited reliable responses to the natural sounds from this experiment (Fig 6B). Two of the components (5 and 6) responded selectively to speech and music, respectively, replicating the selectivity we found previously (last two columns of 6B). Critically, responses to the model-matched sounds were much weaker in these speech- and music-selective components, even for sounds matched on the full model (Fig 6C, last two columns; see S11 Fig for sounds matched on subsets of model features), leading to high NSE values (speech NSE = 0.45; music NSE = 0.55 for the full model, noise-corrected; Fig 6D). By contrast, the other four components, all of which overlapped PAC to varying extents, responded similarly to natural and model-matched sounds constrained by the full model, leading to smaller errors (NSE for Component 1: 0.06, Component 2: 0.12, Component 3: 0.26, Component 4: 0.19) than those for the speech- and music-selective components ($p < 0.001$ for all direct comparisons between the speech- and music-selective components and Components 1, 2, and 4; for Component 3, which had the lowest test-retest reliability, the direct comparison with the music-selective component was significant, $p < 0.01$, and the direct comparison with the speech-selective component was nearly significant, $p = 0.076$; statistics computed via bootstrapping across subjects). These results indicate that selectivity for music and speech cannot be purely explained by standard acoustic features that nonetheless account for much of the voxel response in primary regions.

Our model-matching approach posits that responses should be exactly matched if the model is accurate. If the model is not accurate, the approach makes no prediction about how the responses should differ. Nonetheless, the divergent responses to natural and model-matched sounds in Components 5 and 6 appeared to be largely driven by weaker responses to the model-matched sounds. We verified this observation by comparing the standard deviation of responses to natural and model-matched sounds: the response variation for model-matched sounds decreased sharply in Components 5 and 6, driven by lower overall responses to the model-matched sounds (Fig 6E). In contrast, the noise-corrected correlation remained high (Fig 6F). A similar pattern was also evident in whole-brain maps (S12 Fig): the variation in voxel responses to model-matched sounds constrained by the full model dropped in nonprimary regions (driven by lower responses to the model-matched stimuli), while the correlation remained high. For Components 5 and 6, the high correlations were driven by the fact that model-matched sounds from the component's preferred category produced a higher response than model-matched sounds from other categories (as is evident in Fig 6C). For example, in Component 6, model-matched music produced a lower response than natural music but a

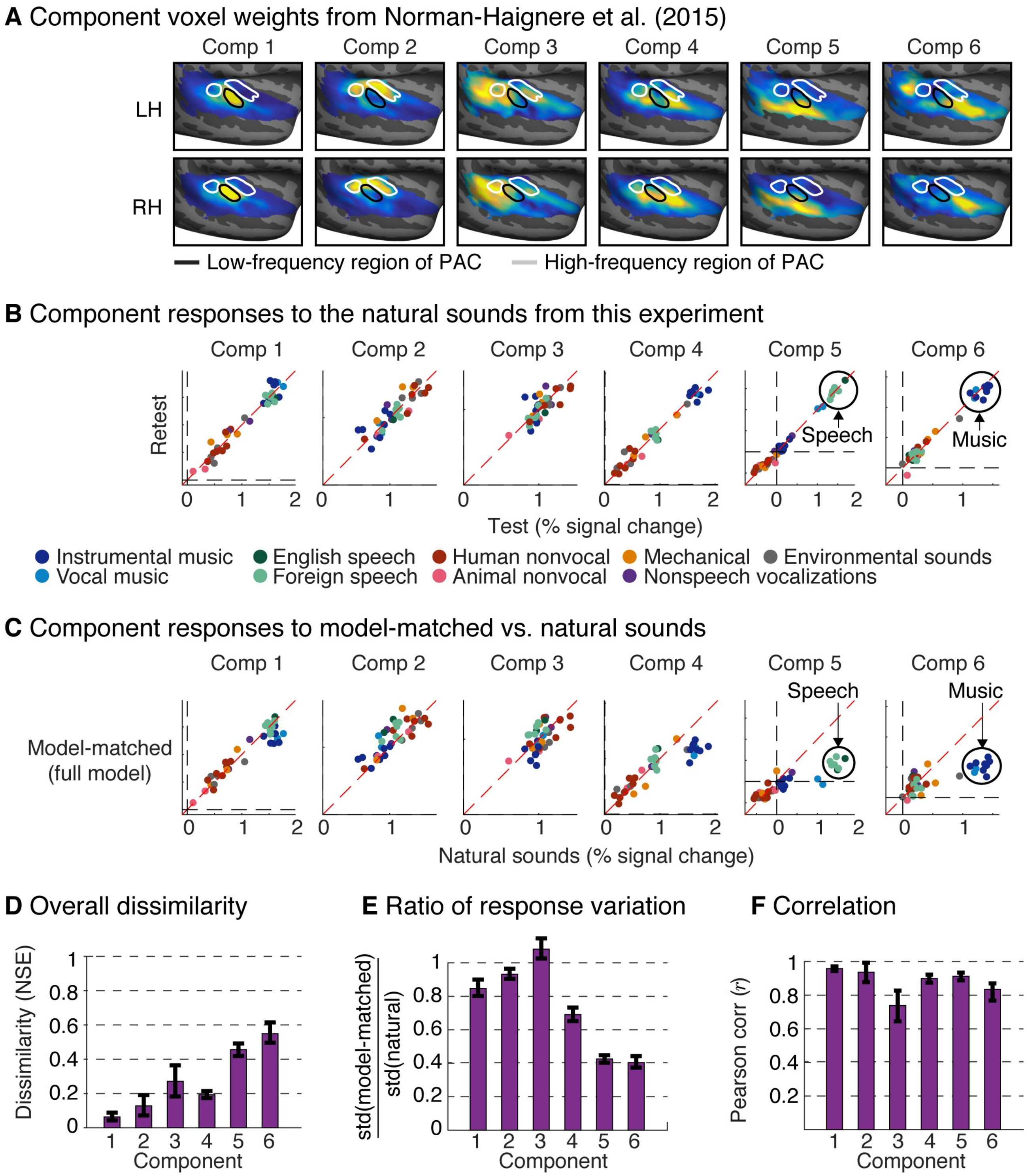

Fig 6. Voxel decomposition of responses to natural and model-matched sounds. Previously, we found that much of the voxel response variance to natural sounds can be approximated as a weighted sum of six canonical response patterns ("components") [45]. This figure shows the response of these components to the natural and model-matched sounds from this experiment. (A) The group component weights from Norman-Haignere and colleagues (2015) [45] are replotted to show where in auditory cortex each component explains the neural response. (B) Test-retest reliability of component responses to the natural sounds from this study. Each data point represents responses to a single sound, with color denoting its semantic category. Components 5 and 6 showed selectivity for speech and music, respectively, as expected (Component 4 also responded most to music because of its selectivity for sounds with pitch). (C) Component responses to natural and model-matched sounds constrained by the complete spectrotemporal model

(see S11 Fig for results using subsets of model features). The speech and music-selective components show a weak response to model-matched sounds, even for sounds constrained by the full model. (D) NSE between responses to natural and model-matched sounds for each component. (E) The ratio of the standard deviation of each component's responses to model-matched and natural sounds (see S12A Fig for corresponding whole-brain maps). (F) Pearson correlation of responses to natural and model-matched sounds (see S12B Fig for corresponding whole-brain maps). All of the metrics in panels D—F are noise-corrected, although the effect of this correction is modest because the component responses are reliable (as is evident in panel B). Error bars correspond to one standard error computed via bootstrapping across subjects. LH, left hemisphere; NSE, normalized squared error; PAC, primary auditory cortex; RH, right hemisphere.

https://doi.org/10.1371/journal.pbio.2005127.g006

higher response than model-matched sounds from other categories ($p < 0.001$, via bootstrapping). The same pattern was evident for Component 6, which responded selectively to speech ($p < 0.001$). This finding suggests that selectivity in nonprimary regions may reflect a mixture of category-specific modulation tuning and responses to higher-order properties specific to music and speech, consistent with prior studies [9,53,61]. The results suggest that the modulation-specific structure driving Components 5 and 6 is correlated across natural sounds with the other properties of music and speech that drive their response. The model-matching approach allows us to see these two contributions to the response, revealing that there is something unique to the response of Components 5 and 6 that is distinct from the other components.

Discussion

We have described a novel approach for evaluating a model of neuronal responses. Given a model, we synthesize a stimulus that yields the same model response as a natural stimulus, and test whether they produce similar neural responses. We applied this approach to test whether voxel responses in human auditory cortex can be explained by a commonly used auditory model based on spectrotemporal modulation. Our results revealed a substantial functional difference between primary and nonprimary regions of human auditory cortex. Many voxels in PAC showed nearly equivalent responses to natural and model-matched sounds constrained by the full spectrotemporal model. We also found that these voxels responded differently when sounds were model matched with only cochlear filter statistics, or with temporal or spectral modulations alone. These findings together suggest that spectrotemporal modulation accounts for much of the voxel response in PAC. By contrast, many voxels in nonprimary regions responded weakly to all of the model-matched sounds, demonstrating that they are only weakly driven by the features captured by the model. This functional difference between primary and nonprimary regions was not clearly evident when the model was evaluated by its response predictions, due to the confounding influences of stimulus-driven correlations across natural stimuli. Model matching thus reveals a novel aspect of functional organization by showing where in the cortex a standard auditory model can explain voxel responses to natural sounds.

Implications for models of auditory cortex

The notion that auditory cortex might be organized hierarchically—i.e., into a series of stages supporting increasingly abstract representations—has been a popular proposal for decades [41,62–64]. Hierarchical organization has some support from anatomical studies [65], and from qualitative observations that responses in nonprimary regions are more complex than those in primary regions [66,67] and more closely aligned with semantically meaningful sound properties [17,45–47,68]. However, there has been little evidence for how primary and nonprimary regions might differ in computational terms [16], and thus it has been unclear what mechanisms underlie the apparent differences in tuning between primary and nonprimary regions.

Most computational models of auditory processing beyond the periphery are based on tuning for modulation [20]. Such models have been used to explain responses throughout the auditory pathway in non-human animals [2,11,26–31,34]. In humans, modulation-based models have been shown to have relatively good predictive accuracy throughout both primary and nonprimary regions [7,12,45], which has led to the hypothesis that sounds are represented in a distributed manner [69]. This view contrasts with the notion of hierarchical organization and, in its most extreme form, suggests that responses to seemingly complex attributes of sound in nonprimary regions (e.g., speech and music selectivity) could reflect the same types of mechanisms used to code sound in PAC.

Our study helps to reconcile the literatures on modulation-based auditory models and hierarchical organization. First, we show that modulation selectivity fails to explain much of the response in nonprimary regions, and that model predictions provide overly optimistic estimates of the model's efficacy. This conclusion follows from the fact that we observed many voxels in nonprimary regions whose response to natural sounds was well predicted by the model and yet produced divergent responses to model-matched sounds. Because the model by definition predicts that natural and model-matched sounds should have equivalent responses, this finding demonstrates a clear model failure.

Conversely, our findings provide further evidence that modulation selectivity is a key feature of functional organization in human PAC [7,32,33,35]. Using both predictions and model matching, we found that the modulation model explains the large majority of the voxel responses in this region. This finding was again not obvious from prior studies using model prediction alone, because the predictions could have been influenced by stimulus-driven correlations, as turned out to be the case in nonprimary regions. By contrast, we found that frequency selectivity, which presumably reflects tonotopy, explained much less response variance in PAC. This finding suggests that modulation selectivity may be a key organizing dimension of PAC.

What features might nonprimary regions of auditory cortex represent? These regions are primarily driven by sound, show signs of having relatively short integration windows [43], and, even when speech selective, respond largely independently of the presence of linguistic structure [43,45], suggesting acoustic rather than linguistic or semantic representations [17]. Moreover, although responses to the model-matched sounds were substantially weaker than responses to natural sounds, the model-matched sounds still drove responses to natural sounds above baseline and were correlated with responses to natural sounds. Thus, one natural hypothesis is that nonprimary regions transform a lower-level acoustic representation, such as the spectrotemporal representation considered here, into a representation that makes behaviorally relevant variables more explicit (e.g., easier to decode). This hypothesis could be tested with hierarchical models that transform the output of modulation filters with additional stages of nonlinear and linear operations [70]. In principle, such models could be fit to existing neural data sets and then evaluated with model-matched stimuli. But because the space of such models is large, some additional constraint is likely to be needed to select models for experimental tests. Such constraints could come from natural sounds and tasks, for example by optimizing for efficient encoding of natural sounds or for performance of ecologically relevant tasks [71–76].

We have recently explored this idea by training a deep neural network to recognize words and musical genres [16] and then comparing the resulting representations with voxel responses. We found that later layers of the network better predicted voxels in nonprimary regions of the cortex, consistent with the notion of hierarchical organization. These predictions could of course be influenced by stimulus-driven correlations, which may explain why the differences in prediction accuracy between layers were modest. Future work could address this question and provide stronger tests of such models by applying model matching to the representation from different layers of a hierarchical model.

Implications and limitations of model matching

The result of our model-matching experiment is an error metric between 0 and 1, indicating the dissimilarity of a neural response to natural and model-matched sounds. What does this number tell us about the type of models that could underlie the neural response? When the error metric is near 1, the models under which responses have been matched are ruled out as descriptions of the voxel response. Because the error metric is noise-corrected, its absolute value is meaningful, and large errors invalidate a model. Our specific implementation matched model responses for all point-wise functions of the filters in question, and thus that family of models is ruled out for voxels with large error.

At the other extreme, errors near 0, like those we observed in PAC, reveal that the voxel responses are consistent with the family of models whose response was matched. The matching procedure employed a specific filter bank, but alternative models might also be matched (for instance, those with filters that can be approximated as linear combinations of the filters used for matching). Small error values thus do not exclude models other than the one we used. However, specific alternative models could be evaluated by measuring their response to the two sets of stimuli used in an experiment (natural and model-matched). Models that give distinct responses to the two stimulus sets could be ruled out for voxels whose responses to the two sets are similar. Conversely, one could also rule out models whose responses to the two sets are similar for voxels whose responses to the two sets are different. We used this approach to investigate different types of spectrotemporal filter banks (S15 Fig), finding that a range of alternative filter banks had matched statistics for the natural and model-matched sounds tested here (see Variants of the spectrotemporal filter model in Materials and methods). This finding suggests that a wide range of spectrotemporal filter models can be ruled out as models of nonprimary auditory cortex. Our stimuli and fMRI data are available, so that alternative models can be evaluated using this approach: https://osf.io/73pfv/.

In other situations, matching with one model may entail matching with another, but not vice versa. This was the case for the four models we compared in Fig 4—the full spectrotemporal model is inclusive of the other models. The higher NSE values observed with the other models provides evidence for the necessity of the spectrotemporal model features.

As with any method for model evaluation, the interpretation of our results is constrained by the resolution of the brain measurements used for evaluation. Because fMRI is primarily sensitive to brain responses that are spatially clustered, our results bear most directly on aspects of cortical tuning that are organized at the scale of voxels. Our results were robust to the exact size of the voxels tested and the amount of spatial smoothing, suggesting that our results hold for spatial scales on the order of millimeters to centimeters. But even small voxels pool activity across neurons and across time, and thus it is possible that voxels with similar responses to natural and model-matched sounds might nonetheless contain neurons that show more divergent responses or that have temporal response properties that differ from the model. This fact may partially explain why electrophysiological recordings in animals have found that linear spectrotemporal filters are insufficient to account for responses in PAC [13,77–79]. Future work could apply model matching to neuronal responses measured electrophysiologically to test models at a finer spatial and temporal scale. For example, one could synthesize model-matched sounds that should yield the same firing rate as a natural sound given a model of an individual neuron's response. At the scale of fMRI voxels, however, linear spectrotemporal filters provide a good description of PAC, potentially because neurons with similar modulation selectivity are spatially clustered.

Because the spatial pooling of fMRI can obscure neural responses that are heterogeneous across nearby neurons, voxel responses to natural and model-matched stimuli could in principle also be more dissimilar than the responses of the underlying neural populations. That is, there could be neural populations that respond similarly to natural and model-matched sounds, but which do not contribute to the voxel NSE because they are not clustered at a coarse enough scale and thus do not differentially drive voxel responses to different sounds within a stimulus set. A high NSE thus demonstrates a model failure (because it implies underlying neurons that respond differently to natural and model-matched sounds), but it does not preclude the possibility that the voxel also contains some neurons that are well described by the model features. We note that these limitations are not specific to the model-matching approach and apply equally to evaluations of models by their predictions of fMRI responses—in both cases, finer-grained brain measurements will enable finer-grained model tests.

Relation to prior work on perceptual metamers and texture synthesis

Our approach to model matching is an extension of methods for texture synthesis originally developed in image processing and computer vision [48,49], and later applied to sound texture [40] and visual texture perception [80,81]. In texture synthesis, the goal is typically to test whether a set of statistical features could underlie perception by testing whether synthetic stimuli with the same statistics are metameric, i.e., whether they look or sound the same as a real-world texture. The implementation of our synthesis procedure is inspired by classic texture synthesis methods [48], but the scientific application differs notably in that we evaluate the model by the similarity of neural responses rather than the similarity that is perceived by a human observer. Indeed, many of the model-matched stimuli sounded unnatural, demonstrating that the modulation spectrum fails to capture higher-order properties of natural sounds to which listeners are sensitive (e.g., the presence of phonemic or melodic structure). This observation reveals the insufficiency of the modulation spectrum as a complete account of perception but does not place strong constraints on whether particular neural stages are well described by the model. The fact that responses to natural and model-matched sounds diverged in nonprimary regions of auditory cortex suggests that those regions may be driven by higher-order structure not made explicit by the modulation model, which we could not have concluded from perceptual observations alone.

The most similar previous approach involved comparing the strength of cortical responses to visual textures synthesized from different classes of statistics of a wavelet filter bank model [81]. Although we also compared cortical responses to sounds synthesized from different model statistics, the key comparison was between responses to individual natural and synthesized sounds, which is critical to identifying regions of the brain that are not well explained by a model.

The modulation filter bank model tested here bears similarities to the texture model of McDermott and Simoncelli [40,82]. The key difference is that dependencies between cochlear frequency channels are captured here by spectral modulation filters rather than the correlations used in the original texture model. In practice, we found that sounds synthesized from the two models were perceptually similar, suggesting that correlations in one stage of representation (the cochlea) can be captured by marginal statistics of a subsequent stage of representation (modulation filters) [40].

Approaches for model testing

Recent years have seen growing interest in the use of computational "encoding models" to test formal theories of sensory processing [5–7,9–14,16–19,71,83]. Because encoding models make quantitative predictions about the neural response, they can be used to test and compare theories of neural coding. The features of the model can then provide insight into the sensory features that are represented in different neural populations [6,7,12,17].

A key challenge of testing encoding models with natural stimuli is that the features of different models are often correlated [17,18], making it difficult to tease apart the unique contribution of any particular model. This problem can be partially overcome by comparing the predictions of two different models but is difficult to eliminate when the features of two models are strongly correlated and when responses can only be measured to a relatively small number of stimuli (as is common with fMRI). Another approach is to alter stimuli so as to decouple different features sets [18,71]. For example, adding varied background noise to natural sounds could help to decouple low- and high-level features of sounds, because noise can alter a sound's low-level features without affecting its perceived identity. However, such approaches are heuristic and do not guarantee that the relevant features will be decorrelated unless the candidate feature sets can be measured with existing models. Model matching is appealing because it provides a way to test the ability of a single model to explain neural responses by imposing the structure of that model alone, decoupling the model from alternative models without needing to specify the many possible alternatives.

Materials and methods

Ethics statement

The study was approved by MIT's Committee on the Use of Humans as Experimental Subjects (protocol 1012004218). All subjects gave written informed consent. The experiments adhere to the Declaration of Helsinki.

Participants

The experiment comprised 41 scanning sessions, each approximately 2 hours. Fifteen subjects participated in the experiment (ages 19–36; five male; all right-handed; one subject, S1, was author SNH). Two different experiment paradigms were tested (hereafter referred to as Paradigm I and Paradigm II). We have chosen to describe these two paradigms as a part of the same experiment because the stimuli and analyses were very similar. In Paradigm I, eight subjects completed a single scanning session, three subjects completed five sessions, and one subject completed three sessions (this subject chose not to return for the fourth and fifth sessions). We chose this approach because it allowed us to compute reliable group maps by averaging across the 12 subjects, as well as reliable individual subject maps using a larger amount of data from the subjects with multiple scan sessions. Five subjects were scanned in Paradigm II. One subject completed two sessions, two subjects completed three sessions, and one subject completed four sessions. One subject (S1) was scanned in both paradigms (when possible we used data from Paradigm II for this subject, because there was a higher quantity of data, and the scan sessions for Paradigm II were higher resolution, as noted below).

Because we aimed to characterize auditory cortex of typical listeners without extensive musical experience, we required that subjects not have received formal musical training in the 5 years preceding their participation in the experiment.

Data acquisition parameters and preprocessing

Data for Paradigm I 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). The functional volumes were designed to provide good spatial resolution in auditory cortex. Each functional volume (i.e., a single 3D image for one participant) included 15 slices oriented parallel to the superior temporal plane and covering the portion of the temporal lobe superior to and including the superior temporal sulcus (3.4 second TR, 30 millisecond TE, 90-degree flip angle; five discarded initial acquisitions). Each slice was 4 mm thick and had an in-plane resolution of 2.1 × 2.1 mm (96 × 96 matrix, 0.4-mm slice gap). iPAT was used to minimize acquisition time (1 second/volume). T1-weighted anatomical images were also collected for each subject (1 mm isotropic voxels).

Data for Paradigm II were collected more recently using a 3T Prisma scanner (also at the McGovern Institute). We used a multiband acquisition sequence (3× acceleration) to reduce slice thickness, while maintaining coverage (36 slices with 2-mm thickness and no gap) and thus reducing voxel size (2 mm isotropic). iPAT was not used. Other acquisition parameters were similar (3.45 second TR, 1.05-second acquisition time, 34 millisecond TE, 90-degree flip angle; three discarded initial acquisitions).

Functional volumes were preprocessed using FSL software and custom MATLAB scripts. Volumes were motion corrected, slice-time corrected, skull stripped, linearly detrended, and aligned to the anatomical volumes (using FLIRT [84] and BBRegister [85]). Volume data were then resampled to the reconstructed cortical surface, computed by FreeSurfer [86], and smoothed on the surface using a 5-mm FWHM kernel to improve SNR (results were similar without smoothing; S3 Fig). Individual subject data were then aligned on the cortical surface to the FsAverage template brain distributed by Freesurfer.

Stimulus presentation and scanning procedure

Our stimulus set was derived from 36 natural sounds, each 10 seconds in duration (Fig 2B). From each natural sound, we synthesized four model-matched sounds, constrained by different subsets of features from a commonly used spectrotemporal filter bank model [20]. The complete stimulus set thus included five conditions (natural sounds + 4 model-matched versions), each with 36 sounds, yielding a total of 180 stimuli.

Scan acquisitions produce a loud noise due to rapid gradient switching. To prevent these noises from interfering with subjects' ability to hear the sounds, we used a "sparse" scanning paradigm [87] that alternated between presenting sounds and acquiring scans, similar to those used in our prior experiments [45,88,89] (S2 Fig). This was achieved by dividing each 10-second stimulus into five 2-second segments (windowed with 25-millisecond linear ramps). These five segments were presented sequentially with a single scan acquired after each segment. The five segments for a particular sound were always presented together in a "block" (the order of the segments within a block was random). Each scan acquisition lasted 1 second in Paradigm I and 1.05 seconds in Paradigm II. There was a 200-millisecond buffer of silence before and after each acquisition. The total duration of each five-segment block was 17 seconds in Paradigm I and 17.25 seconds in Paradigm II. We averaged the responses of the second through fifth acquisitions after the onset of each stimulus block. The first acquisition was discarded to account for the hemodynamic delay. Results were similar when we instead averaged just the second and third time point or just the fourth and fifth time point after stimulus onset, indicating that our results were robust to the averaging window applied to the fMRI measurements (S13 Fig). We chose to use signal averaging rather than a GLM with a standard hemodynamic response function (HRF), because we have found this approach leads to slightly more reliable responses, presumably due to inaccuracies in the standard HRF [90].

In Paradigm I, each model-matched stimulus was presented once per 2-hour scanning session, and the natural stimuli were presented twice so that we could measure the reliability of each voxel's response to natural sounds and noise-correct the NSE metric. Each session was divided into 12 "runs," after which subjects were given a short break (approximately 30 seconds). Each run included six natural sounds and 12 model-matched sounds (three per condition). In Paradigm II, we presented only the model-matched sounds constrained by the complete model, which allowed us to present both the natural and model-matched sounds several times per scan session. Each run included nine natural and nine model-matched sounds. The entire sound set was presented over four consecutive runs. Subjects completed 12 or 16 runs depending on the time constraints of the scan session. Thus, each subject heard each sound between three and four times per session. In both paradigms, there were periods during which no stimulus was presented and only scanner noise was heard, which provided a baseline with which to compare stimulus-driven responses. There were four such "silence" periods per run (each 17 seconds in Paradigm I and 17.25 seconds in Paradigm II). The ordering of stimuli and silence periods was pseudorandom and was designed such that, on average, each condition occurred with roughly the same frequency at each position in a run, and each condition was preceded equally often by every other condition (as in our prior work [88,89]).

Prior to settling on the procedure for Paradigm I, we conducted a pilot experiment in which six of the twelve participants from Paradigm I completed a single session. These sessions featured stimuli from only three of the model-matched conditions (spectral modulation matched stimuli were omitted). These scan sessions were the first of this study, and we limited the number of conditions to make sure the experiment could fit within the allotted 2-hour scanning slot. The runs for these sessions were slightly shorter because there were only nine model-matched stimuli presented per run (there were only three periods of silence per run for these sessions). When analyzing the results, we included the data from these sessions in order to use the maximum amount of data available for each condition, and thus the results for the spectral modulation matched condition were based on less data than the other model-matched conditions. However, because the NSE metric was corrected for noise (see below), differences in the amounts of data across conditions should not bias the results.

Selection of natural stimuli

We used long sounds (10 seconds) so that we could compute time-averaged statistics for filters with relatively long integration periods (i.e., periods of up to 2 seconds). We selected sounds that were likely to produce high response variance in auditory cortical voxels, guided by the results of a prior paper from our lab that measured fMRI responses in auditory cortex to a large set of natural sounds [45]. In our prior study, we found that much of the voxel response variance could be captured by a weighted sum of six response patterns ("components"), and we thus attempted to select sounds that had high response variance along these components. To accomplish this goal, we created a subset of 60 sounds with high component response variance by iteratively discarding sounds in a greedy manner, each time removing the sound that led to the largest increase in response variance, averaged across the six components. Because we needed stimuli that were relatively long in duration, we could not directly use the stimuli from our prior study, which were only 2 seconds in duration. Instead, we created a new stimulus set with 10-second sounds, each of which had the same label (e.g., "finger tapping") as one of the sounds from the 60-sound set.

Model representation

We synthesized sounds based on four different model representations. The simplest model was just based on the output of filters designed to mimic cochlear responses (i.e., a cochleagram). The other three models were based on filters tuned to modulations in this cochleagram representation. Two models were tuned to either temporal modulation or spectral modulation alone, and one was jointly tuned to both temporal and spectral modulation. MATLAB code for measuring and synthesizing sounds from the models described in this paper is available here: https://github.com/snormanhaignere/spectrotemporal-synthesis-v2.

We refer to specific scripts in this repository to clarify our descriptions and ensure that others can replicate our work.

The cochlear representation was computed by convolving the audio waveform of each sound with 120 bandpass filters, spaced equally on an $\mathrm{ERB_N}$-scale between 20 Hz and 10 kHz, with bandwidths chosen to match those measured psychophysically in humans (individual filters had frequency responses that were a half-cycle of the cosine function, in order to exactly tile the frequency spectrum; adjacent filters overlapped by 87.5%) [40] (see wav2-coch_without_filts.m). Each channel was intended to model the response of a different point along the basilar membrane. The envelopes of each filter output were computed using the Hilbert transform, raised to the 0.3 power to mimic cochlear compression/amplification, and downsampled to 400 Hz after applying an anti-aliasing filter. So that we could express the spectral modulation filters that operate on the cochleagram (described below) in units of cycles per octave (as in the original model of Chi and colleagues, 2005 [20]), we interpolated the frequency axis from an ERB-scale to a logarithmic frequency scale (24 cycles/octave), yielding 217 channels.

The modulation-based representations were computed using a bank of multiscale wavelet filters (Fig 1A) that were convolved in time and/or frequency with the cochleagram for each sound (see coch2filtcoch.m). The shapes and bandwidths of the filters were the same as those described by Chi and colleagues (2005). The three sets of filters differed in whether they were tuned to modulation in time, frequency, or both.

The temporal modulation representation was computed using gammatone filters (see filt_temp_mod.m):

$$\psi(t; b_r) = (b_r t)^2 e^{-3.5 b_r t} \sin(2\pi b_r t) \tag{3}$$

where b_r determines the best modulation rate of the filter (i.e., the rate with maximum gain). We used nine filters with octave-spaced best rates: 0.5, 1, 2, 4, 8, 16, 32, 64, and 128 Hz. Each filter was separately convolved in time with each frequency channel of the cochleagram. The output of the model can thus be represented as a set of nine filtered cochleagrams, each of which highlights modulations at a particular temporal rate.

The spectral modulation representation was computed using "Mexican hat" filters, which are proportional to the second derivative of a Gaussian (see filt_spec_mod.m):

$$\phi(f; b_s) = (1 - 2(b_s \pi f)^2) e^{-(b_s \pi f)^2} \tag{4}$$

where b_s determines the best modulation scale of the filter (i.e., the scale with maximum gain). The spectral filters were implemented in the frequency domain using the Fourier representation of a Mexican hat filter:

$$\Phi(\omega; b_s) = w^2 e^{-(w/b_s)^2} \tag{5}$$

We used six filters with octave-spaced scales: 0.25, 0.5, 1, 2, 4, and 8 cycles/octave. Each filter was separately convolved in frequency with each time "slice" of the cochleagram. The output of the model can thus be represented as six filtered cochleagrams, each of which highlights a different range of spectral modulations. Each temporal and spectral filter was scaled so that the power of its best rate/scale was the same for all filters.

The spectrotemporal modulation representation (often referred to as the "full model") was computed primarily from 2D filters that were convolved with the cochleagram in both time and frequency. The filters were instantiated in the 2D Fourier domain (as in the original implementation of Chi and colleagues, 2005 [20]) by taking the outer product of the frequency-domain representations of the temporal and spectral modulation filters described above (see filt_spectemp_mod.m). These filters were then "oriented" so as to be sensitive to upward-right or downward-right modulations. This was accomplished by zeroing either the first and third quadrant of the 2D frequency response (for upward-oriented filters) or the second and fourth quadrant (for downward-oriented filters) (the Nyquist frequency and DC were never zeroed). There were 108 total spectrotemporal filters produced by crossing nine temporal filters with six spectral filters (with best modulation frequencies as described above), and orienting each filter upwards or downwards. Thus, the output of this portion of the model can be represented by 108 filtered cochleagrams (modulo the additional filters described next).

For all three modulation-based representations (temporal, spectral, and spectrotemporal), we included the unfiltered cochleagrams in the representation so that the modulation-based representations would be strictly more expressive. For both the temporal and spectral representations, we also included a filter with power at only the DC (0 Hz or 0 cycles/octaves, respectively). These filters capture the mean of each cochlear frequency channel (for the temporal modulation representation) or the mean of each time slice through the cochleagram (for the spectral modulation representation), and were necessary to reconstruct cochleagrams from the model representation (because all of the other filters were bandpass, with zero power at the DC). For the spectrotemporal modulation representation, the temporal and spectral DC filters were also crossed with the other filters, yielding an additional 15 filters; these filters capture spectrally broadband temporal modulations (i.e., "vertical" modulations) or temporally uniform spectral modulations (i.e., "horizontal" modulations) and have only one orientation. We also added all of the filters from the temporal-only and spectral-only modulation models to the spectrotemporal modulation model so that it would be strictly more expressive than the simpler models.

Finally, two filters that were modulated only in time, and which had very low best-modulation rates (0.125 and 0.25 Hz) were added to the temporal and spectrotemporal modulation representations. These filters were included to replicate the homogeneity of the natural sounds in the model-matched sounds and to improve convergence. Without them, the synthesis process tended to "clump" sound energy at particular time points. The low-rate filters ameliorated this problem by forcing the slow fluctuations in the model-matched sounds to be similar to those in the natural sounds they were matched to.

Model-matching synthesis algorithm

Our model-matching approach, like most algorithms for texture synthesis, starts with a sample of noise, which initially lacks structure, and alters the noise via an iterative procedure to match statistical constraints [40,48,49]—in our case, provided by the histogram of each feature's response (S1 Fig). By initializing with noise, we aim to arrive at a sound that is minimally structured given the imposed constraints.

The model-matching synthesis procedure was initialized with a 10-second sample of Gaussian noise (the same duration as the natural sounds). The algorithm involved three steps (see run_spectrotemporal_synthesis.m): (1) computing the response of each feature from a given model to a natural and noise sound, (2) separately matching the response histogram across time for each model feature [48], and (3) reconstructing a waveform from the modified outputs. These three steps were applied iteratively for reasons described below. For the cochlear representation, we matched the histogram of envelope values for each cochlear frequency channel (see match_coch_hists.m). For the modulation-based representations, we matched the histogram of each frequency channel of each of the filtered cochleagrams (each channel of the filtered cochleagrams represents the output of a single model feature; see match_filtcoch.m), as well as the histograms of the unfiltered cochleagram frequency channels.

The goal of our histogram matching procedure was to modify the distribution of values for one time series so that it had the same distribution of values as that of a target time series, without imposing the same temporal pattern over time. For example, to modify the time series [1 2 3] to match the histogram of [5 1 3], we would like to alter the first time series to be [1 3 5], such that it has the same distribution as the target, but the same relative ordering as the original (smallest, middle, largest). Because the average value of a signal only depends on the distribution of magnitudes and not their ordering, the histogram-matched signals will have the same average value, even if they are transformed by a point-wise function (e.g., $1^2 + 3^2 + 5^2 = 5^2 + 1^2 + 3^2$). Assuming the two time series are represented as vectors of equal length, as was the case for our experiments (because the synthetics were of equal duration), we can histogram match the signals by reassigning the smallest value in the signal to be matched to the smallest value in the target signal, then reassigning the second smallest value in the signal to be matched to the second smallest value in the target, and so on. We can implement this procedure with the following pseudocode:

```
order_original = sortindex(original)
order_target = sortindex(target)
matched[order_original] = target[order_target]
```

where sortindex is a function that takes a vector as input and returns a list of indices into that vector that have been ordered according to the magnitude of the corresponding vector elements (i.e., the indices that would sort the vector). This procedure is a slightly simpler variant of the histogram-matching algorithm described by Heeger and Bergen (1995) and is applicable when matching vectors of equal length.

The details of the reconstruction algorithms have been described previously [20,40]. We reconstruct a waveform from a cochleagram by summing the individual subbands, which are computed by multiplying the envelopes of each cochlear channel (after histogram matching) by their time-varying phases from the previous iteration and then refiltering with the filters used to generate the subbands (as is standard for subband transforms) (see coch2wav_without_filts.m). Similarly, we reconstruct a cochleagram from the modulation domain by adding up the filtered cochleagrams in the 2D Fourier domain and multiplying each cochleagram by the complex conjugate of the filter (to undo phase shifts; see match_filtcoch.m and filtcoch2coch.m). We then divide by the summed power of the modulation filters to correct for the fact that the summed power of the filters is not uniform [20].

In detail, each iteration of the synthesis procedure involved the following steps: (1) compute a cochleagram from the current waveform (2) filter the cochleagram (3) match histograms of the filtered cochleagrams (to those of the target natural sound) (4) reconstruct a cochleagram from the modified filtered cochleagrams (5) match the histograms of the reconstructed cochleagram (6) reconstruct a waveform from the modified cochleagram. For the very first iteration, we matched the histograms of the cochleagrams both before and after matching the histograms of the filtered cochleagrams (rather than just matching after), so that each frequency channel would have approximately the right variance before attempting to match the more detailed modulation properties.

Because the filters whose outputs are being manipulated overlap in the frequency domain, a manipulation such as histogram matching typically introduces inconsistencies between filters, such that when the reconstructed signal is reanalyzed the histograms will generally not remain matched. As such, a single iteration of the matching procedure does not achieve its objective, but iterating the procedure described above generally results in increasingly close matches [40,48,49]. We monitored convergence by measuring the difference in the desired and measured histograms at each iteration (see below) and used 100 iterations, which we found to produce good results.

To avoid wraparound effects due to circular convolution, we padded the cochleagrams in time and frequency prior to convolution with the modulation filters. We padded the cochleagrams with a value equal to the global mean of the cochleagram across both time and frequency so as to minimize the resulting step edge. The amount of padding was chosen to be approximately equal to the duration of ringing in each filter: we padded the cochleagrams in frequency by twice the period of the coarsest spectral modulation filter (eight octaves of padding) and by three times the period of the slowest temporal modulation filter (24 seconds of padding). To ensure that the portion of the signal used for the stimulus was well matched, we applied the histogram-matching procedure twice at each iteration, once to the entire signal, including the padded duration, and once to just the non-padded portion of the signal (in that order).

Assessing the success of the model-matching algorithm

For each model feature, we computed a time-averaged measure of its response amplitude for natural and model-matched sounds. S14 Fig plots these amplitude statistics for example natural and model-matched sounds. For cochlear features, we simply averaged the cochleagram envelope amplitudes across time. For the modulation-tuned features, we computed the standard deviation across time of each feature's response. We then correlated the filter amplitudes for corresponding natural and model-matched sounds across all filters in the model, as a measure of their similarity (S14 Fig, right panel). The mean correlation across sounds was high for all of the model features being matched by the synthesis algorithm ($r^2 > 0.98$), and much higher than the correlation observed for features not constrained by the matching algorithm.

Variants of the spectrotemporal filter model

We investigated the extent to which our results might depend on the particular choice of spectrotemporal filters tested (S15 Fig). Specifically, we created spectrotemporal filters with different properties by either (1) randomizing the temporal and spectral phase to create a diverse range of filter shapes with roughly the same modulation spectrum, (2) halving the filter bandwidths, or (3) randomizing the filters entirely by sampling the filter weights from a Gaussian. Phase randomization was implemented by computing the FFT of each filter's temporal and spectral impulse response (using a window size of twice the period of the filter's center modulation rate/scale), randomizing the phase, transforming back to the signal domain (via the iFFT), and padding with zeros (these variations are implemented in filt_temp_mod.m and filt_spec_mod.m). Narrowing the filter bandwidths was accomplished by doubling the extent of the gammatone envelope, while leaving the carrier frequency unchanged (in the equation below, we set λ_t to 1 for the standard model used by Chi and colleagues [20] and to 0.5 for the half-bandwidth model):

$$\psi(t; b_r, \lambda_t) = (\lambda_t b_r t)^2 e^{-3.5\lambda_t b_r t} \sin(2\pi b_r t) \qquad (6)$$

For the spectral filters, we used a Morlet wavelet, which is similar to the Mexican hat wavelet used by Chi et al. [20], but has a variable bandwidth again determined by the extent of the filter's envelope (we set λ_s to 0.5 for the half-bandwidth model; $\lambda_s = 1$ results in a filter similar to the Mexican hat filter used by Chi and colleagues [20]):

$$\phi(f; b_s, \lambda_s) = e^{-(2\lambda_s b_s f)^2} \cos 2\pi b_s f \tag{7}$$

For the random filters, we varied the size of the filters to mimic the fact that the model filters vary in the amount of time and frequency over which they integrate.

For each filter, we measured the amplitude (standard deviation) of its response to each of the natural and model-matched sounds that we tested in the fMRI experiment (middle panels of S15 Fig), which were constrained only by the original spectrotemporal filters and not the modified variants. We then correlated the filter's amplitude for corresponding natural and model-matched sounds (across the bank of filters for each sound) to assess how well the natural and model-matched sounds were matched (rightmost panel of S15 Fig). For the phase-randomized and half-bandwidth filters, we found that matching the spectrotemporal statistics of the original filters substantially improved how well the modified spectrotemporal filters were matched (median $r^2 > 0.85$), suggesting that matching the statistics of one spectrotemporal model goes a long way toward matching the statistics of other modulation filter models. This result suggests that our findings will generalize to other spectrotemporal filters with different shapes. For random filters, we found that the filter variances were relatively well matched even for sounds that were not matched on the original spectrotemporal filters (median $r^2 = 0.73$ for cochlear-matched sounds), suggesting that random filter variances may be easier to match than the more structured filters in the model.

Envelope-based synthesis algorithm

The stimuli used for the first six pilot scan sessions were synthesized using a slightly different algorithm that was based on matching the histogram of the envelopes of the modulation filter outputs rather than matching the histogram of the raw filter responses. In practice, we found that histogram matching the envelopes produced very similar results to matching the histogram of the raw outputs and thus decided to use the simpler algorithm for the remaining scanning sessions. The voxel responses to stimuli synthesized from the two algorithms was similar, and we thus collapsed across all of the available data for all analyses.

Normalized squared error

We measured the similarity of fMRI responses to natural and model-matched sounds via the mean squared error:

$$\mu\left([\boldsymbol{x} - \boldsymbol{y}]^2\right) = \frac{1}{N}\sum\nolimits_{i=1}^{N} (x_i - y_i)^2 \tag{8}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ represent the vector of responses to natural and model-matched sounds, respectively (here, $N = 36$ because there were 36 natural/model-matched sounds). We normalized the mean squared error so that it would be invariant to the overall scale of the voxel responses and take a value of 0 if the response to natural and model-matched sounds was identical, and 1 if there was no correspondence between responses to natural and model-matched sounds (i.e., if they were independent of each other):

$$\frac{\mu([\boldsymbol{x}-\boldsymbol{y}]^2)}{\mu(\boldsymbol{x}^2)+\mu(\boldsymbol{y}^2)-2\mu(\boldsymbol{x})\mu(\boldsymbol{y})} \tag{9}$$

We refer to this metric as the normalized squared error or NSE. The quantity in the denominator is an estimate of the expected value of the squared error, assuming the two variables are independent:

$$\begin{aligned}\mathbb{E}[(x-y)^2] &= \mathbb{E}[x^2+y^2-2xy]\\ &= \mathbb{E}[x^2]+\mathbb{E}[y^2]-2\mathbb{E}[xy]\\ &= \mathbb{E}[x^2]+\mathbb{E}[y^2]-2\mathbb{E}[x]\mathbb{E}[y]\end{aligned} \tag{10}$$

Noise-correcting the NSE

Model matching makes it possible to falsify a model by showing that neural responses to natural and model-matched stimuli diverge. However, fMRI responses are noisy and thus, even if the true responses to natural and model-matched sounds are identical, the measured fMRI responses will differ somewhat. To account for this fact, we noise-corrected our NSE metric to provide an estimate for what the true NSE would be in the absence of noise and to ensure that differences between regions cannot be explained by differences in voxel reliability. By bootstrapping the noise-corrected NSE, one can estimate a distribution over the true NSE values between natural and model-matched sounds, which can be used to perform statistics (see "Statistics" below). In practice, we observed similar trends with and without correction because voxel responses in both primary and nonprimary regions were similarly reliable (S4 Fig). MATLAB code implementing the noise-correction procedures described below can be downloaded here: https://github.com/snormanhaignere/general-analysis-code. see noise_corrected_similarity.m.

Most noise-correction methods assume that the noise-corrupted response reflects the sum of a noise-free stimulus-driven signal plus noise that is statistically independent of the stimulus-driven signal:

$$\boldsymbol{x} = \boldsymbol{s}_x + \boldsymbol{n}_x \tag{11}$$

$$\boldsymbol{y} = \boldsymbol{s}_y + \boldsymbol{n}_y \tag{12}$$

where in the context of this experiment, $\boldsymbol{x}$ and $\boldsymbol{y}$ are the measured response of a voxel to two sets of sounds (i.e., natural and model-matched sounds), $\boldsymbol{s}_x$ and $\boldsymbol{s}_y$ are the stimulus-driven responses, and $\boldsymbol{n}_x$ and $\boldsymbol{n}_y$ are the noise that contributes to the response measurements. All noise-correction methods require at least two repetitions of the same stimulus so that the effects of the noise can be disentangled from the effects of the stimulus-driven signal. By assumption, these two repetitions only differ in their noise:

$$\boldsymbol{x}_1 = \boldsymbol{s}_x + \boldsymbol{n}_{x1} \tag{13}$$

$$\boldsymbol{x}_2 = \boldsymbol{s}_x + \boldsymbol{n}_{x2} \tag{14}$$

$$\boldsymbol{y}_1 = \boldsymbol{s}_y + \boldsymbol{n}_{y1} \tag{15}$$

$$\boldsymbol{y}_2 = \boldsymbol{s}_y + \boldsymbol{n}_{y2} \tag{16}$$

We would like to estimate the NSE of the stimulus-driven responses, uncorrupted by noise:

$$\begin{aligned} \epsilon(\boldsymbol{s}_x, \boldsymbol{s}_y) &= \frac{\mu([\boldsymbol{s}_x - \boldsymbol{s}_y]^2)}{\mu(\boldsymbol{s}_x^2) + \mu(\boldsymbol{s}_y^2) - 2\mu(\boldsymbol{s}_x)\mu(\boldsymbol{s}_y)} \\ &= \frac{\mu(\boldsymbol{s}_x^2) + \mu(\boldsymbol{s}_y^2) - 2\mu(\boldsymbol{s}_x\boldsymbol{s}_y)}{\mu(\boldsymbol{s}_x^2) + \mu(\boldsymbol{s}_y^2) - 2\mu(\boldsymbol{s}_x)\mu(\boldsymbol{s}_y)} \end{aligned} \tag{17}$$

But we only have available the noise-corrupted responses. From the equation above, it is evident that the NSE depends on three types of statistics: (1) the signal powers ($\mu(\boldsymbol{s}_x^2)$ and $\mu(\boldsymbol{s}_y^2)$), (2) the signal cross-product ($\mu(\boldsymbol{s}_x\boldsymbol{s}_y)$), and (3) the signal means ($\mu(\boldsymbol{s}_x)$ and $\mu(\boldsymbol{s}_y)$). The signal means are unbiased by the noise because, by assumption, the noise is zero mean. The signal cross-product is also unbiased by noise:

$$\begin{aligned} \mathbb{E}[(s_x + n_x)(s_y + n_y)] &= \mathbb{E}[s_x s_y] + \mathbb{E}[s_x n_y] + \mathbb{E}[s_y n_x] + E[n_x n_y] \\ &= \mathbb{E}[s_x s_y] \end{aligned} \tag{18}$$

(we have replaced means with expectations to indicate a theoretical average over infinitely many samples, for which the bias is exactly zero). We thus estimate the signal cross-product and means using the measured cross-product and means of the data without correction:

$$\hat{\mu}\left(\boldsymbol{s}_x\boldsymbol{s}_y\right) = \frac{1}{4}\mu(\boldsymbol{x}_1\boldsymbol{y}_1) + \frac{1}{4}\mu(\boldsymbol{x}_1\boldsymbol{y}_2) + \frac{1}{4}\mu(\boldsymbol{x}_2\boldsymbol{y}_1) + \frac{1}{4}\mu(\boldsymbol{x}_2\boldsymbol{y}_2) \tag{19}$$

$$\hat{\mu}(\boldsymbol{s}_x) = \frac{1}{2}\mu(\boldsymbol{x}_1) + \frac{1}{2}\mu(\boldsymbol{x}_2) \tag{20}$$

$$\hat{\mu}\left(\boldsymbol{s}_y\right) = \frac{1}{2}\mu(\boldsymbol{y}_1) + \frac{1}{2}\mu(\boldsymbol{y}_2) \tag{21}$$

Unlike the mean and the cross-product, the signal power is biased upwards by the noise:

$$\begin{aligned} \mathbb{E}[(s_x + n_x)^2] &= \mathbb{E}[s_x^2] + \mathbb{E}[n_x^2] + 2\mathbb{E}[s_x n_x] \\ &= \mathbb{E}[s_x^2] + \mathbb{E}[n_x^2] \end{aligned} \tag{22}$$

The magnitude of this bias can be estimated using the residual error between two measurements of the same stimulus, which by definition is due exclusively to noise. The expected power of the residual is equal to twice the noise power:

$$\begin{aligned} \mathbb{E}[(x_1 - x_2)^2] &= \mathbb{E}[([s_x + n_{x1}] - [s_x + n_{x2}])^2] \\ &= \mathbb{E}[(n_{x1} - n_{x2})^2] \\ &= \mathbb{E}[n_{x1}^2] + \mathbb{E}[n_{x2}^2] \\ &= 2\mathbb{E}[n_x^2] \end{aligned} \tag{23}$$

Thus, we can estimate the signal power by subtracting off half the residual power from the average power of the noise-corrupted data:

$$\hat{\mu}\left(\boldsymbol{s}_x^2\right) = \frac{1}{2}\mu\left(\boldsymbol{x}_1^2\right) + \frac{1}{2}\mu\left(\boldsymbol{x}_2^2\right) - \frac{1}{2}\mu\left([\boldsymbol{x}_1 - \boldsymbol{x}_2]^2\right) \tag{24}$$

$$\hat{\mu}\left(\boldsymbol{s}_y^2\right) = \frac{1}{2}\mu\left(\boldsymbol{y}_1^2\right) + \frac{1}{2}\mu\left(\boldsymbol{y}_2^2\right) - \frac{1}{2}\mu\left([\boldsymbol{y}_1 - \boldsymbol{y}_2]^2\right) \tag{25}$$

Substituting Eqs 19–21, 24 and 25 into Eq 17 yields the noise-corrected NSE. The noise-corrected NSE, like the raw NSE, is invariant to the overall scale of the data.

Noise-correction requires two independent samples of the same stimulus. In our case, each sample was itself an average across multiple stimulus blocks, and for each stimulus block, we averaged responses across the last four scan acquisitions within the block. Thus, each sample was based on many scan acquisitions (between 12 and 28 acquisitions for individual subject maps, corresponding to between 3 and 7 stimulus block repetitions; group maps were based on 104 scan acquisitions per measurement). In Paradigm I, each natural sound was repeated once per scan while the model-matched sounds were only presented once. We chose this design so that we could present model-matched sounds constrained by different subsets of model features, which would have been infeasible if each model-matched sound was presented twice. To noise-correct the responses, we made the simplifying assumption that the noise power was equal for natural and model-matched sounds, and estimated the noise power from responses to the natural sounds (when multiple scan sessions were available, we first averaged responses across scan sessions). This assumption is natural given that the noise by definition reflects the component of the signal that is not driven by the stimulus. Nonetheless, we tested whether this assumption is appropriate using the data for Paradigm II, in which we repeated responses to both natural and model-matched sounds. In one case, we assumed that the noise power was the same, and calculated the noise power using only the responses to natural sounds. In the other case, we separately calculated the noise power for natural and model-matched sounds. The results were very similar using the two approaches (S16 Fig), which validates the assumption that the noise power is similar for natural and model-matched sounds.

Our noise correction procedure assumes that the noise is uncorrelated across measurements (this assumption was used in Eqs 18 and 22), which is the not the case for fMRI measurements close in time (i.e., <5 seconds) [91]. Here, each measurement corresponds to the average response of the second through fifth scan acquisition after the onset of each stimulus block. Blocks for the same stimulus were never repeated back to back, and even if they were, the two blocks would have been separated by 6.8 seconds, which is longer than the typical autocorrelation of the BOLD signal [91]. In Paradigm II, the same stimuli were never repeated within a run. Thus, it is unlikely that the autocorrelation of the BOLD signal impacted our measures.

Evaluating the noise-corrected NSE with simulated data

Noise-correction inevitably increases the variance of the statistics being corrected, and thus it is critical to have sufficiently reliable responses (which is why we collected a relatively large amount of data for this study). To assess the reliability needed to perform correction, we performed a simulation in which we generated a large number of noisy voxel responses. We based our simulations on Paradigm I, in which only the natural sounds were repeated, but results were similar for simulations that mimicked Paradigm II, in which both natural and model-matched sounds were repeated. For Paradigm I, we had three 36-dimensional response vectors per voxel: two vectors for the 36 natural sounds, which were each presented twice per scan session, and one for the model-matched sounds, which were each presented once per scan session. We thus simulated three 36-dimensional response vectors ($\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{y}_1$) for each voxel ($\boldsymbol{x}_1$, and $\boldsymbol{x}_2$, corresponding to the voxel's response to natural sounds, and $\boldsymbol{y}_1$ to the response to model-matched sounds). Each vector was computed as the weighted combination of a true, noise-free signal ($\boldsymbol{s}_x$, $\boldsymbol{s}_y$) that was constant across repeated measurements plus additive noise that varied across measurements ($\boldsymbol{n}_{x1}$, $\boldsymbol{n}_{x2}$, $\boldsymbol{n}_{y1}$):

$$\boldsymbol{x}_1 = \boldsymbol{s}_x b + \boldsymbol{n}_{x1}(1-b) \tag{26}$$

$$\boldsymbol{x}_2 = \boldsymbol{s}_x b + \boldsymbol{n}_{x2}(1-b) \tag{27}$$

$$\boldsymbol{y}_1 = \boldsymbol{s}_y b + \boldsymbol{n}_{y1}(1-b) \tag{28}$$

We used the weights (b) to control the SNR of the voxel with weights closer to 1 resulting in higher SNR. We sampled b from a uniform distribution between 0 and 1. We sampled the three noise vectors from a zero-mean, unit-variance Gaussian distribution. Our noise-correction algorithms assume that the noise variance is the same for the natural and model-matched sounds (var($\boldsymbol{n}_x$) = var($\boldsymbol{n}_y$)), which we have verified is a reasonable assumption for our data (S16 Fig). We also assume that the noise samples are independent from each other, which we would expect to be the case given that our measurements were spaced far apart in time relative to the autocorrelation of the BOLD signal [91]. Our noise-correction algorithm makes no assumptions about the distribution of errors. Here, we use a Gaussian distribution for simplicity, but results were similar using other noise distributions (e.g., Laplace).

We sampled the true noise-free signals ($\boldsymbol{s}_x$ and $\boldsymbol{s}_y$) in a way that allowed us to vary how similar they were. We did this in two different ways (referred to hereafter as Simulation 1 and Simulation 2). In Simulation 1, we computed $\boldsymbol{s}_x$ and $\boldsymbol{s}_y$ as the weighted sum of a shared response vector ($\boldsymbol{g}$) and a distinct vector unique to x and y ($\boldsymbol{u}_x$, $\boldsymbol{u}_y$):

$$\boldsymbol{s}_x = \boldsymbol{g}c + \boldsymbol{u}_x(1-c) \tag{29}$$

$$\boldsymbol{s}_y = \boldsymbol{g}c + \boldsymbol{u}_y(1-c) \tag{30}$$

By varying c, we manipulated the similarity of the noise-free signals. We sampled c from a uniform distribution, and we sampled $\boldsymbol{g}$, $\boldsymbol{u}_x$, and $\boldsymbol{u}_y$ from a zero-mean, unit-variance Gaussian. The results were similar using other distributions (e.g., the Gamma distribution). Changing the means of these distributions also had little effect on the results.

In Simulation 2, one of the signal vectors was simply a scaled version of the other, in order to mimic weaker responses to model-matched sounds:

$$\boldsymbol{s}_x = \boldsymbol{g} \tag{31}$$

$$\boldsymbol{s}_y = \boldsymbol{g}c \tag{32}$$

For each type of simulation, we sampled 100,000 voxel responses. For each sample, we computed four statistics:

1. the NSE between the noisy signals (using just $\boldsymbol{x}_1$ and $\boldsymbol{y}_1$ for simplicity)

2. the NSE between the true signals ($\boldsymbol{s}_x$ and $\boldsymbol{s}_y$), which is what we would like to infer

3. our estimate of the true NSE, computed by applying our noise-correction algorithm to the noisy data ($\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{y}_1$)
4. the NSE between two independent measurements of the same stimulus ("test-retest"), which provides a measure of the voxel's noise level ($\boldsymbol{x}_1$, $\boldsymbol{x}_2$)

In S17A and S17B Fig, we plot the results from Simulation 1. First, we plot the NSE of the noise-corrupted data versus the NSE of the true signals (S17A Fig, left column). Each point represents the NSE values for a single simulated voxel response, and the results have been binned by the test-retest NSE values of the noise-corrupted signals (from low to high, going from top to bottom of the figure), which provides a measure of the noise level (lower test-retest NSEs corresponding to less noise). Unsurprisingly, as the noise increases, the upwards bias caused by the noise increases. Next, we plot the noise-corrected NSE values versus the NSE values for the true signals (S17A Fig, right column). As expected, noise-correction removes the bias caused by the noise, at the expense of increasing the variance. These effects are quantified in S17B Fig, which plots the median NSE of both the noise-corrupted and noise-corrected values, along with the standard deviation (central 68% of the sampling distribution). At high noise levels (test-retest NSE > 0.4), noise-correction substantially increases the standard deviation of the samples, which makes correction untenable. But for low noise levels (test-retest NSE < 0.4), the method corrects the bias without substantially increasing the standard deviation of the sampling distribution. The results are similar for Simulation 2 (S17C and S17D Fig): at low noise levels (test-retest NSE < 0.4), noise-correction corrects the bias introduced by noise while only modestly increasing the standard deviation. We limited our analyses to voxels with a test-retest NSE of less than 0.4, thus remaining in the regime in which noise-correction is well-behaved. In Paradigm I, we measured reliability using natural sounds, because the model-matched sounds were not repeated. For Paradigm II, we concatenated responses to natural and model-matched sounds and measured the test-retest NSE of the resulting vector.

To directly test whether a test-retest NSE less than 0.4 is sufficient to ensure reliable measures, we measured the consistency of our noise-corrected measures across different subsets of data. Noise-correction requires two independent splits of data, and thus to test the reliability of noise-corrected NSE measures, one needs at least four repetitions of each sound set. For Paradigm II, each subject heard between 6 and 15 repetitions of each sound set, which made it possible to perform this analysis. We averaged responses within four separate splits of data, each with an equal number of repetitions (e.g., assuming 12 repetitions, split 1 included repetitions 1, 5, and 9, split 2 included repetitions 2, 6, and 10, and so on). We then calculated the noise-corrected NSE twice based on splits 1 and 2 and splits 3 and 4. We excluded voxels with a test-retest NSE above 0.4 in splits 1 and 2 (because the test-retest NSE was only determined using splits 1 and 2, splits 3 and 4 provide a fully independent validation of the corrected values). This analysis revealed that the noise-corrected measures were reliable (S18 Fig).

Noise-correcting response variation and correlation measures

In addition to the NSE, we also compared responses (in the six response components as well as in individual voxels) to natural and model-matched sounds by comparing their response variation, as measured by the standard deviation, and by correlating their responses (Fig 6E and 6F, and S12 Fig). We noise-corrected these measures as well. The variance of a noise-corrupted signal is biased upwards by the noise in the same manner as the signal power (Eq 22) and thus can be corrected by subtracting off half of the residual power (the noise-corrected standard deviation can be computed by taking the square root of the noise-corrected variance). The correlation coefficient is given by

$$corr(\boldsymbol{x},\boldsymbol{y}) = \frac{cov(\boldsymbol{x},\boldsymbol{y})}{\sqrt{var(\boldsymbol{x})var(\boldsymbol{y})}} \tag{33}$$

The covariance, which is defined as the cross-product of demeaned variables, is unbiased by noise for the same reason that the raw signal cross-product is unbiased by noise (Eq 18), and thus we only need to correct the signal variance by subtracting off half of the residual power. This approach is similar to the more standard correction procedure of dividing by the square root of the test-retest correlation of the measures [45,92] (and in the limit of infinite data the two are equivalent). However, our approach is applicable when the test-retest reliability can only be measured for a single variable (as was the case for Paradigm I).

We note that the correlation between two variables becomes unstable (and in the limit undefined) as the variance of one variable approaches zero, which poses a problem in nonprimary regions, where we observed weak responses to the model-matched sounds. Thus, it was necessary to exclude voxels that did not have a test-retest correlation to model-matched sounds of at least 0.4, which caused many nonprimary voxels to be excluded in the maps of S12 Fig. This is not an issue with the NSE, because the NSE is well defined as long as either of the two variables being compared have nonzero variance.

Model predictions

Our model assumes that voxels are a weighted sum of time-averaged statistics of the feature responses (Eqs 1 and 2). To predict voxel responses, we must choose a specific set of statistics and voxel weights. For the cochlear model, we used the average magnitude of each filter response's envelope across time as our statistic (yielding 217 features, one per cochlear channel). For the three modulation models (temporal, spectral, and spectrotemporal), we used the standard deviation of each feature's response across time as our statistic, which we found gave better predictions than the power (sum of squares) or variance (sum of squares after demeaning) (we suspect this is because squaring the filters leads to a skewed distribution of values that is harder to linearly align with the voxel responses). We also included the 217 cochlear features in the modulation representation to make the analysis parallel to the model-matching procedure (in which all of the modulation-matched sounds were also matched in their cochlear statistics). For the temporal modulation model, there were a total of 2,170 features (9 rates × 217 audio frequencies + 217 cochlear channels). For the spectral modulation model, there were 1,736 features (7 scales × 217 frequencies + 217 cochlear channels). For the spectrotemporal modulation model, there were 27,559 features (9 rates × 7 scales × 2 orientations × 217 frequencies + 217 cochlear channels). We did not include the temporal-only and spectral-only modulation features in the spectrotemporal modulation model because we found this did not improve prediction accuracy. We also excluded the DC filter from the temporal modulation model because it has zero variance across time.

For all of the models tested, we learned the voxel-specific weights across features via ridge regression, as is standard in the evaluation of encoding models [7,17]. Several of the models tested had a large number of features, which could potentially make it difficult to map the model features to the voxel responses. One option would have been to choose a subset of features or reduce dimensionality with PCA before performing regression. However, we have found that such approaches lead to slightly worse predictions than using a large number of features and regularizing with ridge [16]. Prior to regression, all of the features were normalized (z-scored across sounds), and a bias/ones term was added to account for the mean. For models with both cochlear and modulation features, we separately rescaled the two feature sets so that they would have the same norm and thus contribute similarly to the analysis (otherwise, the modulation features would dominate because there were many more features in the modulation representation).

We used cross-validation across sounds to avoid statistical bias in fitting the weights, as well as to select the optimal regularization parameter. First, we split the response of each voxel to the 36 natural sounds into test and train data. The training data were used to fit the weights and select the regularization parameter (details below), and the test data were used to evaluate the predictions of the model. We used 4-fold cross-validation, splitting the data into four equally sized sets of nine sounds. For each set, we used the remaining sounds to the fit the model (i.e., the 27 sounds from the other three sets), and we averaged the accuracy of the predictions across the 4 folds. We quantified the similarity of the measured and predicted responses using the noise-corrected NSE so that the results could be compared with the model-matching results (details of noise correction are given below).

To select the regularization parameter, we split each training set (27 sounds) again into four approximately equally sized sets (7, 7, 7, and 6). For each set, we used the remaining sounds to fit the weights for a large range of regularization parameters (2^{-100} to 2^{100}, with octave steps) and the left out sounds to evaluate the accuracy of the model as a function of the regularization parameter. We then selected the regularization parameter that led to the best generalization accuracy averaged across the four splits (again using the noise-corrected NSE). Finally, given the selected regularization parameter, we fit the weights using all of the training set.

We used the same procedure for the cross-prediction analyses, but instead of training and testing on natural sounds, we learned the voxel weights on the natural sound responses and tested them on model-matched sounds (and vice versa). MATLAB code implementing these regression analyses can be downloaded here: https://github.com/snormanhaignere/general-analysis-code, see: regress_predictions_from_3way_crossval_noisecorr.m.

Noise-correcting model predictions

In the context of model predictions, we want to estimate the ability of the model to predict voxel responses to left-out stimuli in the absence of noise due to fMRI. Because the predictions are derived from noisy fMRI measurements, it is necessary to correct for the reliability of both the data and predictions [16]. Each natural sound was presented twice in the experiment. For each repetition and each test fold, we measured the response of each voxel to the test sounds and computed a prediction from the model using the training sounds (as described above in "Model predictions"). The same training and test sounds were used for both repetitions. This procedure yielded two samples of the voxel response and two samples of the predicted response for each of the four test folds. We used these two samples to compute the necessary statistics for the noise-corrected NSE (Eqs 19–21, 24 and 25). We used our noise-corrected squared error metric to both quantify the accuracy of the predictions and to select the regularization parameter.

Voxel decomposition

Previously, we found that voxel responses to a diverse set of 165 natural sounds could be approximated by a weighted sum of six canonical response patterns (components) [45]:

$$\boldsymbol{v}_i \approx \sum_{k=1}^{6} \boldsymbol{r}_k w_{k,i} \tag{34}$$

where $\boldsymbol{v}_i$ and $\boldsymbol{r}_k$ are 165-dimensional vectors representing the response of voxel i and component k to the sounds tested, and $w_{k,i}$ represents the weight of component k in voxel i. The component responses ($\boldsymbol{r}_k$) and weights ($w_{k,i}$) were jointly inferred by maximizing the non-Gaussianity of the weights, similar to classical independent component analysis [93]. Fig 6A replots a summary map of the weights from our prior study (averaged across subjects and transformed to a measure of statistical significance).

Six of the subjects from the present experiment also participated in our prior study, and two others participated in a similar experiment in which we measured responses to a subset of 30 sounds from the original 165-sound experiment chosen to best identify the six components (by minimizing the variance of the component weights estimated by regression) (all eight subjects were scanned in Paradigm I). Four of these eight subjects were scanned in the earlier version of the model-matching experiment without the spectral modulation condition, and thus the component responses to the spectral-only model-matched sounds were measured in just these four subjects. For each subject, we learned a set of reconstruction weights ($u_{k,i}$) that when applied to the voxel responses from these two prior studies could approximate the component response profiles:

$$\boldsymbol{r}_k \approx \sum_i \boldsymbol{v}_i u_{k,i} \tag{35}$$

We then simply multiplied the voxel responses from the current experiment by the same reconstruction weights to estimate the component responses to the natural and model-matched stimuli from our current study. The reconstruction weights were estimated using ridge regression, picking the regularization parameter that led to the best prediction accuracy for left-out sounds (using the same cross-validation procedure described in the previous section to select the weights and regularization parameter; we used 5-fold cross-validation here). All voxels with a temporal SNR greater than 30 were used (temporal SNR was defined as the mean of the voxel's time course divided by its standard deviation; results were similar when the analysis was restricted to voxels from the superior temporal plane and gyrus). This analysis was performed separately for every subject, and the inferred component responses were then averaged across subjects (this made it possible to use bootstrapping to compute standard errors and significance; see "Statistics" below). We again quantified the similarity of responses to natural and model-matched sounds using the noise-corrected NSE.

We note that an alternative approach would have been to use the pseudoinverse of the encoding weights ($w_{k,i}$ in Eq 34) as our reconstruction weights [45], rather than learning reconstruction weights via ridge regression. We have consistently found the pseudoinverse approach to be less effective than directly learning the reconstruction weights (i.e., the reconstructed profiles more closely match the target response profile in left-out data when the reconstruction weights are directly optimized for the purpose of reconstruction). The approach of learning separate weights for the purpose of reconstruction is standard in the sensory encoding/decoding literature [94].

Annular analyses

We quantified the similarity of responses to natural and model-matched sounds (or model predictions) by binning voxels based on their distance to PAC, defined either tonotopically or anatomically. Voxels were binned in 5-mm intervals, and we computed the median NSE value across the voxels within each bin. Anatomically, we defined PAC as the center of TE1.1, which is located in posteromedial Heschl's gyrus. We relied on surface-based alignment to map the TE1.1 ROI to the appropriate anatomical region (the presence/absence of duplications along HG is reported in S1 Table and was defined by inspection using the scheme described in

Da Costa and colleagues (2011) [52]). Tonotopically, PAC was defined by hand in individual subjects as the center of the low-frequency reversal of the high-low-high gradient within Heschl's gyrus [51–55]. These maps were derived from responses to pure tones presented in six different frequency ranges (with center frequencies of 200, 400, 800, 1,600, 3,200, and 6,400 Hz). We measured the frequency range that produced the maximum response in voxels significantly modulated by frequency ($p < 0.05$ in a one-way ANOVA across the six ranges); the details of the stimuli and analyses have been described previously [88]. Group tonotopy maps were based on a cohort of 21 subjects who were run in this tonotopy localizer across multiple studies (six of the subjects from this experiment were part of this cohort) [95]. The best-frequency maps from each of these 21 subjects were averaged to form group maps. Voxels in which fewer than three subjects had frequency-modulated voxels were excluded from the map.

Statistics

All of our statistical tests, with the exception of the voxel decomposition analysis (Fig 6), were based on the annular analyses described above. We defined primary and nonprimary regions using the three bins nearest and furthest from PAC, defined anatomically or tonotopically. In every subject and hemisphere, we observed an increase in the NSE between primary and nonprimary regions, which was significant using a sign test ($p < 0.01$). The same was true for comparing NSE values derived from model matching and model prediction: in all eight subjects, the increase in NSE values between primary and nonprimary regions was greater for model-matching than for prediction.

For comparing NSE values between different model-matching conditions, we were only able to compute individual subject maps from the four subjects that were scanned multiple times in Paradigm I. All four subjects tested showed the trends evident in the group map (S7 Fig), but the small number of subjects precluded a random effects analysis. We thus performed statistics on group-averaged responses, bootstrapping across all 12 subjects scanned at least once in Paradigm I. Specifically, we sampled 12 subjects with replacement 1,000 times and averaged responses across the 12 sampled subjects in standardized anatomical coordinates. For each sample, we then recomputed the voxel-wise NSE values and binned these values based on distance to PAC. This procedure yielded 1,000 samples of the NSE for each annular bin and condition. For each contrast (e.g., NSE full model < NSE cochlear-matched sounds), we subtracted the average NSE across all bins between the two conditions being compared and counted the fraction of times this contrast fell below zero. Multiplying this fraction by 2 yielded the reported two-sided p-values. Results were similar when we averaged responses within just PAC (the first three bins), where the model performed best.

An obvious downside of statistics based on group-averaged maps is that individual subjects exhibit idiosyncrasies in their anatomy [96]. Component analysis provides one way to overcome this problem by mapping all of the subjects to a common response space, an approach that is relatively common in EEG studies [97] but is less frequently applied to fMRI analyses [98]. We performed statistics on the component responses by estimating the response of each component in each subject and then bootstrapping across the eight subjects from whom we had component data. We randomly sampled eight subjects with replacement 1,000 times, averaged the component responses for those subjects, and recomputed the noise-corrected NSE. To evaluate significance, we then contrasted NSE values between components or conditions and counted the fraction of samples where this difference fell below zero (multiplying this fraction by 2 yielded the reported two-sided p-values).

Supporting information

S1 Fig. Schematic of model-matching approach. (A) The models considered here were defined by the response time course of a set of model features, each computed by filtering a cochleagram representation of sound (illustrated in Fig 1A). Our model-matching algorithm collapses these time courses across time to form a histogram and then generates a sound with the same histograms as a natural sound. Here, we plot example time courses and histograms for three example natural sounds (left panel) and corresponding model-matched sounds (right panel), in this case generated from the full spectrotemporal filter model used throughout this paper. Different natural sounds produce distinct response time courses and histograms. Corresponding natural and model-matched sounds produce similar response histograms but distinct response time courses. (B) The model-matched sounds were synthesized by modifying a noise signal so as to match the histogram of each feature's response to a natural sound. The algorithm was initialized with Gaussian noise that was initially unstructured and thus produced feature responses with a different histogram and a different time-varying response pattern. The noise sound was then iteratively adjusted so as to match the histogram of each feature to the natural sound, while leaving the temporal pattern unconstrained. This figure plots histograms for one example model feature in response to a natural, noise, and model-matched sound. The histogram-matching algorithm is conceptually similar to a classic visual texture synthesis algorithm [48].
(TIF)

S2 Fig. Schematic of the sparse scanning paradigm used to present stimuli in the experiment. Each 10-second stimulus was subdivided into five 2-second segments. These five segments were presented in a random order, with a 1-second scan acquisition interspersed between each presentation (1.05 seconds for Paradigm II). A short 200-millisecond buffer was present between stimuli and scan acquisitions. The total duration of each "block" of five sounds was 17 seconds (17.25 seconds for Paradigm II). The response to a stimulus was computed as the average response of the second through fifth scan acquisition after block onset (the first acquisition was discarded to account for the hemodynamic lag).
(TIF)

S3 Fig. Results broken down by paradigm and the presence/absence of smoothing. In Paradigm I, only the natural sounds were repeated. In Paradigm II, both natural and model-matched sounds were repeated. A smaller voxel size was employed in Paradigm II (2 mm isotropic instead of 2.1 × 2.1 × 4 mm for Paradigm I). (A) Natural versus model-matched dissimilarity maps computed with and without smoothing. Individual subjects are grouped by paradigm. Subjects are sorted by the reliability of their response to natural sounds for Paradigm I and by the reliability of their response to both natural and model-matched sounds for Paradigm II (measured using the NSE). (B) Annular analyses computed from data with and without smoothing. Each line corresponds to an individual subject and the color indicates the paradigm in which that subject was scanned (orange for Paradigm I and green for Paradigm II). NSE values are averaged across the left and right hemisphere because we observed similar trends in both hemispheres. NSE, normalized squared error.
(TIF)

S4 Fig. Noise-correction and voxel reliability. (A) The uncorrected NSE between responses to natural and model-matched sounds. (B) Corrected NSE maps (same as Fig 3C) replicated here for ease of comparison with the uncorrected maps. (C) Test-retest reliability of voxel responses measured with the NSE. Voxel reliability for Paradigm I (Group, S4, S5, and S6) is based on responses to natural sounds. Voxel reliability for Paradigm II (S1, S2, S3, S7, and S8)

is based on responses to both natural and model-matched sounds (responses to natural and model-matched sounds were combined into a single vector, and we computed the NSE for multiple measurements of this vector). Distance-to-PAC analyses are shown at the bottom of each panel (PAC defined tonotopically). Format is the same as Fig 3C and 3D. NSE, normalized squared error; PAC, primary auditory cortex.
(TIF)

S5 Fig. Dissimilarity versus distance from PAC using an anatomical rather than tonotopic definition of PAC. Voxels are binned based on their distance to the center of anatomical region TE1.1 [58], which is located in posteromedial Heschl's gyrus. Format is the same as Fig 3D. PAC, primary auditory cortex.
(TIF)

S6 Fig. Dissimilarity maps and annular analyses omitting intelligible speech stimuli. Maps plot the NSE between voxel responses to natural and model-matched sounds, omitting English speech and music with English vocals (all subjects were native English speakers). Format is the same as Fig 3C and 3D. NSE, normalized squared error.
(TIF)

S7 Fig. Individual subject maps of dissimilarity between responses to natural and model-matched sounds (NSE) for subsets of model features. Format is the same as Fig 4B. Only subjects scanned in Paradigm I are shown, because Paradigm II did not include model-matched sounds constrained by subsets of model features. NSE, normalized squared error.
(TIF)

S8 Fig. Individual subject prediction error maps based on the full spectrotemporal modulation model (maps for subsets of model features are omitted because of space constraints, but, like the maps for the full model, they resembled those of the group). Format is similar to Fig 5B.
(TIF)

S9 Fig. Prediction accuracy of the full spectrotemporal model using a larger set of 165 natural sounds tested in a prior study [45]. Format similar to Fig 5B and 5C.
(TIF)

S10 Fig. Cross predictions. (A) Prediction error maps of a model trained on natural sounds and tested on model-matched sounds. (B) Prediction error maps of a model trained on model-matched sounds and tested on natural sounds. (C) For comparison, the error of the measured voxel response to natural and model-matched sounds is reproduced here (same as Fig 3C and 3D). Annular analyses summarizing the error as a function of distance to tonotopically defined PAC are shown below each set of maps. Data are shown for subjects scanned in Paradigm II, for whom both natural and model-matched sounds were repeated, which made it possible to noise-correct the predictions. PAC, primary auditory cortex.
(TIF)

S11 Fig. Comparison of responses to natural and model-matched sounds for subsets of model features as well as the full model (Fig 6 only shows results from the full model). (A) Response of each component to each natural and corresponding model-matched sound. (B) NSE between natural and model-matched sounds. (C) Ratio of the standard deviation of responses to model-matched and natural sounds. (D) Correlation of responses to natural and model-matched sounds. NSE, normalized squared error.
(TIF)

S12 Fig. Examination of the nature of the divergent responses to natural and model-matched sounds. (A) Whole-brain maps plotting the variation in responses to natural versus model-matched sounds, measured as the ratio of the standard deviation of responses to the two sound sets. Cool colors indicated less response variation for model-matched sounds. Distance-to-PAC summary analysis is plotted below (PAC defined tonotopically). (B) Maps of the Pearson correlation between responses to natural and model-matched sounds, with distance-to-PAC analysis below. All of the measures have been corrected for noise. Analysis is based on data from Paradigm II, in which we measured responses to natural and model-matched sounds an equal number of times. For the response variation maps (panel A), we included all voxels with a reliable response across both natural and model-matched sounds (test-retest NSE < 0.4). For the correlation maps (panel B), we excluded voxels that did not have a reliable correlation to model-matched sounds (test-rest $r < 0.4$), as was the case in many nonprimary voxels, due to weak responses. For such voxels, it is difficult to estimate a reliable correlation, because the correlation is undefined as the variance of one variable goes to zero. NSE, normalized squared error; PAC, primary auditory cortex.
(TIF)

S13 Fig. Effect of the fMRI averaging window on the dissimilarity of responses to natural and model-matched sounds. Each stimulus was 10 seconds in duration but was split up into five 2-second segments (see S2 Fig). After each segment, a single scan acquisition was collected. Analyses in the main text were based on the average response of the second through fifth acquisitions after the onset of each stimulus block (first acquisition was discarded to account for the hemodynamic delay). Here, we test the sensitivity of the results to the averaging window by restricting the analysis to data averaged across acquisitions 2 and 3 (panel A) or 4 and 5 (panel B). Compare with Fig 3C and 3D.
(TIF)

S14 Fig. Validation of the model-matching synthesis procedure via comparison of time-averaged statistics for natural and model-matched sounds. (A) Cochleagrams for a natural sound (a speech excerpt) and four corresponding model-matched sounds. (B—E) Each model was defined by a set of feature responses. Panels plot a time-averaged measure of the amplitude of each feature's response to the example natural and model-matched sounds shown in panel A. The right-most panel plots the correlation of the filter amplitudes across all model filters for corresponding natural and model-matched sounds. Each dot corresponds to a single pair of natural/model-matched sounds. (B) Amplitude of each cochlear frequency channel envelope, averaged across time. Cochlear channel power is matched in all four conditions, as desired/expected. (C) Temporal modulation amplitude (standard deviation of each temporal modulation feature across time) for example natural and model-matched sounds. Modulation amplitude is plotted as a function of the filter's preferred audio frequency and temporal modulation rate. (D) Spectral modulation amplitude plotted as a function of the filter's preferred audio frequency and spectral modulation scale. (E) Spectrotemporal modulation amplitude plotted as a function of temporal modulation rate and spectral modulation scale for an example audio frequency channel (centered at 200 Hz).
(TIF)

S15 Fig. Comparison of how well the natural and model-matched sounds are matched when evaluated using spectrotemporal filters that differed from those used to generate the model-matched sounds. In each case, we plot an example filter from the model (left), the amplitude (standard deviation) of the filter responses as a function of the temporal rate and

spectral scale for an example audio frequency channel (centered at 200 Hz) (middle), and the correlation of the amplitude across all of the filters for the natural and model-matched sounds (right) (format similar to S14E Fig). (A) The original spectrotemporal filters from Chi and colleagues (2005) that were used to constrain the model-matched sounds (same as S14E Fig). (B) Spectrotemporal filters with randomized temporal and spectral phases. (C) A model with narrower bandwidths and more filters to compensate (these filters are broader in extent when visualized in the time-frequency plane). (D) A random filter basis with variable temporal and spectral extent. In all four cases, the measured modulation power is similar for the natural and model-matched sounds. This suggests that voxels with similar responses to natural and model-matched sounds are compatible with a wide range of spectrotemporal modulation filters, and that a wide range of such filters are ruled out as descriptions of voxels that give different responses to natural and model-matched sounds, such as those we observed in nonprimary regions.
(TIF)

S16 Fig. Comparison of noise-correction using noise estimates exclusively from responses to natural sounds or from both natural and model-matched sounds. Noise-correction required estimating the power of the noise for natural and model-matched sounds. For Paradigm I, only responses to natural sounds were repeated in each scan. Using data from Paradigm II, we tested whether it is necessary to separately estimate the noise power for natural and model-matched sounds or whether one can assume they are equal. (A) Noise-corrected NSE value computed by assuming the noise power for natural and model-matched sounds is equal, using only responses to natural sounds to compute it. (B) Noise-corrected NSE values computed by separately estimating the noise power for natural and model-matched sounds (same maps as those in Fig 3C). Results are similar in both cases. NSE, normalized squared error.
(TIF)

S17 Fig. Results of noise-correction simulations. (A) Each dot corresponds to a single simulated voxel. The noise-corrupted and noise-corrected NSE values are plotted against the NSE values of the true signals uncorrupted by noise. Results have been grouped by the reliability of the simulated voxel responses, as measured by the test-retest NSE of the voxel responses (from high to low reliability, going from top to bottom). (B) The median and standard deviation (central 68% of samples) of the noise-corrupted or noise-corrected NSE values. (C–D) Same as for panels A and B, but for Simulation II (see “Evaluating the noise-corrected NSE with simulated data” in Materials and methods for details of the two simulations). NSE, normalized squared error.
(TIF)

S18 Fig. Test-retest reliability of the noise-corrected NSE. Each dot corresponds to a single voxel. The Spearman rank correlation is shown at the top of each plot for each subject. Results are shown for subjects scanned in Paradigm II, for which there was sufficient data to compute two separate estimates of the noise-corrected NSE (which requires four splits of data). NSE, normalized squared error.
(TIF)

S1 Table. The presence/absence of duplications along Heschl’s gyrus for each hemisphere of each subject that was scanned multiple times in the experiment. Categories were determined by inspection using the scheme described in Da Costa and colleagues (2011) [52].
(XLSX)

Author Contributions

Conceptualization: Sam V. Norman-Haignere, Josh H. McDermott.

Funding acquisition: Sam V. Norman-Haignere, Josh H. McDermott.

Investigation: Sam V. Norman-Haignere, Josh H. McDermott.

Methodology: Sam V. Norman-Haignere, Josh H. McDermott.

Resources: Josh H. McDermott.

Software: Sam V. Norman-Haignere, Josh H. McDermott.

Supervision: Josh H. McDermott.

Visualization: Sam V. Norman-Haignere.

Writing – original draft: Sam V. Norman-Haignere, Josh H. McDermott.

Writing – review & editing: Sam V. Norman-Haignere, Josh H. McDermott.

References

1. Simoncelli EP, Olshausen BA. Natural image statistics and neural representation. Annu Rev Neurosci. 2001; 24: 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193 PMID: 11520932
2. Woolley SM, Fremouw TE, Hsu A, Theunissen FE. Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. Nat Neurosci. 2005; 8: 1371–1379. https://doi.org/10.1038/nn1536 PMID: 16136039
3. Smith EC, Lewicki MS. Efficient auditory coding. Nature. 2006; 439: 978–982. https://doi.org/10.1038/nature04485 PMID: 16495999
4. Sharpee T, Rust NC, Bialek W. Analyzing neural responses to natural signals: maximally informative dimensions. Neural Comput. 2004; 16: 223–250. https://doi.org/10.1162/089976604322742010 PMID: 15006095
5. Naselaris T, Kay KN, Nishimoto S, Gallant JL. Encoding and decoding in fMRI. Neuroimage. 2011; 56: 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073 PMID: 20691790
6. Huth AG, Nishimoto S, Vu AT, Gallant JL. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron. 2012; 76: 1210–1224. https://doi.org/10.1016/j.neuron.2012.10.014 PMID: 23259955
7. Santoro R, Moerel M, De Martino F, Goebel R, Ugurbil K, Yacoub E, et al. Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput Biol. 2014; 10 (1):e1003412. https://doi.org/10.1371/journal.pcbi.1003412 PMID: 24391486
8. Theunissen FE, Elie JE. Neural processing of natural sounds. Nat Rev Neurosci. 2014; 15: 355–366. https://doi.org/10.1038/nrn3731 PMID: 24840800
9. Di Liberto GM, O'Sullivan JA, Lalor EC. Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol. 2015; 25: 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030 PMID: 26412129
10. Kriegeskorte N. Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu Rev Vis Sci. 2015; 1: 417–446. https://doi.org/10.1146/annurev-vision-082114-035447 PMID: 28532370
11. Thorson IL, Liénard J, David SV. The essential complexity of auditory receptive fields. PLoS Comput Biol. 2015; 11: e1004628. https://doi.org/10.1371/journal.pcbi.1004628 PMID: 26683490
12. Hullett PW, Hamilton LS, Mesgarani N, Schreiner CE, Chang EF. Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. J Neurosci. 2016; 36: 2014–2026. https://doi.org/10.1523/JNEUROSCI.1779-15.2016 PMID: 26865624
13. Kozlov AS, Gentner TQ. Central auditory neurons have composite receptive fields. Proc Natl Acad Sci. 2016; 113: 1441–1446. https://doi.org/10.1073/pnas.1506903113 PMID: 26787894
14. Willmore BDB, Schoppe O, King AJ, Schnupp JWH, Harper NS. Incorporating midbrain adaptation to mean sound level improves models of auditory cortical processing. J Neurosci. 2016; 36: 280–289. https://doi.org/10.1523/JNEUROSCI.2441-15.2016 PMID: 26758822

15. Yamins DL, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci. 2016; 19: 356–365. https://doi.org/10.1038/nn.4244 PMID: 26906502

16. Kell AJ, Yamins DL, Shook EN, Norman-Haignere SV, McDermott JH. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron. 2018; 98(3). https://doi.org/10.1016/j.neuron.2018.03.044

17. de Heer WA, Huth AG, Griffiths TL, Gallant JL, Theunissen FE. The hierarchical cortical organization of human speech processing. J Neurosci. 2017; 3267–16. https://doi.org/10.1523/JNEUROSCI.3267-16.2017 PMID: 28588065

18. Groen II, Greene MR, Baldassano C, Fei-Fei L, Beck DM, Baker CI. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. Elife. 2018; 7: e32962. https://doi.org/10.7554/eLife.32962 PMID: 29513219

19. Meyer AF, Williamson RS, Linden JF, Sahani M. Models of neuronal stimulus-response functions: elaboration, estimation, and evaluation. Front Syst Neurosci. 2017; 10. https://doi.org/10.3389/fnsys.2016.00109 PMID: 28127278

20. Chi T, Ru P, Shamma SA. Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am. 2005; 118: 887–906. https://doi.org/10.1121/1.1945807 PMID: 16158645

21. Dau T, Kollmeier B, Kohlrausch A. Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. J Acoust Soc Am. 1997; 102: 2892–2905. https://doi.org/10.1121/1.420344 PMID: 9373976

22. Elliott TM, Theunissen FE. The modulation transfer function for speech intelligibility. PLoS Comput Biol. 2009; 5: e1000302. https://doi.org/10.1371/journal.pcbi.1000302 PMID: 19266016

23. Patil K, Pressnitzer D, Shamma S, Elhilali M. Music in our ears: the biological bases of musical timbre perception. PLoS Comput Biol. 2012; 8: e1002759. https://doi.org/10.1371/journal.pcbi.1002759 PMID: 23133363

24. McDermott JH, Schemitsch M, Simoncelli EP. Summary statistics in auditory perception. Nat Neurosci. 2013; 16: 493–498. https://doi.org/10.1038/nn.3347 PMID: 23434915

25. McWalter R, McDermott JH. Adaptive and selective time averaging of auditory scenes. Curr Biol. 2018; 28: 1405–1418. https://doi.org/10.1016/j.cub.2018.03.049 PMID: 29681472

26. deCharms CR, Blake DT, Merzenich MM. Optimizing sound features for cortical neurons. Science. 1998; 280: 1439–1444. https://doi.org/10.1126/science.280.5368.1439 PMID: 9603734

27. Theunissen FE, Sen K, Doupe AJ. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J Neurosci. 2000; 20: 2315–2331. https://doi.org/10.1523/JNEUROSCI.20-06-02315.2000 PMID: 10704507

28. Depireux DA, Simon JZ, Klein DJ, Shamma SA. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol. 2001; 85: 1220–1234. https://doi.org/10.1152/jn.2001.85.3.1220 PMID: 11247991

29. Miller LM, Escabí MA, Read HL, Schreiner CE. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol. 2002; 87: 516–527. https://doi.org/10.1152/jn.00395.2001 PMID: 11784767

30. Nelson PC, Carney LH. A phenomenological model of peripheral and central neural responses to amplitude-modulated tones. J Acoust Soc Am. 2004; 116: 2173–2186. https://doi.org/10.1121/1.1784442 PMID: 15532650

31. Fishman YI, Steinschneider M. Temporally dynamic frequency tuning of population responses in monkey primary auditory cortex. Hear Res. 2009; 254: 64–76. https://doi.org/10.1016/j.heares.2009.04.010 PMID: 19389466

32. Schönwiesner M, Zatorre RJ. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc Natl Acad Sci. 2009; 106: 14611–14616. https://doi.org/10.1073/pnas.0907682106 PMID: 19667199

33. Barton B, Venezia JH, Saberi K, Hickok G, Brewer AA. Orthogonal acoustic dimensions define auditory field maps in human cortex. Proc Natl Acad Sci. 2012; 109: 20738–20743. https://doi.org/10.1073/pnas.1213381109 PMID: 23188798

34. Rabinowitz NC, Willmore BD, Schnupp JW, King AJ. Spectrotemporal contrast kernels for neurons in primary auditory cortex. J Neurosci. 2012; 32: 11271–11284. https://doi.org/10.1523/JNEUROSCI.1715-12.2012 PMID: 22895711

35. Herdener M, Esposito F, Scheffler K, Schneider P, Logothetis NK, Uludag K, et al. Spatial representations of temporal and spectral sound cues in human auditory cortex. Cortex. 2013; 49: 2822–2833. https://doi.org/10.1016/j.cortex.2013.04.003 PMID: 23706955

36. Mesgarani N, Cheung C, Johnson K, Chang EF. Phonetic feature encoding in human superior temporal gyrus. Science. 2014; 343: 1006–1010. https://doi.org/10.1126/science.1245994 PMID: 24482117

37. Mesgarani N, Slaney M, Shamma S, others. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. Audio Speech Lang Process IEEE Trans On. 2006; 14: 920–930. https://doi.org/10.1109/TSA.2005.858055

38. Greenberg S, Carvey H, Hitchcock L, Chang S. Temporal properties of spontaneous speech—a syllable-centric perspective. J Phon. 2003; 31: 465–485. https://doi.org/10.1016/j.wocn.2003.09.005

39. Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D. Temporal modulations in speech and music. Neurosci Biobehav Rev. 2017; 81:181–187. https://doi.org/10.1016/j.neubiorev.2017.02.011

40. McDermott JH, Simoncelli EP. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron. 2011; 71: 926–940. https://doi.org/10.1016/j.neuron.2011.06.032 PMID: 21903084

41. Wessinger CM, VanMeter J, Tian B, Van Lare J, Pekar J, Rauschecker JP. Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. J Cogn Neurosci. 2001; 13: 1–7. https://doi.org/10.1162/089892901564108 PMID: 11224904

42. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. Nature. 2000; 403: 309–312. https://doi.org/10.1038/35002078 PMID: 10659849

43. Overath T, McDermott JH, Zarate JM, Poeppel D. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat Neurosci. 2015; 18: 903–911. https://doi.org/10.1038/nn.4021 PMID: 25984889

44. Binder JR, Frost JA, Hammeke TA, Bellgowan PSF, Springer JA, Kaufman JN, et al. Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex. 2000; 10: 512–528. https://doi.org/10.1093/cercor/10.5.512 PMID: 10847601

45. Norman-Haignere SV, Kanwisher NG, McDermott JH. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron. 2015; 88: 1281–1296. https://doi.org/10.1016/j.neuron.2015.11.035 PMID: 26687225

46. Leaver AM, Rauschecker JP. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J Neurosci. 2010; 30: 7604–7612. https://doi.org/10.1523/JNEUROSCI.0296-10.2010 PMID: 20519535

47. Angulo-Perkins A, Aubé W, Peretz I, Barrios FA, Armony JL, Concha L. Music listening engages specific cortical regions within the temporal lobes: Differences between musicians and non-musicians. Cortex. 2014; 59: 126–137. https://doi.org/10.1016/j.cortex.2014.07.013 PMID: 25173956

48. Heeger DJ, Bergen JR. Pyramid-based texture analysis/synthesis. Proceedings of the 22nd annual conference on Computer graphics and interactive techniques. ACM; 1995. pp. 229–238. https://doi.org/10.1145/218380.218446

49. Portilla J, Simoncelli EP. A parametric texture model based on joint statistics of complex wavelet coefficients. Int J Comput Vis. 2000; 40: 49–70. https://doi.org/10.1023/A:1026553619983

50. Balas BJ. Texture synthesis and perception: Using computational models to study texture representations in the human visual system. Vision Res. 2006; 46: 299–309. https://doi.org/10.1016/j.visres.2005.04.013 PMID: 15964047

51. Humphries C, Liebenthal E, Binder JR. Tonotopic organization of human auditory cortex. NeuroImage. 2010; 50: 1202–1211. https://doi.org/10.1016/j.neuroimage.2010.01.046 PMID: 20096790

52. Costa SD, van der Zwaag W, Marques JP, Frackowiak RSJ, Clarke S, Saenz M. Human primary auditory cortex follows the shape of heschl's gyrus. J Neurosci. 2011; 31: 14067–14075. https://doi.org/10.1523/JNEUROSCI.2000-11.2011 PMID: 21976491

53. Moerel M, De Martino F, Formisano E. Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. J Neurosci. 2012; 32: 14205–14216. https://doi.org/10.1523/JNEUROSCI.1388-12.2012 PMID: 23055490

54. Baumann S, Petkov CI, Griffiths TD. A unified framework for the organization of the primate auditory cortex. Front Syst Neurosci. 2013; 7: 11. https://doi.org/10.3389/fnsys.2013.00011 PMID: 23641203

55. Leaver AM, Rauschecker JP. Functional topography of human auditory cortex. J Neurosci. 2016; 36: 1416–1428. https://doi.org/10.1523/JNEUROSCI.0226-15.2016 PMID: 26818527

56. Schönwiesner M, Dechent P, Voit D, Petkov CI, Krumbholz K. Parcellation of human and monkey core auditory cortex with fMRI pattern classification and objective detection of tonotopic gradient reversals. Cereb Cortex. 2015; 25: 3278–3289. https://doi.org/10.1093/cercor/bhu124 PMID: 24904067

57. Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, et al. A multi-modal parcellation of human cerebral cortex. Nature. 2016; 536: 171–178. https://doi.org/10.1038/nature18933 PMID: 27437579

58. Morosan P, Rademacher J, Schleicher A, Amunts K, Schormann T, Zilles K. Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. Neuroimage. 2001; 13: 684–701. https://doi.org/10.1006/nimg.2000.0715 PMID: 11305897

59. Dick F, Tierney AT, Lutti A, Josephs O, Sereno MI, Weiskopf N. In vivo functional and myeloarchitectonic mapping of human primary auditory areas. J Neurosci. 2012; 32: 16095–16105. https://doi.org/10.1523/JNEUROSCI.1712-12.2012 PMID: 23152594

60. Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature. 2016; 532: 453–458. https://doi.org/10.1038/nature17637 PMID: 27121839

61. Santoro R, Moerel M, De Martino F, Valente G, Ugurbil K, Yacoub E, et al. Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. Proc Natl Acad Sci. 2017; 114: 4799–4804. https://doi.org/10.1073/pnas.1617622114 PMID: 28420788

62. Carrasco A, Lomber SG. Evidence for hierarchical processing in cat auditory cortex: nonreciprocal influence of primary auditory cortex on the posterior auditory field. J Neurosci. 2009; 29: 14323–14333. https://doi.org/10.1523/JNEUROSCI.2905-09.2009 PMID: 19906979

63. Rauschecker JP, Scott SK. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci. 2009; 12: 718–724. https://doi.org/10.1038/nn.2331 PMID: 19471271

64. Okada K, Rong F, Venezia J, Matchin W, Hsieh I-H, Saberi K, et al. Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cereb Cortex. 2010; 20: 2486–2495. https://doi.org/10.1093/cercor/bhp318 PMID: 20100898

65. Kaas JH, Hackett TA. Subdivisions of auditory cortex and processing streams in primates. Proc Natl Acad Sci. 2000; 97: 11793–11799. https://doi.org/10.1073/pnas.97.22.11793 PMID: 11050211

66. Rauschecker JP, Tian B, Hauser M. Processing of complex sounds in the macaque nonprimary auditory cortex. Science. 1995; 268(5207): 111–114. https://doi.org/10.1126/science.7701330

67. Recanzone GH, Cohen YE. Serial and parallel processing in the primate auditory cortex revisited. Behav Brain Res. 2010; 206: 1–7. https://doi.org/10.1016/j.bbr.2009.08.015 PMID: 19686779

68. Hickok G, Poeppel D. The cortical organization of speech processing. Nat Rev Neurosci. 2007; 8: 393–402. https://doi.org/10.1038/nrn2113 PMID: 17431404

69. Staeren N, Renvall H, De Martino F, Goebel R, Formisano E. Sound categories are represented as distributed patterns in the human auditory cortex. Curr Biol. 2009; 19: 498–502. https://doi.org/10.1016/j.cub.2009.01.066 PMID: 19268594

70. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521: 436. https://doi.org/10.1038/nature14539 PMID: 26017442

71. Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci. 2014; 111: 8619–8624. https://doi.org/10.1073/pnas.1403112111 PMID: 24812127

72. Eickenberg M, Gramfort A, Varoquaux G, Thirion B. Seeing it all: Convolutional network layers map the function of the human visual system. NeuroImage. 2017; 152: 184–194. https://doi.org/10.1016/j.neuroimage.2016.10.001 PMID: 27777172

73. Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A. Comparison of deep neural networks to spatiotemporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep. 2016; 6: 27755. https://doi.org/10.1038/srep27755 PMID: 27282108

74. Khaligh-Razavi S-M, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol. 2014; 10: e1003915. https://doi.org/10.1371/journal.pcbi.1003915 PMID: 25375136

75. Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, et al. Deep convolutional models improve predictions of macaque V1 responses to natural images. bioRxiv. 2017; 201764. https://doi.org/10.1101/201764

76. Güçlü U, van Gerven MA. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J Neurosci. 2015; 35: 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015 PMID: 26157000

77. Sahani M, Linden JF. How linear are auditory cortical responses? Advances in neural information processing systems. 2003. pp. 125–132.

78. Atencio CA, Sharpee TO, Schreiner CE. Cooperative nonlinearities in auditory cortical neurons. Neuron. 2008; 58: 956–966. https://doi.org/10.1016/j.neuron.2008.04.026 PMID: 18579084

79. Sadagopan S, Wang X. Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex. J Neurosci. 2009; 29: 11192–11202. https://doi.org/10.1523/JNEUROSCI.1286-09.2009 PMID: 19741126

80. Balas B, Nakano L, Rosenholtz R. A summary-statistic representation in peripheral vision explains visual crowding. J Vis. 2009; 9: 13–13. https://doi.org/10.1167/9.12.13 PMID: 20053104

81. Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, Movshon JA. A functional and perceptual signature of the second visual area in primates. Nat Neurosci. 2013; 16: 974–981. https://doi.org/10.1038/nn.3402 PMID: 23685719

82. McDermott JH, Oxenham AJ, Simoncelli EP. Sound texture synthesis via filter statistics. Applications of Signal Processing to Audio and Acoustics, 2009 WASPAA'09 IEEE Workshop on. IEEE; 2009. pp. 297–300. https://doi.org/10.1109/ASPAA.2009.5346467

83. David SV, Shamma SA. Integration over multiple timescales in primary auditory cortex. J Neurosci. 2013; 33: 19154–19166. https://doi.org/10.1523/JNEUROSCI.2270-13.2013 PMID: 24305812

84. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Med Image Anal. 2001; 5: 143–156. https://doi.org/10.1016/S1361-8415(01)00036-6 PMID: 11516708

85. Greve DN, Fischl B. Accurate and robust brain image alignment using boundary-based registration. Neuroimage. 2009; 48: 63. https://doi.org/10.1016/j.neuroimage.2009.06.060 PMID: 19573611

86. Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis: I. Segmentation and surface reconstruction. NeuroImage. 1999; 9: 179–194. https://doi.org/10.1006/nimg.1998.0395 PMID: 9931268

87. Hall DA, Haggard MP, Akeroyd MA, Palmer AR, Summerfield AQ, Elliott MR, et al. Sparse temporal sampling in auditory fMRI. Hum Brain Mapp. 1999; 7: 213–223. https://doi.org/10.1002/(SICI)1097-0193(1999)7:3<213::AID-HBM5>3.0.CO;2-N PMID: 10194620

88. Norman-Haignere S, Kanwisher N, McDermott JH. Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. J Neurosci. 2013; 33: 19451–19469. https://doi.org/10.1523/JNEUROSCI.2880-13.2013 PMID: 24336712

89. Norman-Haignere SV, Albouy P, Caclin A, McDermott JH, Kanwisher NG, Tillmann B. Pitch-responsive cortical regions in congenital amusia. J Neurosci. 2016; 36(10):2986–2994. https://doi.org/10.1523/JNEUROSCI.2705-15.2016

90. Kay K, Rokem A, Winawer J, Dougherty R, Wandell B. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. Brain Imaging Methods. 2013; 7: 247. https://doi.org/10.3389/fnins.2013.00247 PMID: 24381539

91. Woolrich MW, Ripley BD, Brady M, Smith SM. Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage. 2001; 14: 1370–1386. https://doi.org/10.1006/nimg.2001.0931 PMID: 11707093

92. Schoppe O, Harper NS, Willmore BDB, King AJ, Schnupp JWH. Measuring the performance of neural models. Front Comput Neurosci. 2016; 10: 10. https://doi.org/10.3389/fncom.2016.00010 PMID: 26903851

93. Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. Neural Netw IEEE Trans On. 1999; 10(3): 626–634. https://doi.org/10.1109/72.761722

94. Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, et al. Reconstructing speech from human auditory cortex. PLoS Biol. 2012; 10: e1001251. https://doi.org/10.1371/journal.pbio.1001251 PMID: 22303281

95. Norman-Haignere S, McDermott JH. Distortion products in auditory fMRI research: measurements and solutions. NeuroImage. 2016; 129: 401–413. https://doi.org/10.1016/j.neuroimage.2016.01.050 PMID: 26827809

96. Saxe R, Brett M, Kanwisher N. Divide and conquer: a defense of functional localizers. Neuroimage. 2006; 30: 1088–1096. https://doi.org/10.1016/j.neuroimage.2005.12.062 PMID: 16635578

97. de Cheveigné A, Wong DD, Di Liberto GM, Hjortkjær J, Slaney M, Lalor E. Decoding the auditory brain with canonical component analysis. NeuroImage. 2018; 172: 206–216. https://doi.org/10.1016/j.neuroimage.2018.01.033 PMID: 29378317

98. Chen P-HC, Chen J, Yeshurun Y, Hasson U, Haxby J, Ramadge PJ. A reduced-dimension fMRI shared response model. Advances in Neural Information Processing Systems. 2015. pp. 460–468.