

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

Sam V. Norman-Haignere^{1,2,3,*}, Josh H. McDermott¹

1. Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

2. HHMI Postdoctoral Fellow of the Life Sciences Research Institute

3. Laboratoire des Systèmes Perceptifs, Département d'études cognitives, ENS, PSL University, CNRS, Paris France

* Corresponding Author: snormanhaignere@gmail.com

Short Title: Model-matched stimuli reveal distinct cortical computations

Abstract

A central goal of sensory neuroscience is to construct models that can explain neural responses to complex, 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 stimuli that yield the same model response as a natural stimulus, and test whether the natural and “model-matched” stimulus elicit the same neural response. 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 shown that this model has good predictive power throughout auditory cortex, but this finding could reflect stimulus-driven correlations. We observed that fMRI voxel responses to natural and model-matched stimuli were nearly equivalent in primary auditory cortex, but that non-primary regions showed highly divergent responses to the two sound sets, suggesting that neurons in non-primary regions extract higher-order properties not made explicit by traditional models. This dissociation between primary and non-primary regions was not clear from model predictions due to the influence of stimulus-driven response correlations. Our methodology enables stronger tests of sensory models and could be broadly applied in other domains.

Author Summary

Modeling neural responses to natural stimuli is a core goal of sensory neuroscience. Here we propose a new approach for testing sensory models: we synthesize a “model-matched” stimulus that yields the same model response as a natural stimulus, and test whether it produces the same neural response. We used model-matching to test whether a standard model of auditory cortex can explain human cortical responses measured with fMRI. Model-matched stimuli produced nearly equivalent voxel responses in primary auditory cortex, but highly divergent responses in non-primary regions. This dissociation was not evident using more standard approaches for model testing, and suggests that non-primary regions compute higher-order stimulus properties not captured by traditional models. The methodology could be broadly applied in other domains.

Introduction

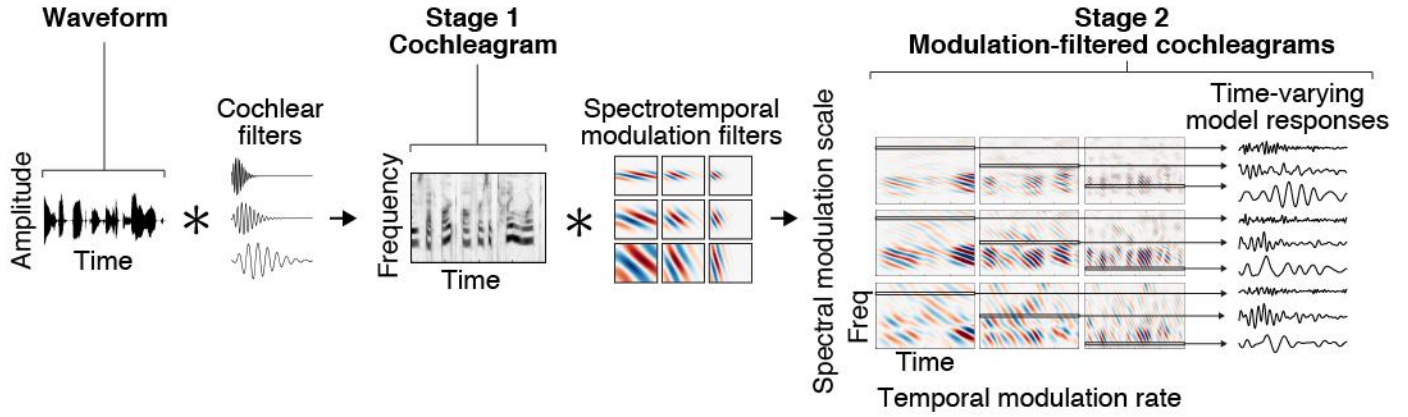
One definition of understanding a neural system is to be able to build a model that can predict its responses. Responses to natural stimuli are of particular interest, both because natural stimuli are complex and varied, and thus provide a strong test of a model, and because sensory systems are presumably 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^{4–15}.

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)¹⁶. 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^{4,17}, but have been less noted in cognitive neuroscience^{16,18}.

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, respectively¹⁹ (Figure 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^{20–23} and to explain neural responses throughout the auditory pathway^{2,7,9,11,12,24–35}. But in natural stimuli, the responses of these second-stage filters are often correlated with other sound properties, such as semantic categories (Figure 1B)³⁶, 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 patterning³⁷, music has distinctive temporal modulations reflective of its beat structure³⁸, and both speech and music have characteristic spectral modulations due to harmonic frequency structure¹⁹. However, speech, music, and other natural sounds also have many unique properties that are not captured by spectrotemporal modulation alone³⁹. Thus, if a neuron responds more to speech than to other sounds, 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 stimuli that yield 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.

A Schematic of two-stage filterbank model



B Cochleograms and modulation energy for example natural sounds

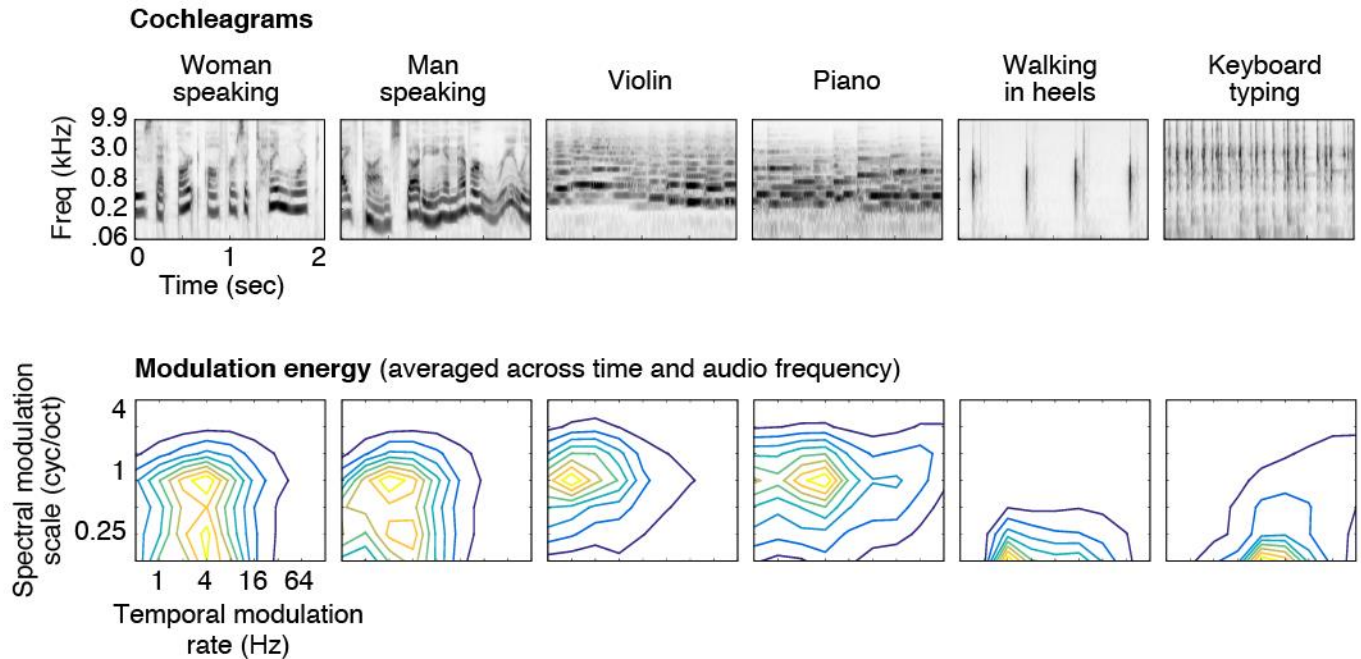


Figure 1. Illustration of the auditory model tested in this study. (A) The model consists of two cascaded stages of filtering. In the first stage, a cochleagram is computed by convolving each sound with audio filters tuned to different frequencies, extracting the temporal envelope of the resulting filter 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 output 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) Cochleograms 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.

We demonstrate the method by using it to evaluate whether a common filter bank model of the 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 non-primary regions^{16,40,41}, such as selectivity for voice⁴², speech^{43,44} and music^{45–47}. At the same time, other studies have demonstrated that the standard filter bank model has good predictive accuracy throughout primary and non-primary regions^{7,12,41,45}, raising the possibility that primary and non-primary 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 Figure 1A¹⁹. 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 non-primary areas. This functional difference between primary and non-primary regions was much less evident using conventional model predictions due to the effect of stimulus-driven response correlations. Our findings provide novel evidence for functional differentiation between primary and non-primary auditory cortex, and suggest that non-primary 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.

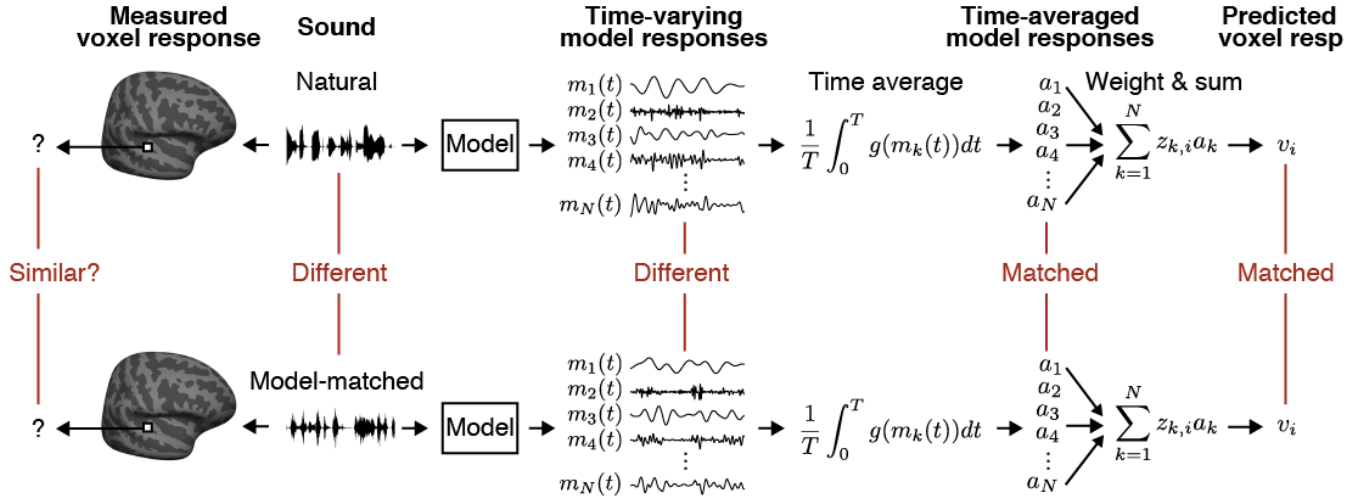
Results

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 (Figure 1A&2A). In general, the response of these features will differ across natural sounds, both in their temporal pattern and their time-averaged properties (Figure S1A). 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 temporal pattern is different. To test this prediction, we iteratively modified a noise stimulus, which was initially unstructured, so as to match the time-averaged model responses (Figure S1B), similar to methods for texture synthesis^{39,48–50}. Because the temporal patterns of the model responses are unconstrained, the model-matched sounds differ from the natural sounds they were matched to.

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 time scale 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) (Figure 2A):

A Schematic of voxel model-matching approach



B 36 natural sounds tested

- | | |
|-------------------|-----------------------|
| 1. Woman speaking | 19. Biting & chewing |
| 2. Man speaking | 20. Finger tapping |
| 3. Spanish | 21. Walking on leaves |
| 4. French | 22. Scratching |
| 5. Italian | 23. Walking in heels |
| 6. German | 24. Writing on paper |
| 7. Hindi | 25. Heart beat |
| 8. Russian | 26. Cicadas |
| 9. Big band music | 27. Crickets |
| 10. Bluegrass | 28. Baby Crying |
| 11. Cello | 29. Breathing |
| 12. Orchestra | 30. Clock ticking |
| 13. Piano | 31. Siren |
| 14. Saxophone | 32. Keyboard Typing |
| 15. Violin | 33. Chimes |
| 16. Latin music | 34. Chopping food |
| 17. Country song | 35. Crumpling paper |
| 18. R&B song | 36. Keys jingling |

- Category labels
- English speech
 - Foreign speech
 - Instr. music
 - Vocal music

- Human nonvocal
- Animal nonvocal
- NonSpeech voc
- Mechanical
- Env. sounds

C Cochleagrams of example natural sounds and corresponding model-matched sounds

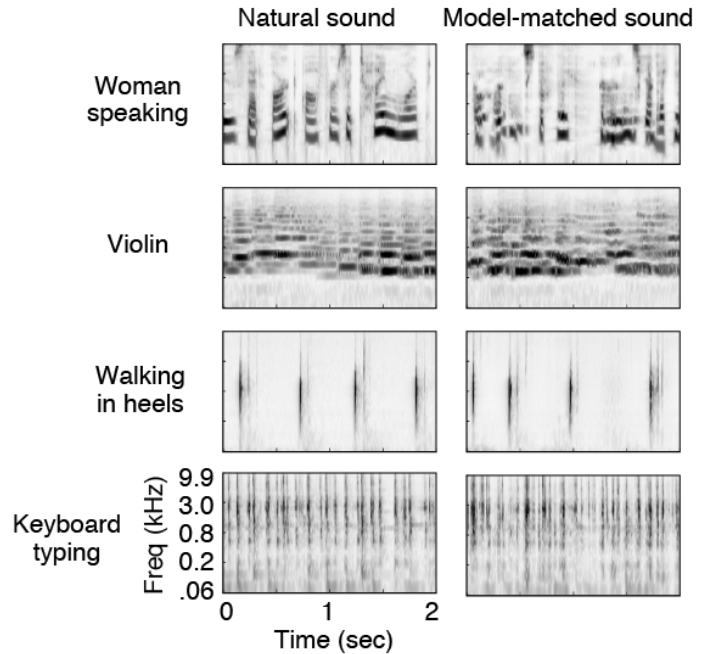


Figure 2. Model-matching methodology and experimental stimuli. (A) The logic of the model-matching procedure as applied to fMRI. The models we consider are defined by the time-varying response of a set of model features ($m_k(t)$) to a sound (as in the auditory model shown in Figure 1A). Because fMRI is thought to pool activity across neurons and time, we modeled fMRI voxel responses as weighted sums of time-averaged model responses (equations 1-2) (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 regardless of how these time-averaged activities are weighted (for voxels containing neurons that can be approximated by the model features). The temporal response pattern of the model features was otherwise unconstrained. As a consequence the model-matched sounds were distinct from the natural sounds they were matched to. (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 study⁴⁵. Font color denotes membership in one of 9 semantic categories (as determined by human listeners⁴⁵). (C) Cochleagrams are shown for four natural and model-matched sounds constrained by the spectrotemporal modulation model shown in Figure 1A¹⁸.

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

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

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 equations 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 (Figure 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 equation 2 are matched; see Figure 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 (Figure S1; see *Model-matching synthesis algorithm* in the 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 modulation¹⁹ in a spectrogram-like representation of sound termed a cochleagram (Figure 1A), produced by passing a sound signal 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 timecourses 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](http://mcdermottlab.mit.edu/svnh/model-matching/Stimuli%20from%20Model-Matching%20Experiment.html) for examples). This observation demonstrates that the time-averaged properties of the model's features, which approximately capture the modulation spectrum (Figure 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.

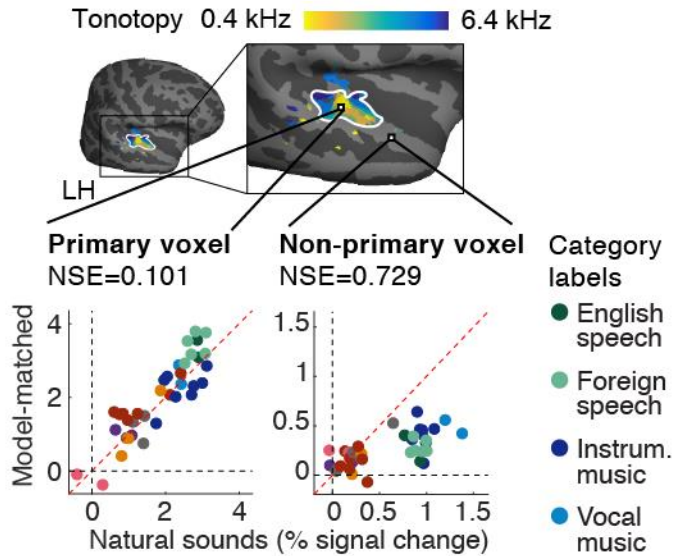
Comparing fMRI responses to natural and model-matched sounds

We measured fMRI responses to a diverse set of 36 natural sounds and their corresponding model-matched sounds (Figure 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 (Figure S2; see *Stimulus presentation and scanning procedure* in the Methods). The model-matched sounds were constrained by all of the features from the two-stage filter bank model shown in Figure 1A (see below for results from sounds constrained by simpler models). We first plot the response of two example voxels from a single subject (Figure 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).

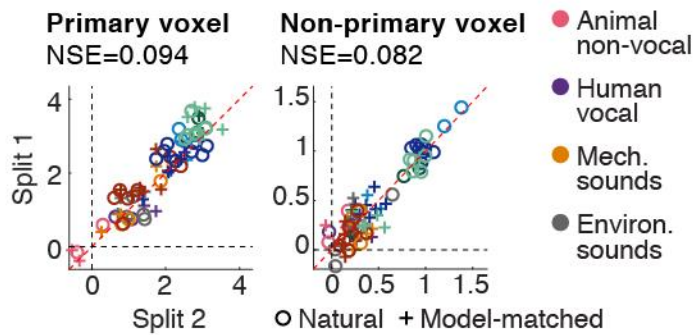
As shown in Figure 3A, the response of the primary voxel to natural and model-matched sounds was similar. By contrast, the non-primary voxel responded notably less to the model-matched sounds. We quantified the dissimilarity of responses to natural and model-matched sounds by computing the squared error 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 the 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 non-primary 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.

For these example voxels, the normalized squared error (NSE) was higher for the non-primary voxel (NSE=0.729) than the primary voxel (NSE=0.101), reflecting the fact that the non-primary 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, since a similar NSE value was observed between two different measurements of the voxel’s response to natural sounds (NSE=0.094) (Figure 3B). By contrast in the non-primary 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 lower SNR.

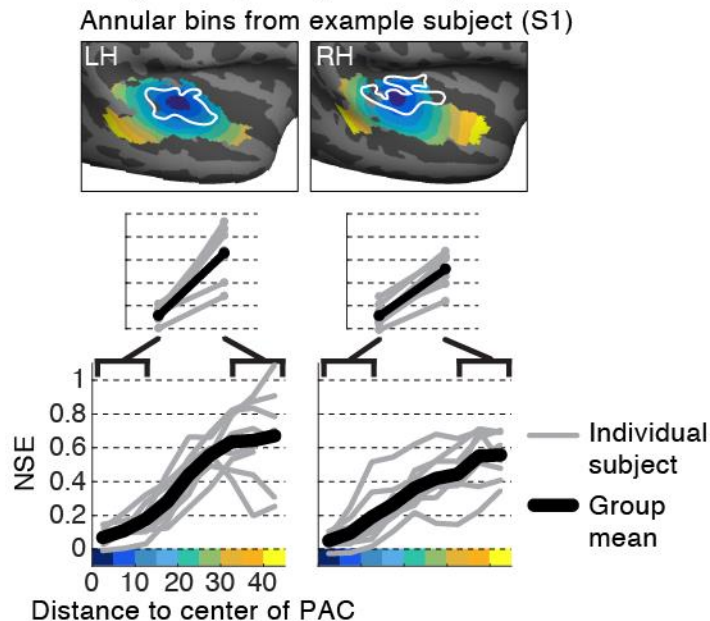
A Example voxel responses to natural and model-matched sounds



B Voxel reliability



D Dissimilarity vs. distance to to PAC (tonotopically defined)



C Dissimilarity of voxel responses to natural vs. model-matched sounds

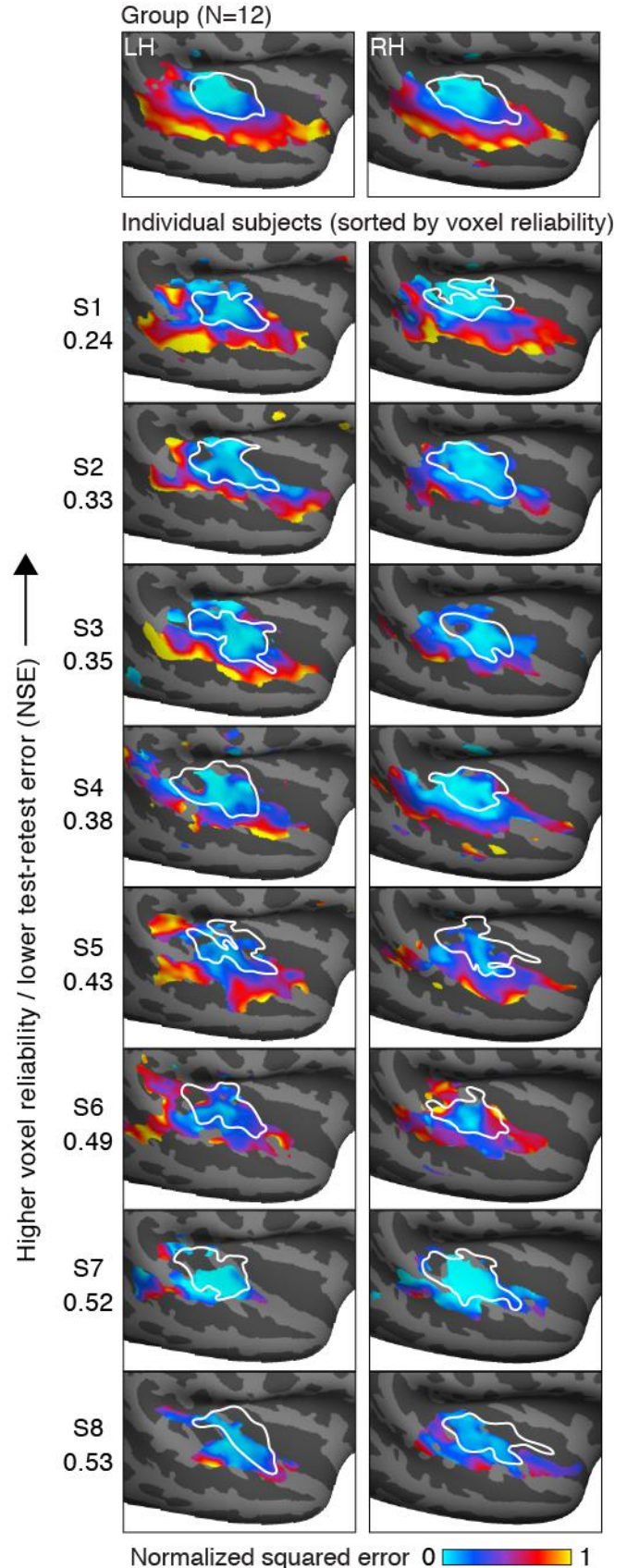


Figure 3. Voxel responses to natural and model-matched sounds. (A) Responses to natural and model-matched sounds from two example voxels from a single subject. One voxel is drawn from the low-frequency region of primary auditory cortex (defined tonotopically) and one from outside of PAC. A tonotopic 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 non-primary 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) (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 non-primary 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 Figure S4 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 primary auditory cortex, defined tonotopically. Only voxels with a reliable response are 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. (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 Figure S5 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 non-primary auditory cortex were defined as the average NSE value across the three bins closest and furthest from PAC (inset). In every subject and hemisphere, we observed larger NSE values in non-primary regions.

We quantified these effects across voxels by plotting the NSE between responses to natural and model-matched sounds for each voxel (Figures 3C). Maps were computed from voxel responses in eight individual subjects who were scanned substantially more than the other subjects (see *Participants* in the 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 (Figure S3), and so we describe them together (see Methods for details; subjects S1, S2, S3, S7, S8 were scanned in Paradigm I; subjects S4, S5, S6 were scanned in Paradigm II; group results are based on data from Paradigm I; S1 participated in both paradigms but only data from Paradigm II is shown). 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 the 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 normalized squared error* in the Methods), but the results were similar without correction (Figure S4).

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 regions far from PAC. 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 non-primary regions, plausibly because non-primary 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 non-primary voxels by binning the NSE of voxels from individual subjects based on their distance to PAC. Similar results were observed for tonotopic (Figure 3D) and anatomical definitions of PAC (Figure S5; PAC was defined either as the center of the high-low-high gradient, or as the center of anatomical region TE1.1⁵⁸, in posteromedial HG). To directly compare primary and non-primary regions we then averaged NSE values within the three bins nearest and furthest from PAC (Figure 3D, inset). This analysis revealed that responses to natural and model-matched sounds became more dissimilar in non-primary regions in both the left and right hemisphere of every subject tested, leading to a highly significant difference between primary and non-primary regions ($p < 0.01$ via sign test for both hemispheres and for both tonotopic and anatomical definitions of PAC). The gradient between primary and non-primary regions was observed in both scanning paradigms regardless of smoothing (Figure S3), 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 Figure S6). 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 non-primary regions (Figure S4C). As a consequence, the increase in the NSE between natural and model-matched sounds between primary and non-primary regions was significantly greater than the change in voxel reliability. This was true using both corrected and uncorrected values for the natural vs. 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 Figure S3 for a breakdown by paradigm). Thus, our results demonstrate that the modulation filter bank model is worse at accounting for voxel responses in non-primary regions.

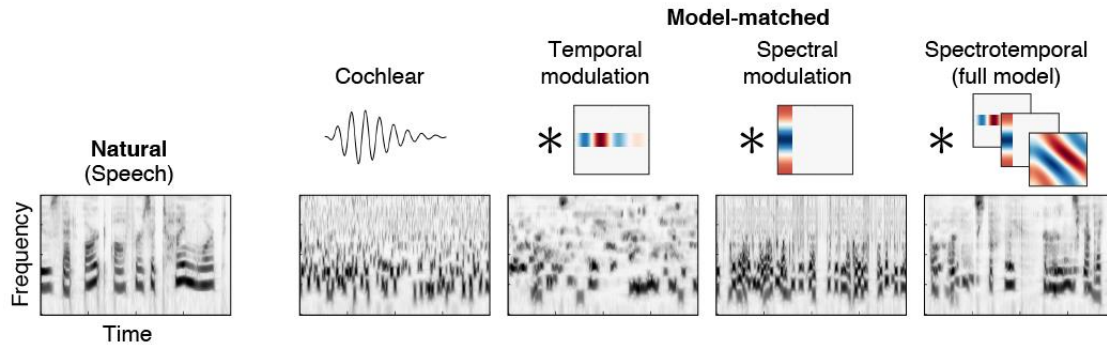
Comparing responses to sounds matched on subsets of model features

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 (Figure 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.

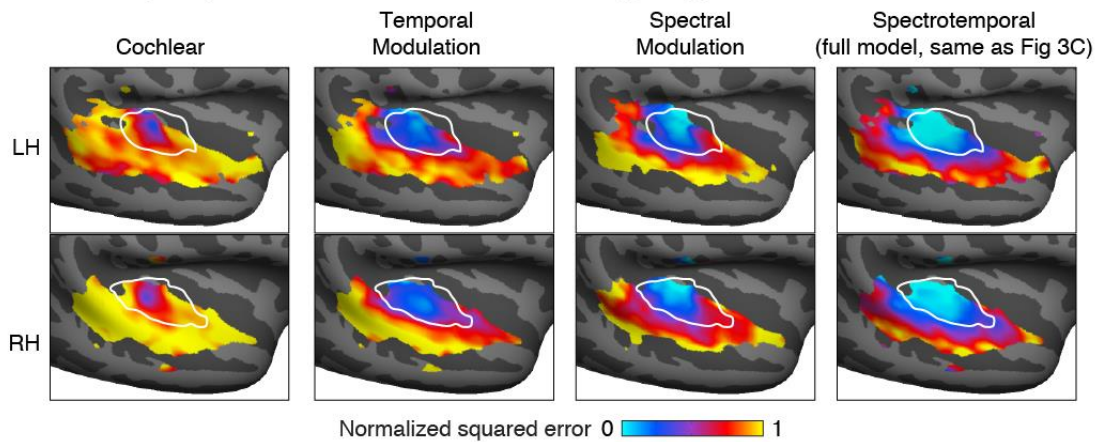
The results of this analysis suggest that all of the model features are necessary to account for voxel responses to natural sounds in primary auditory cortex (Figure 4B&C; Figure S7). Responses to model-matched sounds constrained just by cochlear statistics differed substantially from responses to natural sounds even in PAC, leading to significantly larger NSE values than those observed for the full model ($p < 0.001$ in PAC via bootstrapping across subjects; see *Statistics* in the 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^{7,30,31}. In non-primary 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 absolute NSE values were high for all of the models tested, indicating that the modulation model fails to account for a substantial fraction of non-primary responses.

A Model-matching using subsets of model features



B Dissimilarity maps for subsets of model features (group)



C Dissimilarity vs. distance to PAC (individual subjects, tonotopically defined PAC)

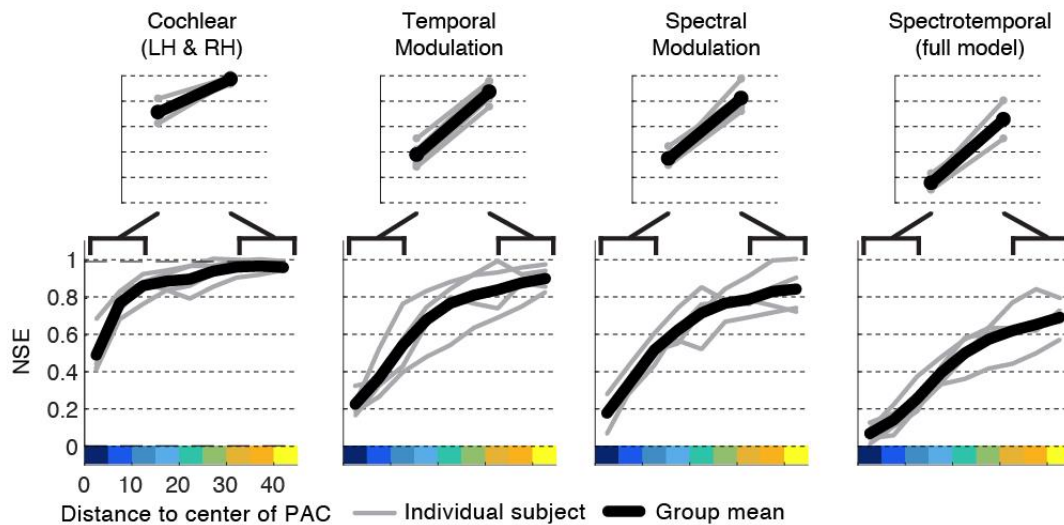


Figure 4. Comparison of responses to model-matched sounds constrained by different models. (A) Cochleagrams for an example natural sound and several corresponding model-matched sounds constrained by subsets of features from the full two-stage model. Cochlear-matched sounds were constrained by time-averaged 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 Figure 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 normalized squared error between responses to natural and model-matched sounds, constrained by each of the four models. Format is the same as panel 3C. See Figure S7 for maps from individual subjects. (C) Dissimilarity between responses to natural and model-matched vs. 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.

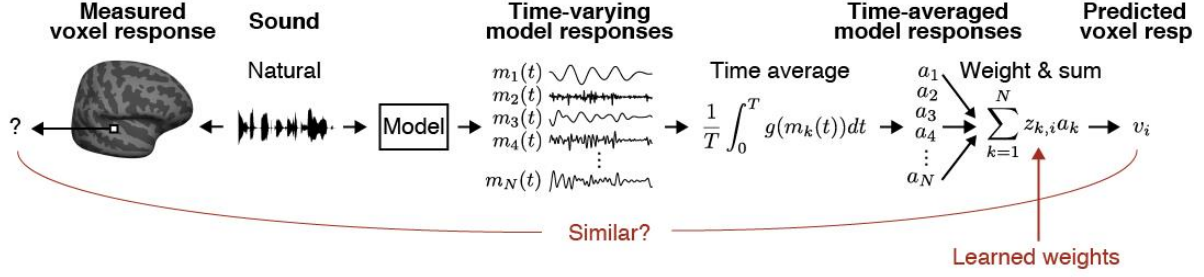
Predicting responses to natural sounds from model features

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, since 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.

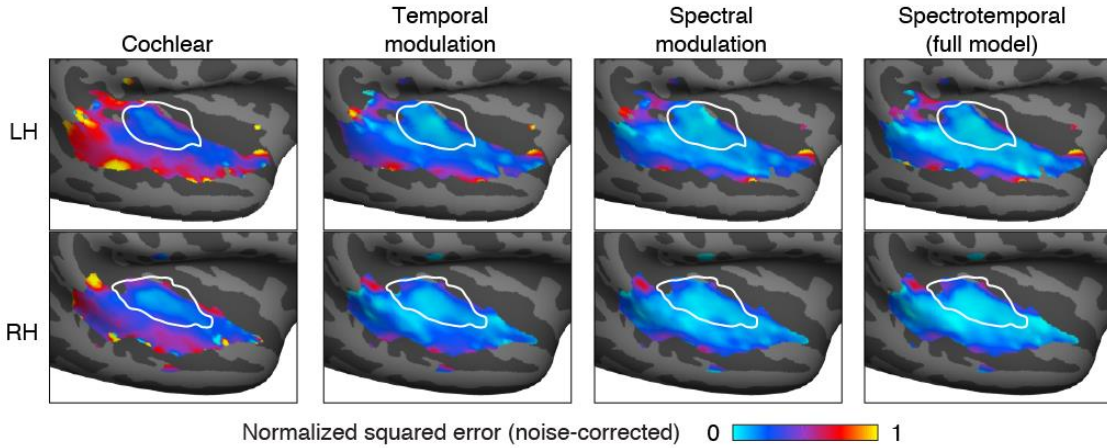
We attempted to predict responses to the 36 natural sounds from time-averaged statistics of the same model features used to generate the model-matched sounds (Figure 5A; see Figure S8 for individual-subject prediction error maps for the full spectrotemporal model). Specifically, we used ridge regression to predict voxel responses from the amplitude of each model feature's response to each natural sound^{6,7}, 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, like 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 from those to natural sounds would thus suggest that prediction-based analyses overestimate of 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 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 (Figure 5B&C), leading to smaller NSEs for model predictions compared with model-matched stimulus responses. This difference was particularly pronounced in non-primary 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 vs. model matching) and region (primary vs. non-primary) ($p < 0.01$ via sign test for both tonotopic and anatomical definitions of PAC). Because the natural and model-matched sounds were matched in the features used for prediction, the divergent responses to the two sound sets implies 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 functional organization not clearly evident from model predictions, by demonstrating the failure of the filter bank model to account for non-primary responses.

A Schematic of model prediction via regression



B Model prediction errors (group)



C Prediction and model-matching errors vs. distance to PAC (tonotopically defined)

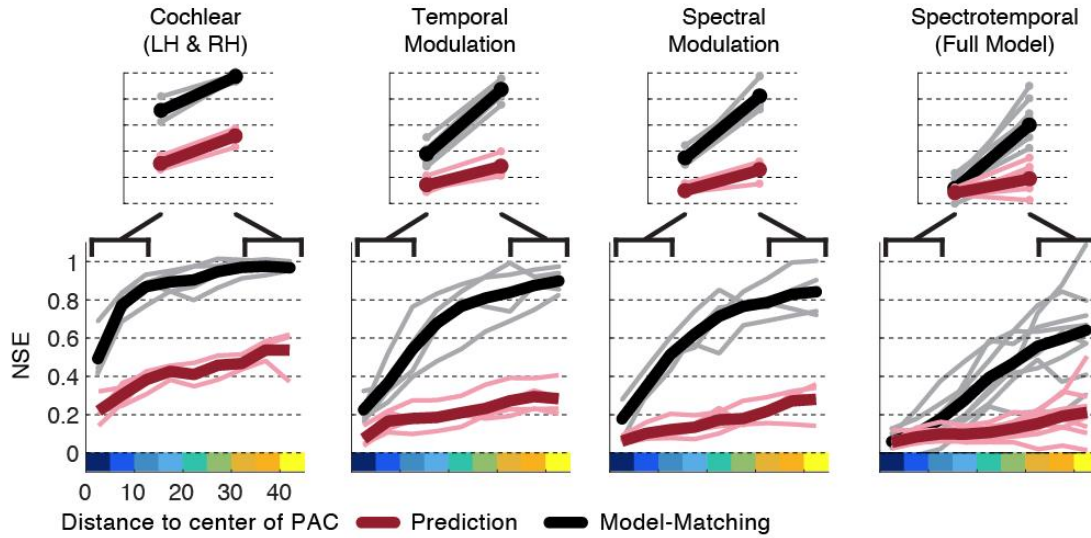


Figure 5. Predicted responses to natural sounds via regression using the same auditory model used to constrain the model-matched sounds. (A) Schematic of regression procedure used to predict neural responses from model features. For each natural sound, we computed the response timecourse for each activity 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 Figures 3&4) between measured and predicted responses to natural sounds for the corresponding models shown in Figure 4 (see Figure S8 for maps from individual subjects). (C) Prediction error vs. 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 re-plotted from Figure 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 8 subjects shown in Figure 3C. Results for model subsets (cochlear, temporal modulation and spectral modulation) are based on data from 4 subjects that were scanned in Paradigm I (sounds constrained by subsets of model features were not tested in Paradigm II).

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^{7,45,60}. 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 sounds⁴⁵. We observed similar results with this larger sound set, with relatively good prediction accuracy for the full spectrotemporal model throughout primary and non-primary auditory cortex (Figure S9).

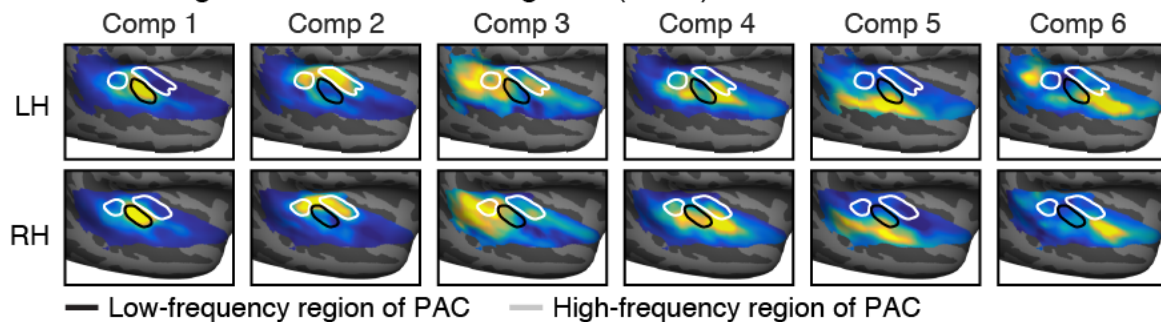
Another way to assess the utility of the model-matching approach is to train a model to predict natural sounds, and then test its predictive accuracy on model-matched sounds (and vice-versa). In practice, this approach yielded similar results to directly comparing responses to natural and model-matched sounds: good cross-predictions in primary auditory cortex, but poor cross-predictions in non-primary auditory cortex (Figure S10). 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 non-primary regions, but it provides a consistency check of the two types of analyses.

Voxel decomposition of responses to natural and model-matched sounds

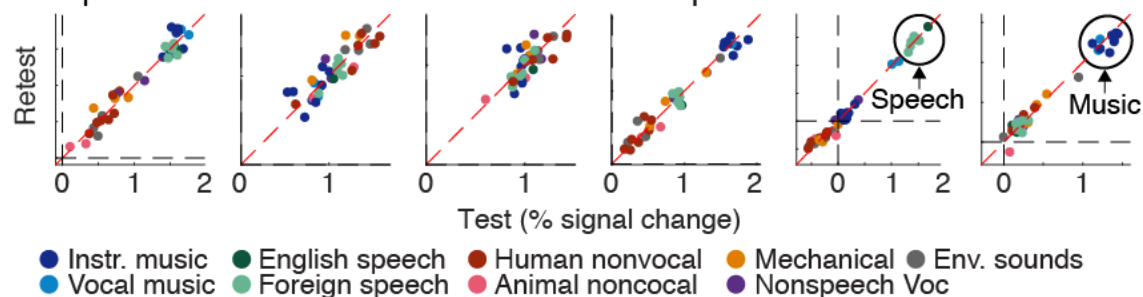
All of our analyses described thus far were performed on individual voxels, summarized with maps plotting the normalized squared error between each voxel's response to natural and model-matched sounds. But these error maps do not reveal in what respect the responses to natural and model-matched sounds differ, and due to 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 (Figure 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 the Methods).

Voxel decomposition of responses to natural and model-matched sounds

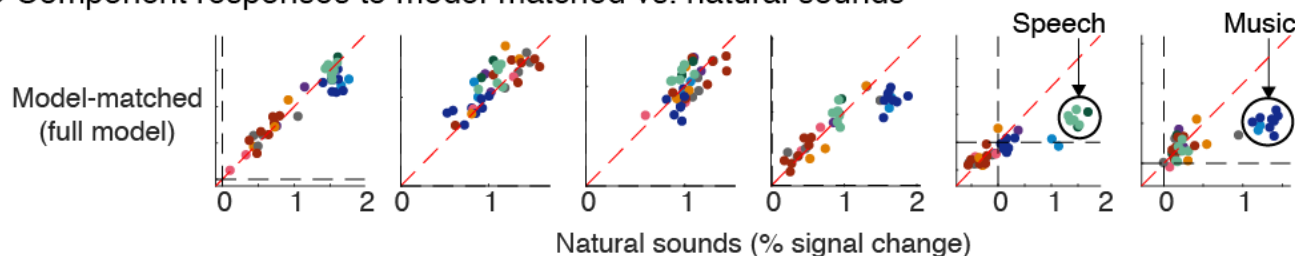
A Component voxel weights from Norman-Haignere (2015)



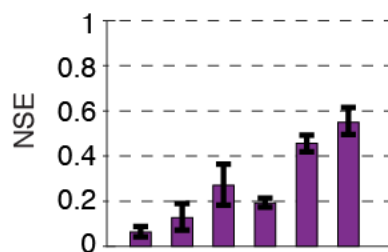
B Component responses to the natural sounds from this experiment



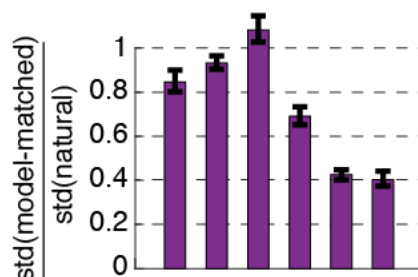
C Component responses to model-matched vs. natural sounds



D Overall dissimilarity



E Ratio of response variation



F Correlation

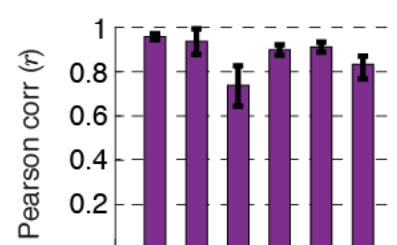


Figure 6. Voxel decomposition of responses to natural and model-matched sounds. Previously we found that much of the voxel response variance to natural sounds can be approximated as a weighted sum of six canonical response patterns (“components”) ⁴⁵. This figure shows the response of these components to the natural and model-matched sounds from this experiment. (A) The group components weights from Norman-Haignere et al. (2015) ⁴⁵ are re-plotted to show where in the 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 due to its selectivity for sounds with pitch). (C) Component responses to natural and model-matched sounds constrained by the complete spectrotemporal model (see Figure S11 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) Normalized squared error 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 Figure S12A for corresponding whole-brain maps). (F) Pearson correlation of responses to natural and model-matched sounds (see Figure S12B 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.

We found that all six components exhibited reliable responses to the natural sounds from this experiment (Figure 6B). Two of the components (5 & 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 (Figure 6C, bottom row, last two columns; see Figure S11 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) (Figure 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 & 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, due to lower overall responses to the model-matched sounds (Figure 6E). In contrast, the noise-corrected correlation remained high (Figure 6F). A similar pattern was also evident in whole-brain maps (Figure S12): the variation in voxel responses to model-matched sounds constrained by the full model dropped in non-primary regions (driven by lower responses to the “preferred” model-matched stimuli) while the correlation remained high. For Components 5 & 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 Figure 6D). For example, in Component 6, model-matched music produced a lower response than natural music, but a higher response than model-matched sounds from other categories ($p < 0.001$; via bootstrapping). The same pattern was evident for Component 5, which responded selectively to speech ($p < 0.001$). This finding suggests that selectivity in non-primary 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^{53,61}. The results suggest that the modulation-specific structure driving Components 5 & 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 & 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 non-primary 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 non-primary 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 non-primary 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^{40,62–64}. Hierarchical organization has some support from anatomical studies⁶⁵, and from qualitative observations that responses in non-primary regions are more complex than those in primary regions^{66,67} and more closely aligned with semantically meaningful sound properties^{16,45–47,68}. However, there has been little evidence for how primary and non-primary regions might differ in computational terms⁴¹, and thus it has been unclear what mechanisms underlie the apparent differences in tuning between primary and non-primary regions.

Most computational models of auditory processing beyond the periphery are based on tuning for modulation^{19,20}. Such models have been used to explain responses throughout the auditory pathway in non-human animals^{2,11,24–29,32,33}. In humans, modulation-based models have been shown to have relatively good predictive accuracy throughout both primary and non-primary regions^{7,12,45}, which has led to the hypothesis that sounds are represented in a distributed manner⁶⁹. 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 non-primary 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 these two prior literatures. First, we show that modulation selectivity fails to explain much of the response in non-primary 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 non-primary regions whose response to natural sounds was well predicted by the model and yet produced divergent responses to model-matched sounds. Since the model by definition predicts that natural and model-matched sounds should have equivalent responses, this finding demonstrates a clear model failure. Since our predictions were based on linear regression, there must be correlations across natural sounds between the model's features and whatever sound properties actually drive the voxel, and these correlations are what allow the model to accurately predict the voxel responses to natural sounds.

Conversely, our findings provide further evidence that modulation selectivity is a key feature of functional organization in human PAC^{7,30,31}. 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, since the predictions could have been influenced by stimulus-driven correlations, as turned out to be the case in non-primary 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 non-primary regions of auditory cortex represent? These regions are primarily driven by sound, show signs of having relatively short integration windows⁴³, and even when speech selective, respond largely independently of the presence of linguistic structure^{43,45}, suggesting acoustic rather than linguistic or semantic representations¹⁶. 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 non-primary regions transform a lower-level acoustic representation like the spectrotemporal representation considered here, into an acoustic 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 operations⁷⁰. 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, by optimizing for efficient encoding of natural sounds^{71,72}, or for performance of ecologically relevant tasks^{73–78}.

We have recently explored this idea by training a deep neural network to recognize words and musical genres⁴¹ and then comparing the resulting representations to voxel responses. We found that later layers of the network better predicted voxels in non-primary 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 primary auditory cortex, 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 (Figure S15), 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 the Methods). This finding suggests that a wide range of spectrotemporal filter models can be ruled out as models of non-primary auditory cortex. Our stimuli and fMRI data are available on our lab website, so that alternative models can be evaluated using this approach:

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 Figure 4 – the full spectrotemporal model is inclusive of the other models. The weaker correlations 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. Since 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 primary auditory cortex^{13,79–81}. 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^{48,82}, and later applied to sound texture³⁹ and visual texture perception^{49,83}. 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 methods⁴⁸, 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 that listeners are sensitive to (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 not does 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 non-primary 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 model⁸⁴. 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 & Simoncelli^{39,85}. 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)³⁹.

Approaches for model testing

Recent years have seen growing interest in the use of computational “encoding models” to test formal theories of sensory processing^{5–7,10–14,16,17,73,86}. 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^{6,7,12,16}.

A key challenge of testing encoding models with natural stimuli is that the features of different models are often correlated^{16,18}, 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 sets⁷³. 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 de-correlated 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.

Author Contributions

SNH and JHM designed the stimuli and experiment together and wrote the paper. SNH implemented all of the experimental procedures, synthesis algorithms, and analyses.

Competing Financial Interests Statement

The authors declare no competing financial interests.

Methods

Participants

The experiment comprised 41 scanning sessions, each approximately two hours. Fifteen subjects participated in the experiment (ages 19-36; 5 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 4th and 5th sessions). We chose this approach because it allowed us to compute reliable group maps by averaging across the twelve 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 the auditory cortex of typical listeners without extensive musical experience, we required that subjects not have received formal musical training in the five years preceding their participation in the experiment. The study was approved by MIT's Committee on the Use of Humans as Experimental Subjects; all subjects gave informed consent.

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 s TR, 30 ms TE, 90 degree flip angle; 5 discarded initial acquisitions). Each slice was 4 mm thick and had an in-plane resolution of 2.1 x 2.1 mm (96 x 96 matrix, 0.4 mm slice gap). iPAT was used to minimize acquisition time (1 sec/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 (3x acceleration) to reduce slice thickness while maintaining coverage (36 slices with 2mm thickness and no gap), and thus reducing voxel size (2 mm isotropic). iPAT was not used. Other acquisition parameters were similar (3.45 s TR, 1.05 sec acquisition time, 34 ms TE, 90 degree flip angle; 3 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 FLIRT⁸⁷ and BBRegister⁸⁸). Volume data were then resampled to the reconstructed cortical surface computed by FreeSurfer⁸⁹, and smoothed on the surface using a 5mm FWHM kernel to improve SNR (results were similar without smoothing; Figure S3). 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 (Figure 2B). From each natural sound, we synthesized four model-matched sounds, constrained by different subsets of features from a model of auditory cortex based on modulation filtering¹⁹. The complete stimulus set

thus included 5 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 paradigm⁹⁰ that alternated between presenting sounds and acquiring scans, similar to those used in our prior experiments^{45,91,92} (Figure S2). This was achieved by dividing each 10-second stimulus into five 2-second segments (windowed with 25 ms 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 ms 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 response of the 2nd through 5th acquisition 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 timepoint or just the fourth and fifth timepoint after stimulus onset, indicating that our results were robust to the averaging window applied to the fMRI measurements (Figure S13). 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 HRF⁹³.

In Paradigm I, each model-matched stimulus was presented once per two-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 normalized squared error metric. Each session was divided into 12 "runs", after which, subjects were given a short break (~30 seconds). Each run included 6 natural sounds and 12 model-matched sounds (3 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 9 natural and 9 model-matched sounds. The entire sound set was presented over 4 consecutive runs. Subjects completed 12 or 16 runs depending on the time constraints of the scan session. Thus, each subject heard each sound between 3 or 4 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^{91,92}).

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 3 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 9 model-matched stimuli presented per run (there were only 3 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. But 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 sounds⁴⁵. In our prior study, we found that much of the 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.

The cochlear representation was computed by convolving the audio waveform of each sound with 120 bandpass filters, spaced equally on an 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%)³⁹. 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 Chi et al., 2005¹⁹, 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 multi-scale wavelet filters (Figure 1A). The shapes and bandwidths of the filters were the same as those described by Chi et al. (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:

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

where b_r determines the best modulation rate of the filter (i.e. the rate with maximum gain). We used 9 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 9 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:

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

where b_s determines the best modulation scale of the filter (i.e. the scale with maximum gain). We used 6 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 6 filtered cochleagrams, each of which highlights a different range of spectral modulations.

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 described in Chi et al., 2005¹⁹) by taking the outer-product of the frequency-domain representations of the temporal and spectral modulation filters described above. 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 9 temporal filters with 6 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 which 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^{39,48,82}, in our case provided by the histogram of each feature’s response (Figure S1). 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: (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 feature⁴⁸ (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. 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), 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 they 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). Since 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 re-assigning the smallest value in the signal-to-be-matched to the smallest value in the target signal, then re-assigning 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 which have been ordered according to the magnitude of the corresponding vector elements (i.e. the indices that would sort the vector, as in the numpy/python function argsort). 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^{19,39}. 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). 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). 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 uniform¹⁹.

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 re-analyzed 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 (analyze, histogram-match, reconstruct) generally results in increasingly close matches^{39,48,82}. 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 wrap-around 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 (8 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.

MATLAB code for synthesis algorithm will be made available upon publication.

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. Figure S14 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 (Figure S14, 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 (Figure S15). 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. Narrowing the filter bandwidths was accomplished by halving the temporal and spectral extent of the filters in the frequency domain as well as doubling the number of filters to ensure that all modulation rates and scales were represented by the model. 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 Figure S15), 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 Figure S15). 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:

$$(5) \quad \mu([\mathbf{x} - \mathbf{y}]^2) = \frac{1}{N} \sum_i^N (x_i - y_i)^2$$

where $\mathbf{x}$ and $\mathbf{y}$ 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):

$$(6) \quad \epsilon(\mathbf{x}, \mathbf{y}) = \frac{\mu[(\mathbf{x} - \mathbf{y})^2]}{\mu(\mathbf{x}^2) + \mu(\mathbf{y}^2) - 2\mu(\mathbf{x})\mu(\mathbf{y})}$$

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:

$$(7) \quad \begin{aligned} \mathbf{E}[(x - y)^2] &= \mathbf{E}[x^2 + y^2 - 2xy] \\ &= \mathbf{E}[x^2] + \mathbf{E}[y^2] - 2\mathbf{E}[xy] \\ &= \mathbf{E}[x^2] + \mathbf{E}[y^2] - 2\mathbf{E}[x][y] \end{aligned}$$

Noise-correcting the normalized squared error

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 have 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 voxels responses in both primary and non-primary regions were similarly reliable (Figure S4).

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:

$$(8) \quad \mathbf{x} = \mathbf{s}_x + \mathbf{n}_x$$

$$(9) \quad \mathbf{y} = \mathbf{s}_y + \mathbf{n}_y$$

where in the context of this experiment, $\mathbf{x}$ and $\mathbf{y}$ are the measured response of a voxel to two sets of sounds (i.e. natural and model-matched sounds), $\mathbf{s}_x$ and $\mathbf{s}_y$ are the stimulus-driven responses, and $\mathbf{n}_x$ and $\mathbf{n}_y$ 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:

$$(10) \quad \mathbf{x}_1 = \mathbf{s}_x + \mathbf{n}_{x1}$$

$$(11) \quad \mathbf{x}_2 = \mathbf{s}_x + \mathbf{n}_{x2}$$

$$(12) \quad \mathbf{y}_1 = \mathbf{s}_y + \mathbf{n}_{y1}$$

$$(13) \quad \mathbf{y}_2 = \mathbf{s}_y + \mathbf{n}_{y2}$$

We would like to estimate the NSE of the stimulus-driven component of the uncorrupted responses:

$$(14) \quad \epsilon(\mathbf{s}_x, \mathbf{s}_y) = \frac{\mu[(\mathbf{s}_x - \mathbf{s}_y)^2]}{\mu(\mathbf{s}_x^2) + \mu(\mathbf{s}_y^2) - 2\mu(\mathbf{s}_x)\mu(\mathbf{s}_y)} \\ = \frac{\mu[\mathbf{s}_x^2] + \mu[\mathbf{s}_y^2] - 2\mu[\mathbf{s}_x\mathbf{s}_y]}{\mu(\mathbf{s}_x^2) + \mu(\mathbf{s}_y^2) - 2\mu(\mathbf{s}_x)\mu(\mathbf{s}_y)}$$

But are restricted to measuring the NSE of the noise-corrupted response. From the equation above, it is evident that the NSE depends on three types of statistics: (1) the signal powers ($\mu[\mathbf{s}_x^2]$ and $\mu[\mathbf{s}_y^2]$) (2) the signal cross-product ($\mu[\mathbf{s}_x\mathbf{s}_y]$) and (3) the signal means ($\mu[\mathbf{s}_x]$). The signal means are unbiased by the noise, since by assumption the noise is zero mean. The signal cross-product is also unbiased by noise:

$$(15) \quad \mathbf{E}[(s_x + n_x)(s_y + n_y)] = \mathbf{E}[s_x s_y] + \mathbf{E}[s_x n_y] + \mathbf{E}[s_y n_x] + \mathbf{E}[n_x n_y] \\ = \mathbf{E}[s_x s_y]$$

(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:

$$(16) \quad \hat{\mu}[\mathbf{s}_x \mathbf{s}_y] = \frac{1}{4}\mu[\mathbf{x}_1 \mathbf{y}_1] + \frac{1}{4}\mu[\mathbf{x}_1 \mathbf{y}_2] + \frac{1}{4}\mu[\mathbf{x}_2 \mathbf{y}_1] + \frac{1}{4}\mu[\mathbf{x}_2 \mathbf{y}_2]$$

$$(17) \quad \hat{\mu}[\mathbf{s}_x] = \frac{1}{2}\mu[\mathbf{x}_1] + \frac{1}{2}\mu[\mathbf{x}_2]$$

$$(18) \quad \hat{\mu}[\mathbf{s}_y] = \frac{1}{2}\mu[\mathbf{y}_1] + \frac{1}{2}\mu[\mathbf{y}_2]$$

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

$$(19) \quad \mathbf{E}[(s_x + n_x)^2] = \mathbf{E}[s_x^2] + \mathbf{E}[n_x^2] + 2\mathbf{E}[s_x n_x] \\ = \mathbf{E}[s_x^2] + \mathbf{E}[n_x^2]$$

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:

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

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

$$(21) \quad \hat{\mu}[\mathbf{s}_x^2] = \frac{1}{2}\mu[\mathbf{x}_1^2] + \frac{1}{2}\mu[\mathbf{x}_2^2] - \frac{1}{2}\mu[(\mathbf{x}_1 - \mathbf{x}_2)^2]$$

$$(22) \quad \hat{\mu}[\mathbf{s}_y^2] = \frac{1}{2}\mu[\mathbf{y}_1^2] + \frac{1}{2}\mu[\mathbf{y}_2^2] - \frac{1}{2}\mu[(\mathbf{y}_1 - \mathbf{y}_2)^2]$$

Substituting equations 16-18 and 21-22 into equation 14 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 (Figure S16), 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 equation 15), which is not the case for fMRI measurements close in time (i.e. <5 seconds)⁹⁴. Here, each measurement corresponds to the average response of the 2nd through 5th 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 signal⁹⁴. 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 where 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 ($\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{y}_1$) for each voxel ($\mathbf{x}_1$ and $\mathbf{x}_2$ corresponding to the voxel's response to natural sounds, and $\mathbf{y}_1$ to the response to model-matched sounds). Each vector was computed as the weighted combination of a true, noise-free signal ($\mathbf{s}_x$, $\mathbf{s}_y$) that was constant across repeated measurements plus additive noise that varied across measurements ($\mathbf{n}_{x1}$, $\mathbf{n}_{x2}$, $\mathbf{n}_{y1}$):

$$(23) \quad \mathbf{x}_1 = \mathbf{s}_x b + \mathbf{n}_{x1}(1 - b)$$

$$(24) \quad \mathbf{x}_2 = \mathbf{s}_x b + \mathbf{n}_{x2}(1 - b)$$

$$(25) \quad \mathbf{y}_1 = \mathbf{s}_y b + \mathbf{n}_{y1}(1 - b)$$

We used the weights (b) to control the SNR of the voxel with weights closer to 1 resulting in higher SNR. 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 ($\text{var}(\mathbf{n}_x) = \text{var}(\mathbf{n}_y)$), which we have verified is a reasonable assumption for our data (Figure S16). 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 signal⁹⁴. 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 ($\mathbf{s}_x$ and $\mathbf{s}_y$), 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 $\mathbf{s}_x$ and $\mathbf{s}_y$ as the weighted sum of a shared response vector ($\mathbf{g}$) and a distinct vector unique to x and y ($\mathbf{u}_x$, $\mathbf{u}_y$):

$$(26) \quad \mathbf{s}_x = \mathbf{g}c + \mathbf{u}_x(1 - c)$$

$$(27) \quad \mathbf{s}_y = \mathbf{g}c + \mathbf{u}_y(1 - c)$$

Thus by varying c , we could vary the similarity of the voxel responses. We sampled c from a uniform distribution, and we sampled $\mathbf{g}$, $\mathbf{u}_x$, and $\mathbf{u}_y$ 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:

$$(28) \quad \mathbf{s}_x = \mathbf{g}$$

$$(29) \quad \mathbf{s}_y = \mathbf{g}c$$

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 $\mathbf{x}_1$ and $\mathbf{y}_1$ for simplicity).
- (2) the NSE between the true signals ($\mathbf{s}_x$ and $\mathbf{s}_y$), 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 ($\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{y}_1$).
- (4) the NSE between two independent measurements of the same stimulus (“test-retest”), which provides a measure of the voxel’s noise level ($\mathbf{x}_1$, $\mathbf{x}_2$).

In Figure S17A&B, we plot the results from Simulation 1. First, we plot the NSE of the noise-corrupted data vs. the NSE of the true signals (Figure S17A, 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 vs. the NSE values for the true signals (Figure S17A, 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 Figure S17B, which plots the median NSE of both the noise-corrupted and noise-corrected values along with its 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 (Figure S17C&D): 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 where noise correction is well-behaved. In Paradigm I, we measured reliability using natural sounds, since 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 4 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, 9, split 2 included repetitions 2, 6, 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 (Figure S18).

Noise-correcting response variation and correlation measures

In addition to the normalized squared error, 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 (Figure

6E&F and Figure S12). 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 (equation 19), 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:

$$(30) \quad \text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{\sqrt{\text{var}(\mathbf{x})\text{var}(\mathbf{y})}}$$

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 (equation 15), 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^{45,95} (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). MATLAB code implementing these noise-correction procedures can be downloaded here:

<https://github.com/snormanhaignere/general-analysis-code>
see noise_corrected_similarity.m

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 non-primary regions where we observed weak responses to the model-matched sounds. Thus, it was necessary to exclude voxels that did not have test-retest correlation to model-matched sounds of at least 0.4, which caused many non-primary voxels to be excluded in the maps of Figure S12. 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 non-zero variance.

Model predictions with natural sounds

Our model assumes that voxels are a weighted sum of time-averaged statistics of the feature responses (equations 1-2). 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 which are 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 (where all of the modulation-matched sounds were also matched in their cochlear statistics). For the temporal modulation model, there were a total of 2170 features (9 rates x 217 audio frequencies + 217 cochlear channels). For the spectral modulation model there were 1736 features (7 scales x 217 frequencies + 217 cochlear channels). For the spectrotemporal modulation model, there were 27559 features (9 rates x 7 scales x 2 orientations x 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^{7,16}. 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 ridge⁴¹. 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 re-scaled 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 was used to fit the weights and select the regularization parameter (details below), and the test data was used to evaluate the predictions of the model. We used 4-fold cross-validation, splitting the data into four equally sized sets of 9 sounds. For each set, we used the remaining sounds to 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 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.m`

`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 predictions⁴¹. 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 (equations 16-18 and 21-22). 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 a six canonical response patterns (components)⁴⁵:

$$(31) \quad \mathbf{v}_i \approx \sum_{k=1}^6 \mathbf{r}_k w_{k,i}$$

where $\mathbf{v}_i$ and $\mathbf{r}_k$ 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 ($\mathbf{r}_k$) and weights ($w_{k,i}$) were jointly inferred by maximizing the non-Gaussianity of the weights, similar to classical independent component analysis⁹⁶. Figure 6A re-plots 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 where 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 8 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 modulation-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:

$$(32) \quad \mathbf{r}_k \approx \sum_i \mathbf{v}_i u_{k,i}$$

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 timecourse divided by its standard deviation; results were similar when the analysis was restricted to voxels from the superior temporal plane / 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 equation 31) as our reconstruction weights⁴⁵, 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 literature⁹⁷.

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 Table S1; defined by inspection using the scheme described in Da Costa et al. (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, 1600, 3200, & 6400). We measured the frequency range that produced the maximum response in voxels significantly modulated by frequency ($p < 0.05$ in a 1-way ANOVA across the 6 ranges); the details of the stimuli and analyses have been described previously⁹¹. Group tonotopy maps were based on a cohort of 21 subjects who were run in this tonotopy localizer across multiple studies (6 of the subjects from this experiment were part of this cohort)⁹⁸. 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 (Figure 6) were based on the annular analyses described above. We defined primary and non-primary 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 non-primary 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 non-primary 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 (Figure S7), 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 1000 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 1000 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 (yielding the reported 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 anatomy⁹⁹. 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 studies¹⁰⁰ but is less frequently applied to fMRI analyses¹⁰¹. We performed stats on the component responses by estimating the response of each component in each subject, and then bootstrapping across the eight subjects in which we had component data. We randomly sampled 8 subjects with replacement 1000 times, averaged the component responses for those subjects, and recomputed the noise-corrected NSE. We then contrasted NSE values between components or conditions to assess significance.

References

1. Simoncelli, E. P. & Olshausen, B. A. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24, 1193–1216 (2001).
2. Woolley, S. M., Fremouw, T. E., Hsu, A. & Theunissen, F. E. Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nat. Neurosci.* 8, 1371–1379 (2005).
3. Smith, E. C. & Lewicki, M. S. Efficient auditory coding. *Nature* 439, 978–982 (2006).
4. Sharpee, T., Rust, N. C. & Bialek, W. Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Comput.* 16, 223–250 (2004).
5. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *Neuroimage* 56, 400–410 (2011).
6. Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76, 1210–1224 (2012).
7. Santoro, R. *et al.* Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLoS Comput. Biol.* 10, (2014).
8. Theunissen, F. E. & Elie, J. E. Neural processing of natural sounds. *Nat. Rev. Neurosci.* 15, 355–366 (2014).
9. Di Liberto, G. M., O’Sullivan, J. A. & Lalor, E. C. Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Curr. Biol.* 25, 2457–2465 (2015).
10. Kriegeskorte, N. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annu. Rev. Vis. Sci.* 1, 417–446 (2015).
11. Thorson, I. L., Liénard, J. & David, S. V. The Essential Complexity of Auditory Receptive Fields. *PLOS Comput. Biol.* 11, e1004628 (2015).
12. Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E. & Chang, E. F. Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *J. Neurosci.* 36, 2014–2026 (2016).
13. Kozlov, A. S. & Gentner, T. Q. Central auditory neurons have composite receptive fields. *Proc. Natl. Acad. Sci.* 113, 1441–1446 (2016).
14. Willmore, B. D. B., Schoppe, O., King, A. J., Schnupp, J. W. H. & Harper, N. S. Incorporating Midbrain Adaptation to Mean Sound Level Improves Models of Auditory Cortical Processing. *J. Neurosci.* 36, 280–289 (2016).
15. Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365 (2016).
16. de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L. & Theunissen, F. E. The hierarchical cortical organization of human speech processing. *J. Neurosci.* 3267–16 (2017).
17. Meyer, A. F., Williamson, R. S., Linden, J. F. & Sahani, M. Models of Neuronal Stimulus-Response Functions: Elaboration, Estimation, and Evaluation. *Front. Syst. Neurosci.* 10, (2017).
18. Groen, I. I. *et al.* Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife* 7, e32962 (2018).
19. Chi, T., Ru, P. & Shamma, S. A. Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* 118, 887–906 (2005).
20. Dau, T., Kollmeier, B. & Kohlrausch, A. Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *J. Acoust. Soc. Am.* 102, 2892–2905 (1997).
21. Elliott, T. M. & Theunissen, F. E. The modulation transfer function for speech intelligibility. *PLoS Comput Biol* 5, e1000302 (2009).
22. Patil, K., Pressnitzer, D., Shamma, S. & Elhilali, M. Music in our ears: the biological bases of musical timbre perception. *PLoS Comput Biol* 8, e1002759 (2012).

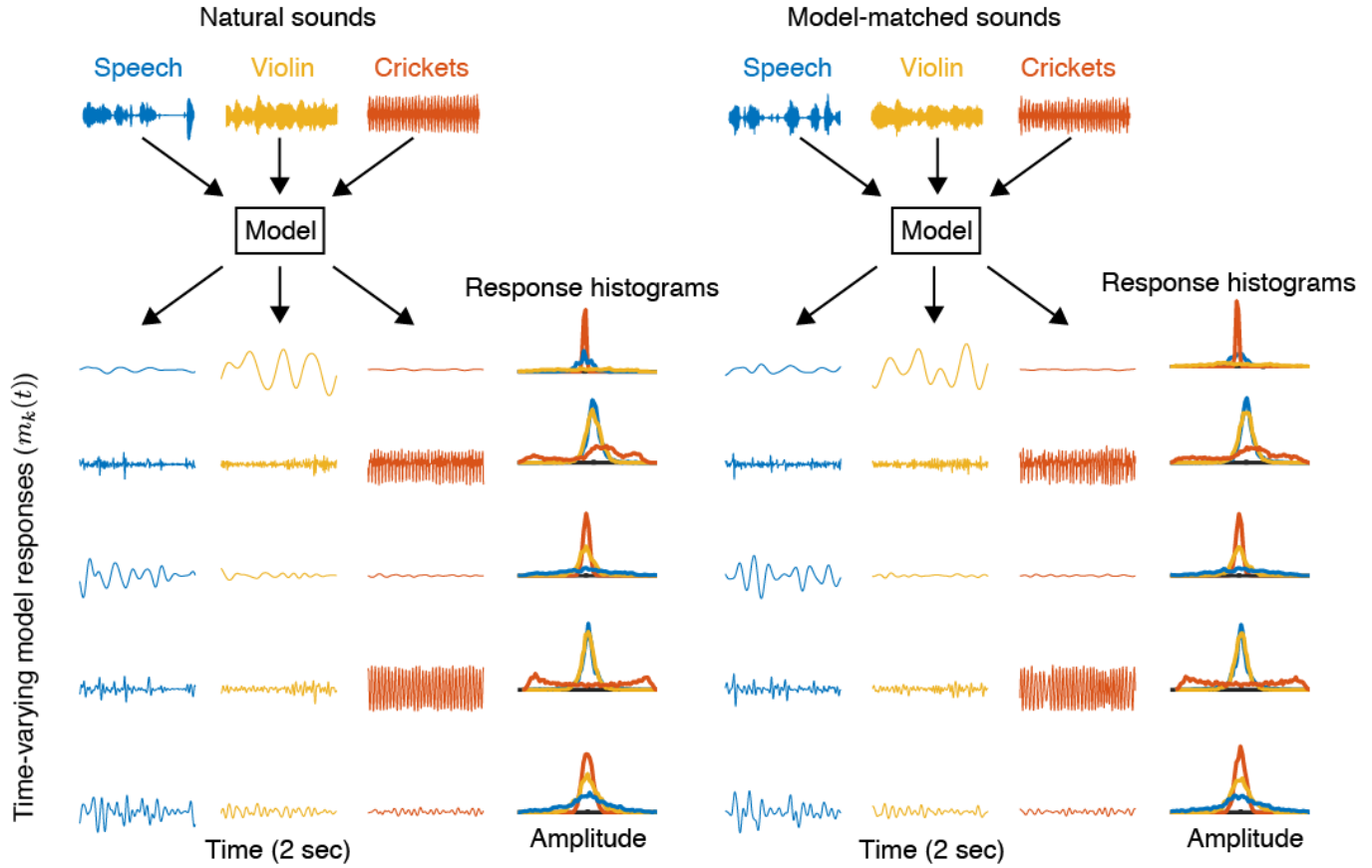
23. McDermott, J. H., Schemitsch, M. & Simoncelli, E. P. Summary statistics in auditory perception. *Nat. Neurosci.* 16, 493–498 (2013).
24. deCharms, C. R., Blake, D. T. & Merzenich, M. M. Optimizing sound features for cortical neurons. *science* 280, 1439–1444 (1998).
25. Theunissen, F. E., Sen, K. & Doupe, A. J. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J. Neurosci.* 20, 2315–2331 (2000).
26. Depireux, D. A., Simon, J. Z., Klein, D. J. & Shamma, S. A. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *J. Neurophysiol.* 85, 1220–1234 (2001).
27. Miller, L. M., Escabí, M. A., Read, H. L. & Schreiner, C. E. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *J. Neurophysiol.* 87, 516–527 (2002).
28. Nelson, P. C. & Carney, L. H. A phenomenological model of peripheral and central neural responses to amplitude-modulated tones. *J. Acoust. Soc. Am.* 116, 2173–2186 (2004).
29. Fishman, Y. I. & Steinschneider, M. Temporally dynamic frequency tuning of population responses in monkey primary auditory cortex. *Hear. Res.* 254, 64–76 (2009).
30. Schönwiesner, M. & Zatorre, R. J. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proc. Natl. Acad. Sci.* 106, 14611–14616 (2009).
31. Barton, B., Venezia, J. H., Saberi, K., Hickok, G. & Brewer, A. A. Orthogonal acoustic dimensions define auditory field maps in human cortex. *Proc. Natl. Acad. Sci.* 109, 20738–20743 (2012).
32. Chechik, G. & Nelken, I. Auditory abstraction from spectro-temporal features to coding auditory entities. *Proc. Natl. Acad. Sci.* 109, 18968–18973 (2012).
33. Rabinowitz, N. C., Willmore, B. D., Schnupp, J. W. & King, A. J. Spectrotemporal contrast kernels for neurons in primary auditory cortex. *J. Neurosci.* 32, 11271–11284 (2012).
34. Herdener, M. *et al.* Spatial representations of temporal and spectral sound cues in human auditory cortex. *Cortex* 49, 2822–2833 (2013).
35. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006–1010 (2014).
36. 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* 14, 920–930 (2006).
37. Greenberg, S., Carvey, H., Hitchcock, L. & Chang, S. Temporal properties of spontaneous speech—a syllable-centric perspective. *J. Phon.* 31, 465–485 (2003).
38. Ding, N. *et al.* Temporal modulations in speech and music. *Neurosci. Biobehav. Rev.* (2017).
39. McDermott, J. H. & Simoncelli, E. P. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron* 71, 926–940 (2011).
40. Wessinger, C. M. *et al.* Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *J. Cogn. Neurosci.* 13, 1–7 (2001).
41. Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V. & McDermott, J. H. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron* (2018).
42. Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P. & Pike, B. Voice-selective areas in human auditory cortex. *Nature* 403, 309–312 (2000).
43. Overath, T., McDermott, J. H., Zarate, J. M. & Poeppel, D. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat. Neurosci.* 18, 903–911 (2015).
44. Binder, J. R. *et al.* Human Temporal Lobe Activation by Speech and Nonspeech Sounds. *Cereb. Cortex* 10, 512–528 (2000).
45. Norman-Haignere, S. V., Kanwisher, N. G. & McDermott, J. H. Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron* 88, 1281–1296 (2015).
46. Leaver, A. M. & Rauschecker, J. P. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *J. Neurosci.* 30, 7604–7612 (2010).

47. Angulo-Perkins, A. *et al.* Music listening engages specific cortical regions within the temporal lobes: Differences between musicians and non-musicians. *Cortex* 59, 126–137 (2014).
48. Heeger, D. J. & Bergen, J. R. Pyramid-based texture analysis/synthesis. in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques* 229–238 (ACM, 1995).
49. Freeman, J. & Simoncelli, E. P. Metamers of the ventral stream. *Nat. Neurosci.* 14, 1195–1201 (2011).
50. Balas, B. J. Texture synthesis and perception: Using computational models to study texture representations in the human visual system. *Vision Res.* 46, 299–309 (2006).
51. Humphries, C., Liebenthal, E. & Binder, J. R. Tonotopic organization of human auditory cortex. *NeuroImage* 50, 1202–1211 (2010).
52. Costa, S. D. *et al.* Human primary auditory cortex follows the shape of heschl's gyrus. *J. Neurosci.* 31, 14067–14075 (2011).
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.* 32, 14205–14216 (2012).
54. Baumann, S., Petkov, C. I. & Griffiths, T. D. A unified framework for the organization of the primate auditory cortex. *Front. Syst. Neurosci.* 7, 11 (2013).
55. Leaver, A. M. & Rauschecker, J. P. Functional topography of human auditory cortex. *J. Neurosci.* 36, 1416–1428 (2016).
56. Schönwiesner, M., Dechent, P., Voit, D., Petkov, C. I. & Krumbholz, K. Parcellation of human and monkey core auditory cortex with fMRI pattern classification and objective detection of tonotopic gradient reversals. *Cereb. Cortex* 25, 3278–3289 (2015).
57. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* 536, 171–178 (2016).
58. Morosan, P. *et al.* Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. *Neuroimage* 13, 684–701 (2001).
59. Dick, F. *et al.* In vivo functional and myeloarchitectonic mapping of human primary auditory areas. *J. Neurosci.* 32, 16095–16105 (2012).
60. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532, 453–458 (2016).
61. Santoro, R. *et al.* Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proc. Natl. Acad. Sci.* 114, 4799–4804 (2017).
62. Carrasco, A. & Lomber, S. G. Evidence for Hierarchical Processing in Cat Auditory Cortex: Nonreciprocal Influence of Primary Auditory Cortex on the Posterior Auditory Field. *J. Neurosci.* 29, 14323–14333 (2009).
63. Rauschecker, J. P. & Scott, S. K. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat. Neurosci.* 12, 718–724 (2009).
64. Okada, K. *et al.* Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. *Cereb. Cortex* 20, 2486–2495 (2010).
65. Kaas, J. H. & Hackett, T. A. Subdivisions of auditory cortex and processing streams in primates. *Proc. Natl. Acad. Sci.* 97, 11793–11799 (2000).
66. Rauschecker, J. P., Tian, B., Hauser, M. & others. Processing of complex sounds in the macaque nonprimary auditory cortex. *Sci.-N. Y. THEN Wash.* 111–111 (1995).
67. Recanzone, G. H. & Cohen, Y. E. Serial and parallel processing in the primate auditory cortex revisited. *Behav. Brain Res.* 206, 1–7 (2010).
68. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402 (2007).
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.* 19, 498–502 (2009).
70. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* 521, 436 (2015).
71. Mlynski, W. & McDermott, J. H. Learning midlevel auditory codes from natural sound statistics. *Neural Comput.* 30, 631–669 (2018).

72. Lewicki, M. S. Efficient coding of natural sounds. *Nat. Neurosci.* 5, 356 (2002).
73. Yamins, D. L. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci.* 111, 8619–8624 (2014).
74. Eickenberg, M., Gramfort, A., Varoquaux, G. & Thirion, B. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage* 152, 184–194 (2017).
75. Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* 6, 27755 (2016).
76. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* 10, e1003915 (2014).
77. Cadena, S. A. *et al.* Deep convolutional models improve predictions of macaque V1 responses to natural images. *bioRxiv* 201764 (2017).
78. Güçlü, U. & van Gerven, M. A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35, 10005–10014 (2015).
79. Sahani, M. & Linden, J. F. How linear are auditory cortical responses? in *Advances in neural information processing systems* 125–132 (2003).
80. Atencio, C. A., Sharpee, T. O. & Schreiner, C. E. Cooperative nonlinearities in auditory cortical neurons. *Neuron* 58, 956–966 (2008).
81. Sadagopan, S. & Wang, X. Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex. *J. Neurosci.* 29, 11192–11202 (2009).
82. Portilla, J. & Simoncelli, E. P. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* 40, 49–70 (2000).
83. Balas, B., Nakano, L. & Rosenholtz, R. A summary-statistic representation in peripheral vision explains visual crowding. *J. Vis.* 9, 13–13 (2009).
84. Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P. & Movshon, J. A. A functional and perceptual signature of the second visual area in primates. *Nat. Neurosci.* 16, 974–981 (2013).
85. McDermott, J. H., Oxenham, A. J. & Simoncelli, E. P. Sound texture synthesis via filter statistics. in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on* 297–300 (IEEE, 2009).
86. David, S. V. & Shamma, S. A. Integration over multiple timescales in primary auditory cortex. *J. Neurosci.* 33, 19154–19166 (2013).
87. Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* 5, 143–156 (2001).
88. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* 48, 63 (2009).
89. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage* 9, 179–194 (1999).
90. Hall, D. A. *et al.* Sparse temporal sampling in auditory fMRI. *Hum. Brain Mapp.* 7, 213–223 (1999).
91. Norman-Haignere, S., Kanwisher, N. & McDermott, J. H. Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *J. Neurosci.* 33, 19451–19469 (2013).
92. Norman-Haignere, S. V. *et al.* Pitch-Responsive Cortical Regions in Congenital Amusia. *J Neurosci* (2016).
93. Kay, K., Rokem, A., Winawer, J., Dougherty, R. & Wandell, B. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. *Brain Imaging Methods* 7, 247 (2013).
94. Woolrich, M. W., Ripley, B. D., Brady, M. & Smith, S. M. Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage* 14, 1370–1386 (2001).
95. Schoppe, O., Harper, N. S., Willmore, B. D. B., King, A. J. & Schnupp, J. W. H. Measuring the Performance of Neural Models. *Front. Comput. Neurosci.* 10, 10 (2016).
96. Hyvarinen, A. Fast and robust fixed-point algorithms for independent component analysis. *Neural Netw. IEEE Trans. On* 10, 626–634 (1999).

97. Pasley, B. N. *et al.* Reconstructing speech from human auditory cortex. *PLoS Biol.* 10, e1001251 (2012).
98. Norman-Haignere, S. & McDermott, J. H. Distortion products in auditory fMRI research: measurements and solutions. *NeuroImage* 129, 401–413 (2016).
99. Saxe, R., Brett, M. & Kanwisher, N. Divide and conquer: a defense of functional localizers. *Neuroimage* 30, 1088–1096 (2006).
100. de Cheveigné, A. *et al.* Decoding the auditory brain with canonical component analysis. *NeuroImage* 172, 206–216 (2018).
101. Chen, P.-H. C. *et al.* A reduced-dimension fMRI shared response model. in *Advances in Neural Information Processing Systems* 460–468 (2015).

A) Example model responses to natural and model-matched sounds



B) Schematic of histogram-matching algorithm

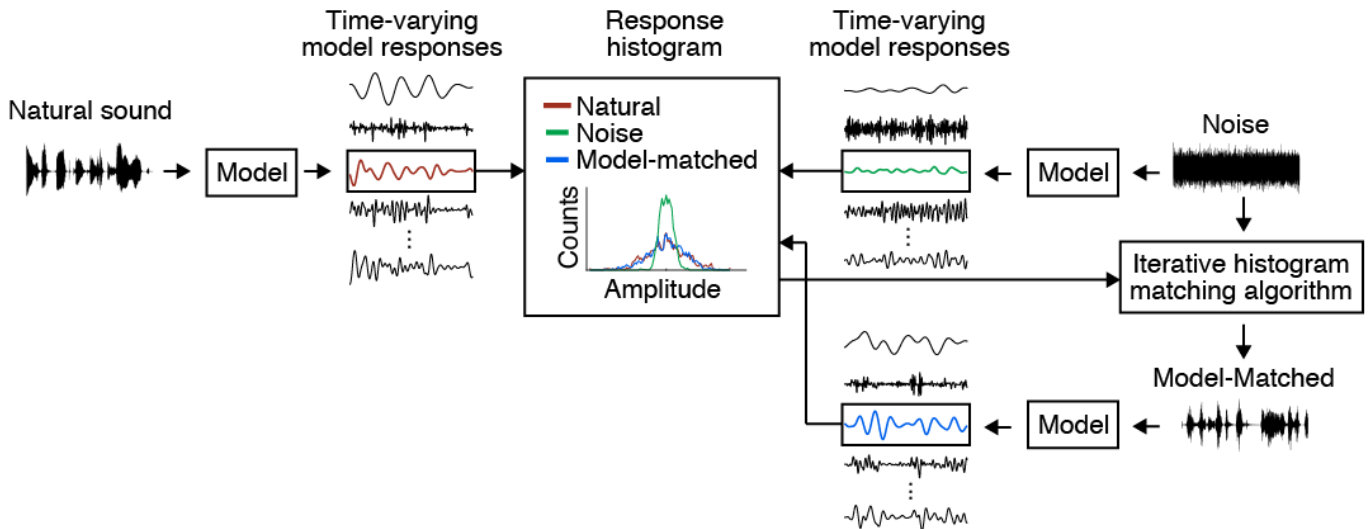


Figure S1. Schematic of model-matching approach. (A) All of the models considered are defined by the response timecourse of a set of model features, each computed by filtering a cochleagram representation of sound (illustrated in Figure 1A). Our model-matching algorithm collapses these timecourses across time to form a histogram, and then generates a sound with the same histograms as a natural sound. Here we plot example timecourses and histograms for three example natural sounds (left panel) and corresponding model-matched sounds (right panel). Different natural sounds

produce distinct response timecourses and histograms. Corresponding natural and model-matched sounds produce similar response histograms, but distinct response timecourses. (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 algorithm⁴⁸.

Schematic of block design used to present sounds in the MRI scanner

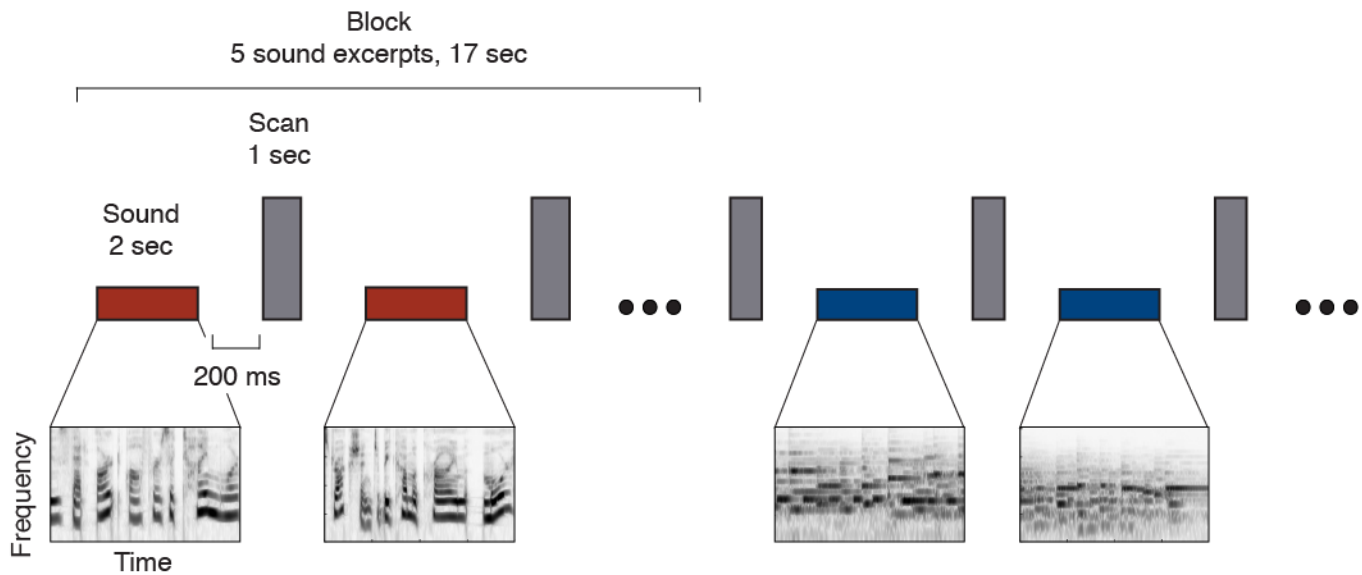
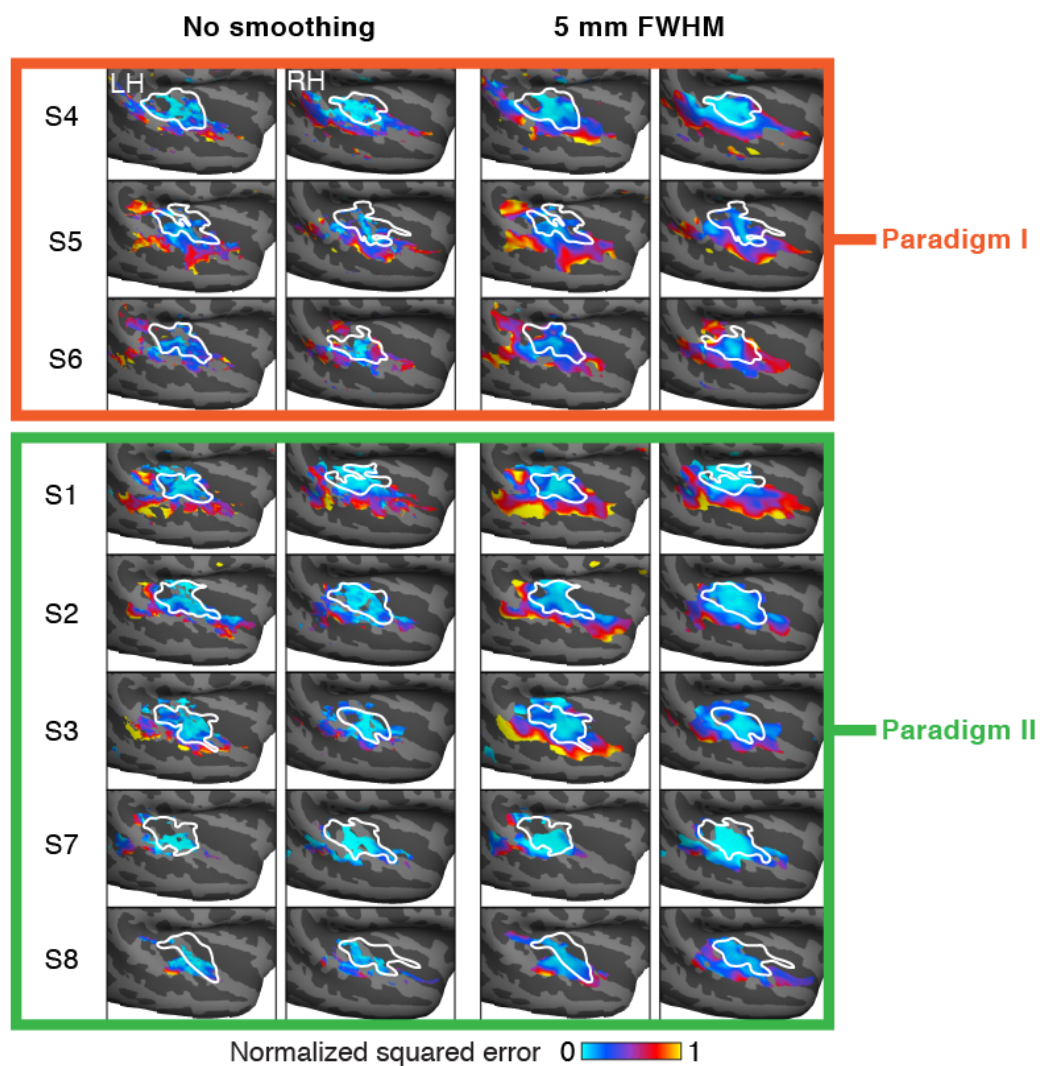


Figure S2. Schematic of the sparse scanning paradigm used to present stimuli in the experiment. Each 10-second stimulus was subdivided into five 2-second segments. These five segments were presented in a random order with a 1-second scan acquisition interspersed between each presentation (1.05 seconds for Paradigm II). A short 200-ms buffer was present between stimuli and scan acquisitions. The total duration of each “block” of 5 sounds was 17 seconds (17.25 seconds for Paradigm II). The response to a stimulus was computed as the average response of the 2nd through 5th scan acquisition after block onset (the first acquisition was discarded to account for the hemodynamic lag).

A Dissimilarity maps for different paradigms and smoothing kernels



B Dissimilarity vs. distance to PAC

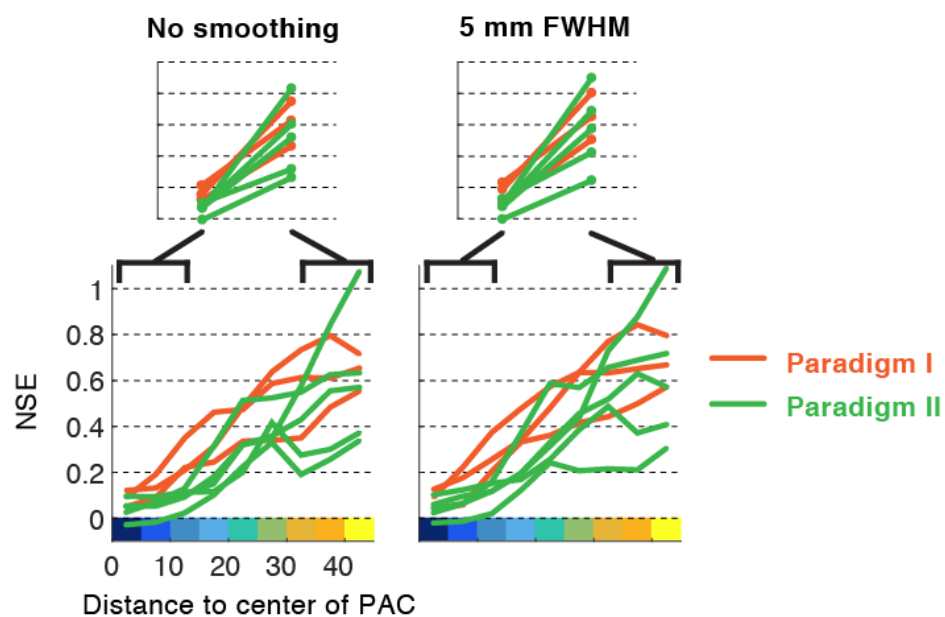


Figure S3. Results broken down by paradigm and the presence/absence of smoothing. In Paradigm I, only the natural sounds were repeated. In Paradigm II, both natural and model-matched sounds were repeated. A smaller voxel size was employed in Paradigm II (2 mm isotropic instead of 2.1 x 2.1 x 4 mm for Paradigm I). (A) Natural vs. 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 (orange for Paradigm I and green for Paradigm II). NSE values are averaged across the left and right hemisphere since we observed similar trends in both hemispheres.

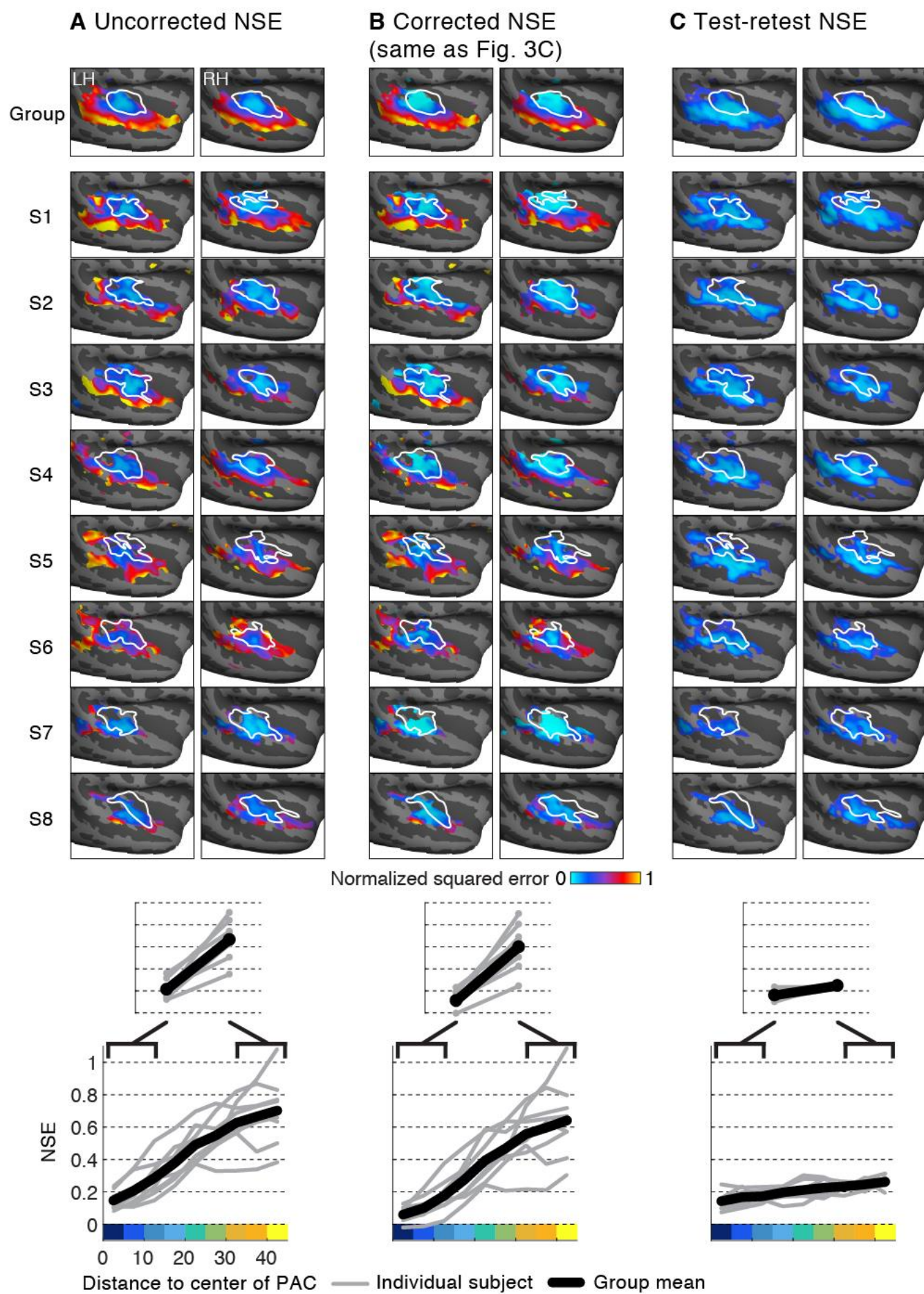


Figure S4. Noise correction and voxel reliability. (A) The uncorrected NSE between responses to natural and model-matched sounds. (B) Corrected NSE maps (same as Figures 3C) replicated here for ease of comparison with the uncorrected maps. (C) Test-retest reliability of voxel responses measured with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

Dissimilarity vs. distance to
anatomically-defined PAC
(medial HG / TE1.1)

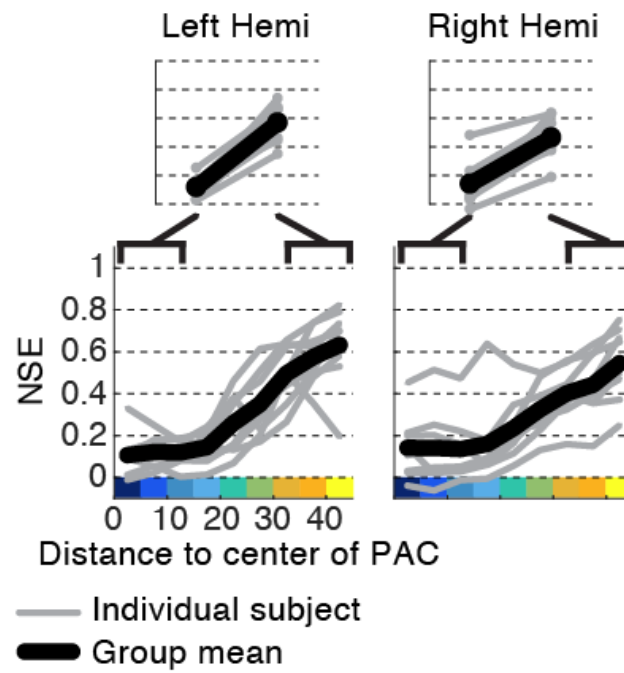


Figure S5. Dissimilarity vs. distance from PAC using an anatomical rather than tonotopic definition of PAC. Voxels are binned based on their distance to center of anatomical region TE1.1⁵⁸, which is located in posteromedial HG. Format is the same as Figure 3D.

Dissimilarity maps leaving out intelligible speech stimuli

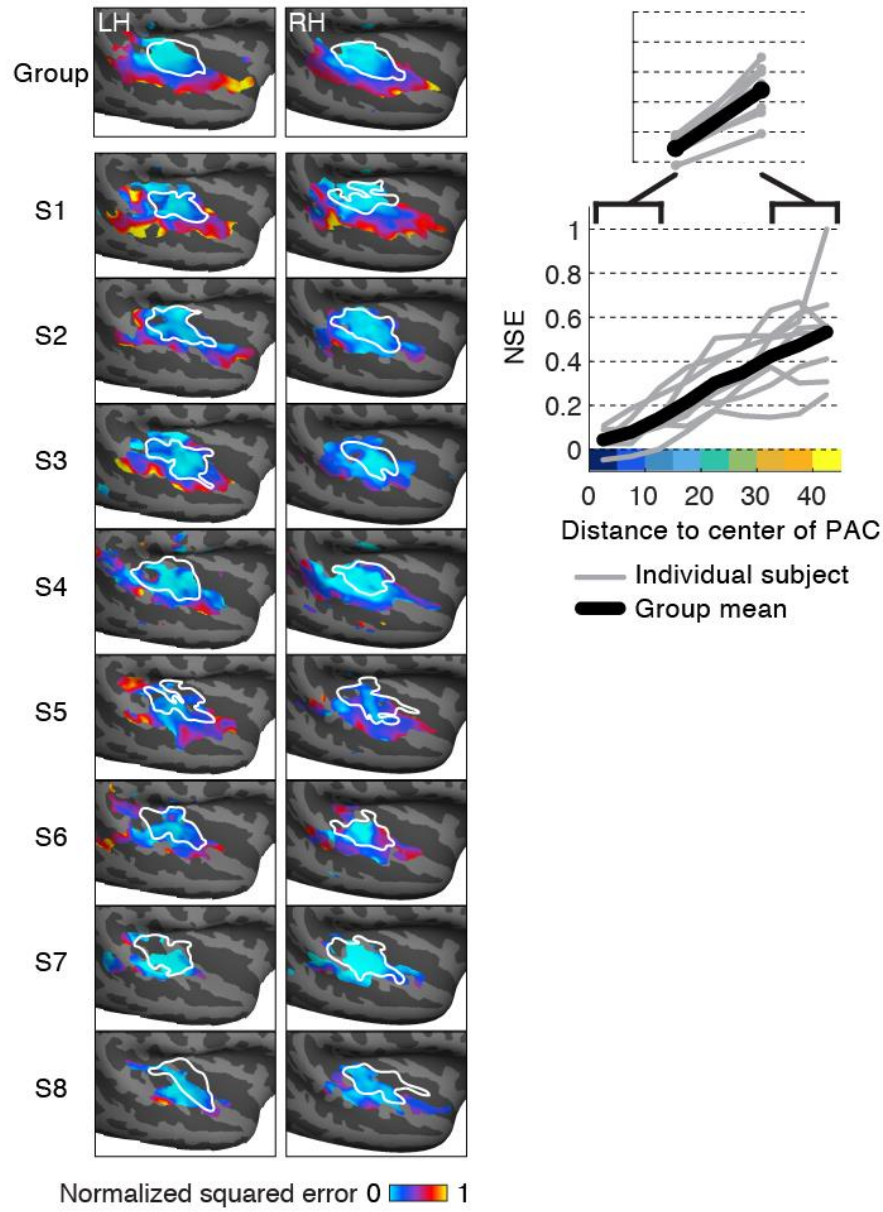


Figure S6. Dissimilarity maps and annular analyses omitting intelligible speech stimuli. Maps plot the normalized squared error (NSE) between voxel responses to natural and model-matched sounds omitting English speech and music with English vocals (all subjects were native English speakers). Format is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

Individual subject dissimilarity maps for subsets of model features

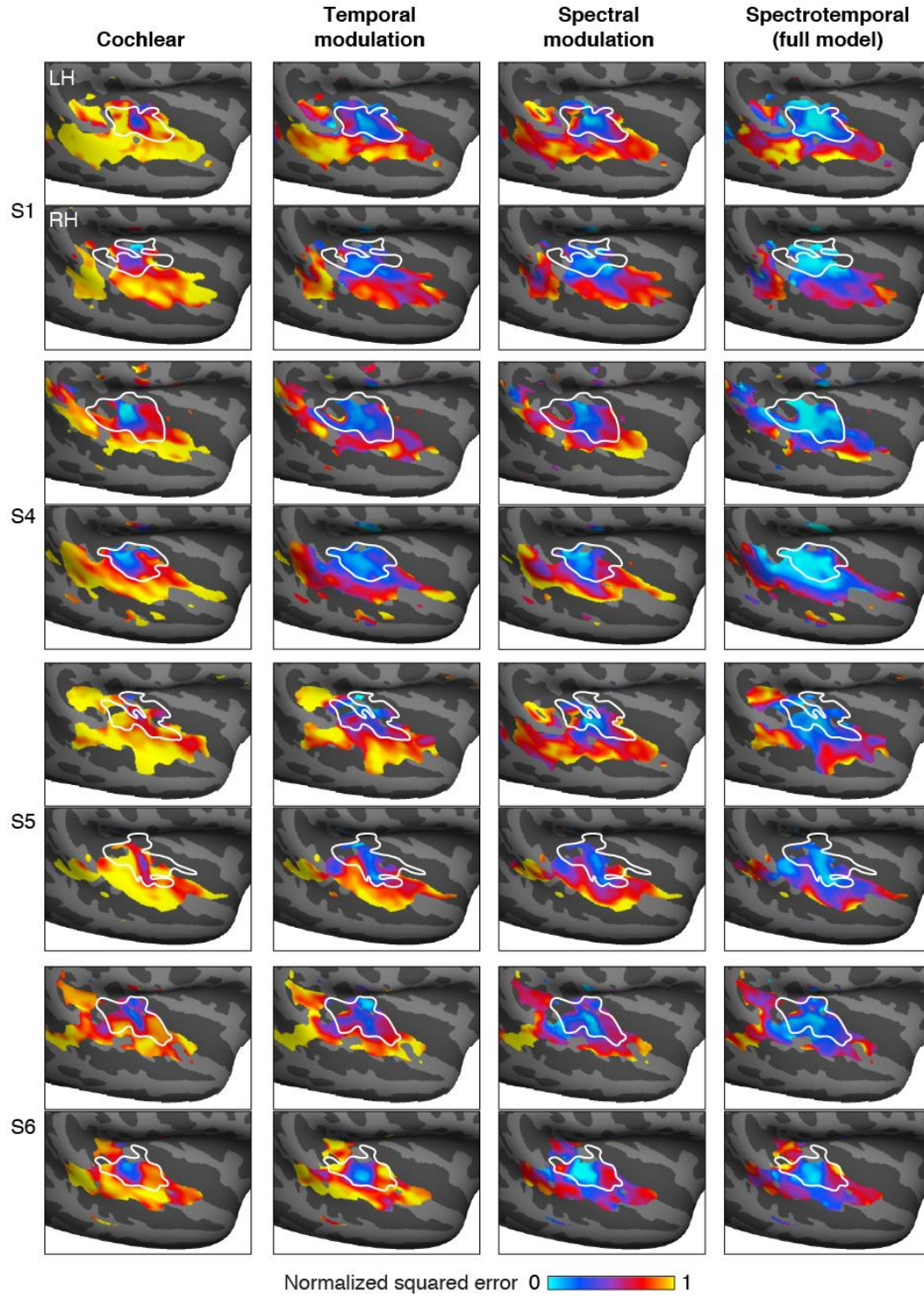


Figure S7. Individual subject maps of dissimilarity between responses to natural and model-matched sounds (NSE) for subsets of model features. Format the same as Figure 4B. Only subjects scanned in Paradigm I are shown, since Paradigm II did not include model-matched sounds constrained by subsets of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

Individual-subject prediction errors
(full spectrotemporal model)

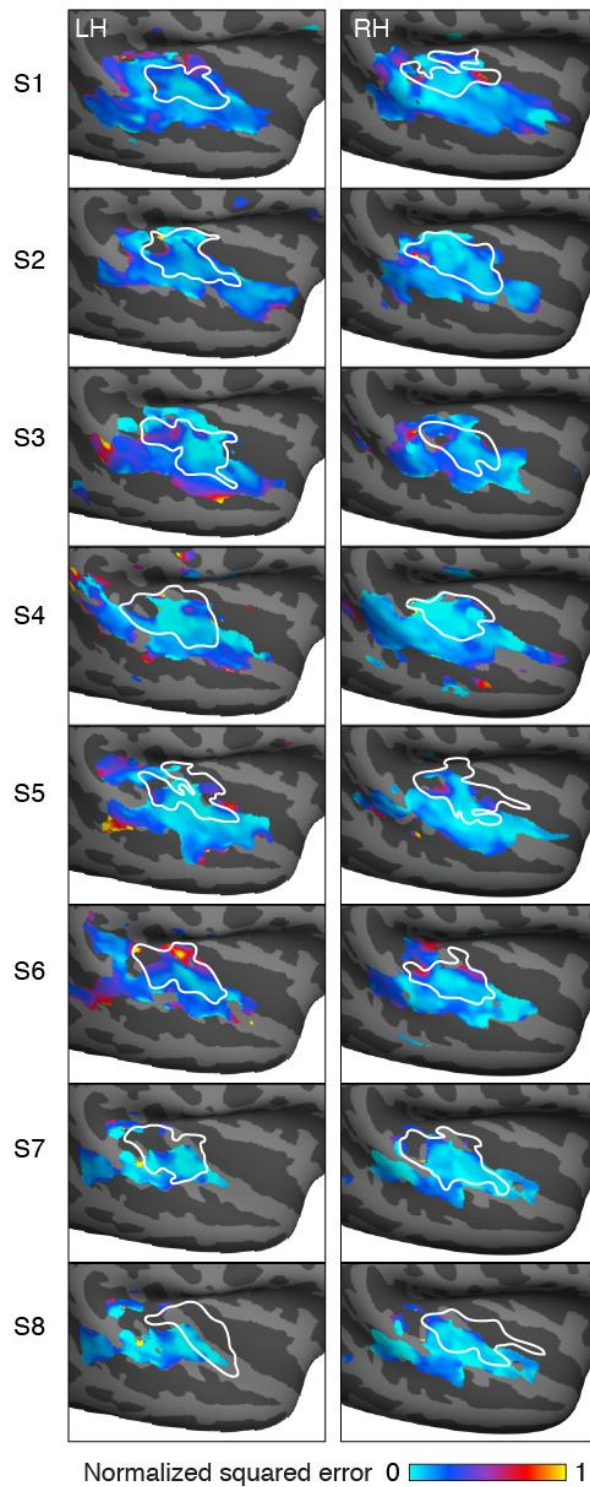
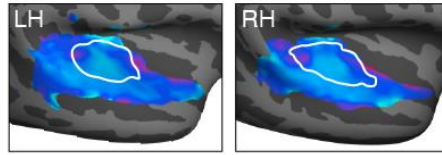


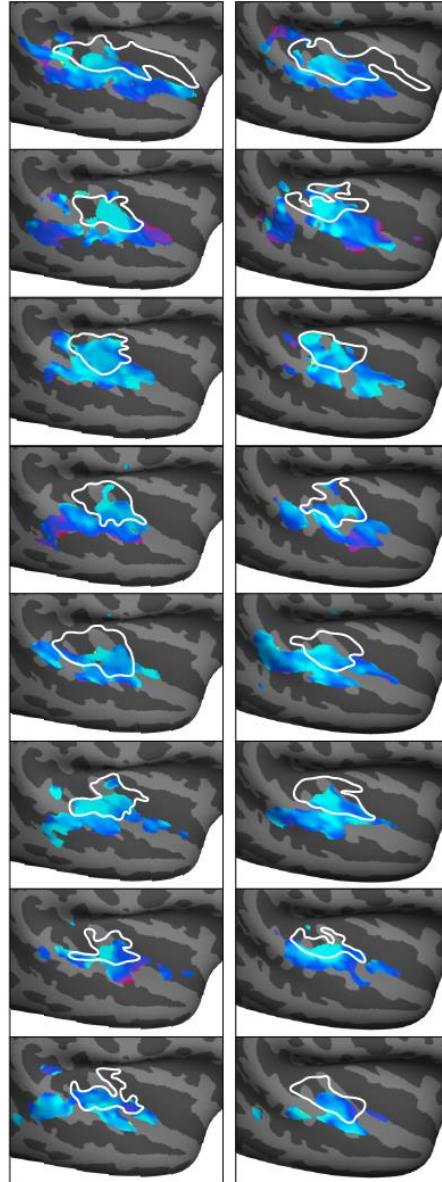
Figure S8. Individual subject prediction error maps based on the full spectrotemporal modulation model (maps for subsets of model features are omitted due to space constraints, but, like the maps for the full model, they resembled those of the group). Format the same as Figure 3C.

Prediction errors for larger 165 sound set (from full model)

Group



Individual subjects



Normalized squared error 0 1

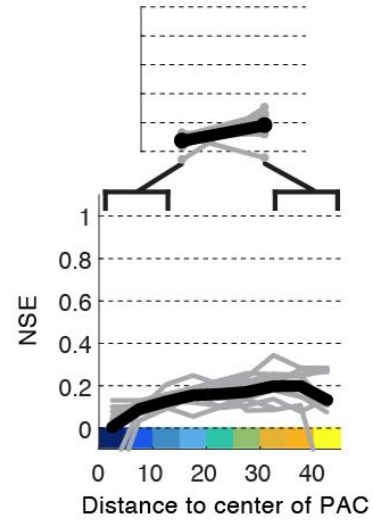


Figure S9. Prediction accuracy of the full spectrotemporal model using a larger set of 165 natural sounds tested in a prior study. Format is the same as Figure 4B&C.

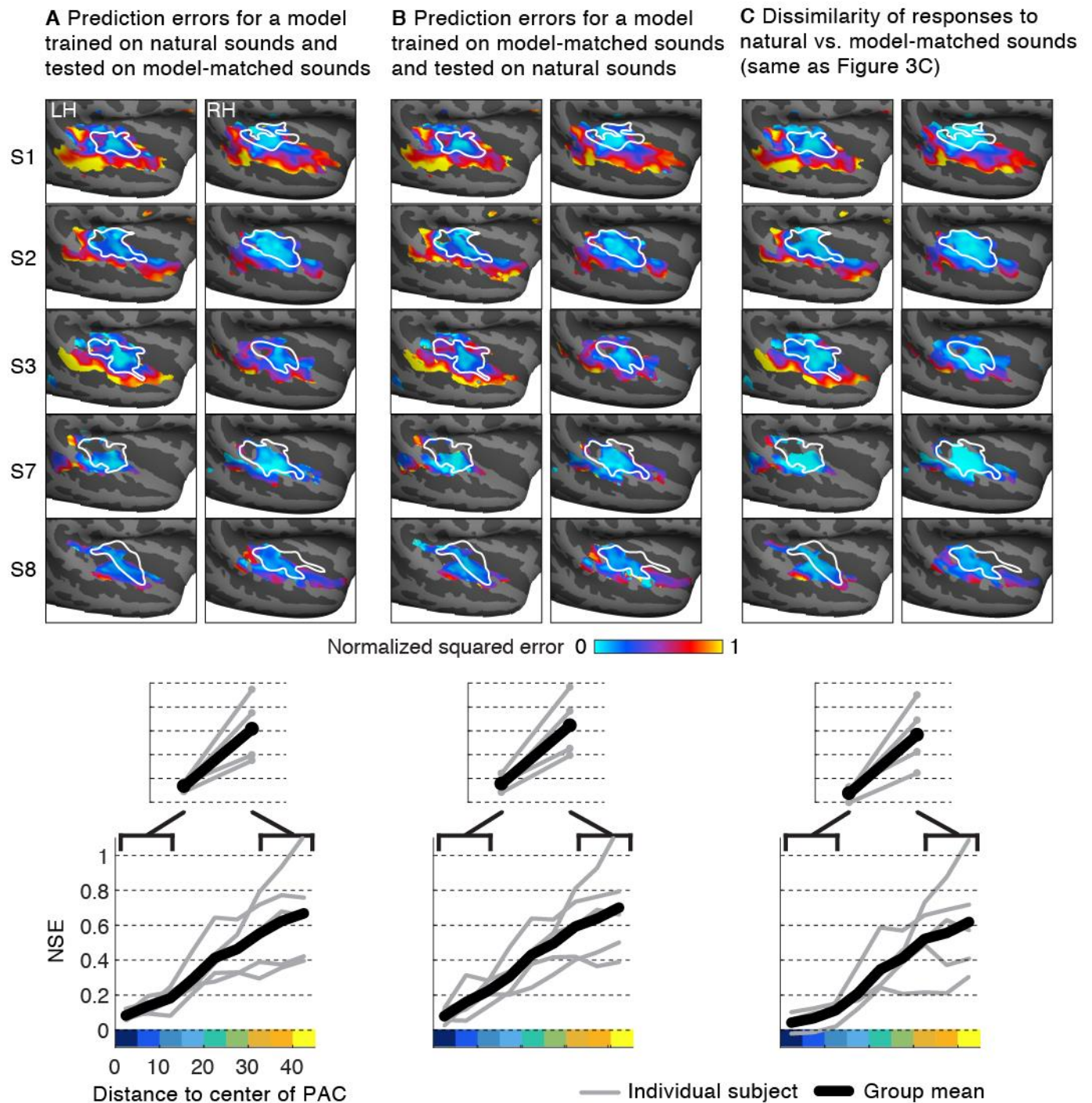
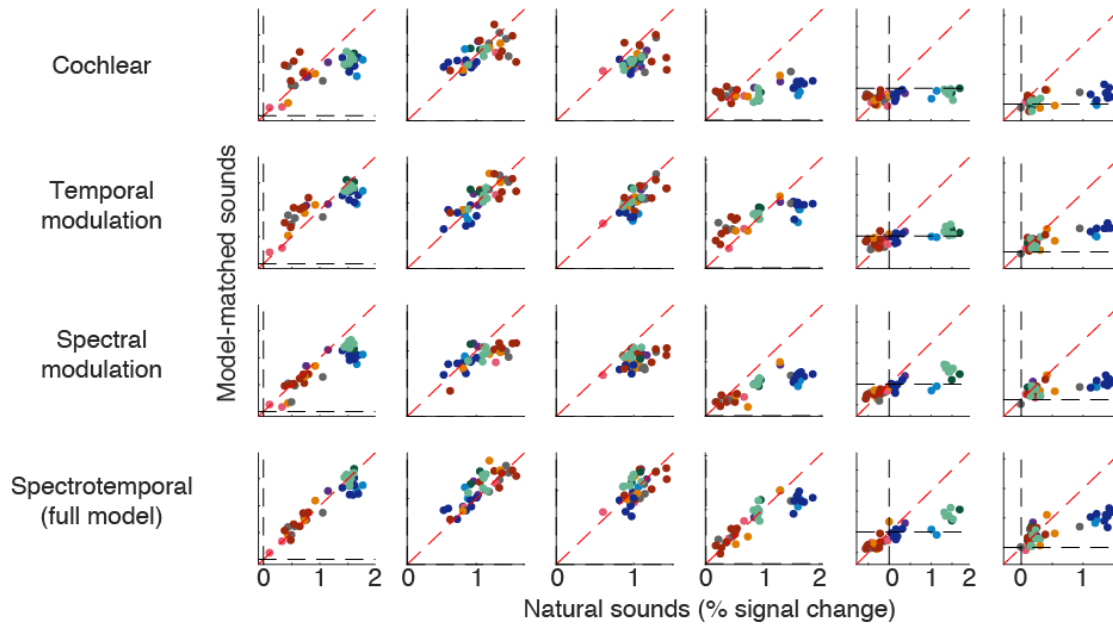
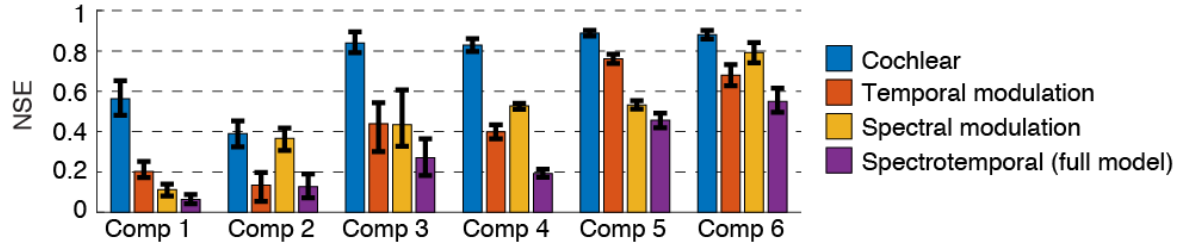


Figure S10. (A) Prediction error maps of a model trained on natural sounds and tested on model-matched sounds. (B) Prediction error maps of a model trained on model-matched sounds and tested on natural sounds. (C) For comparison, the error of the measured voxel response to natural and model-matched sounds is reproduced here (same as Figure 3C&D). 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 which both natural and model-matched sounds were repeated, which made it possible to noise-correct the predictions.

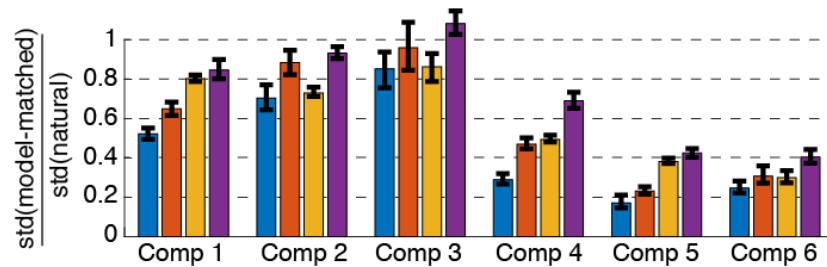
A Component responses to model-matched sounds constrained by subsets of model features



B Overall dissimilarity



C Ratio of response variation



D Correlation

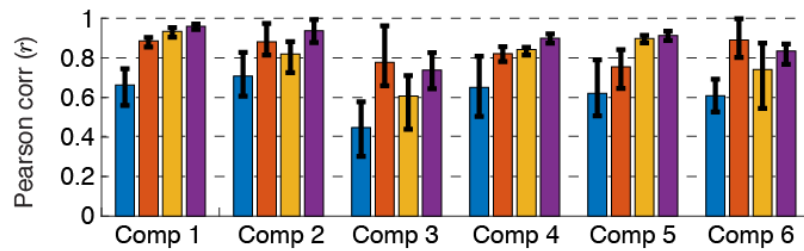
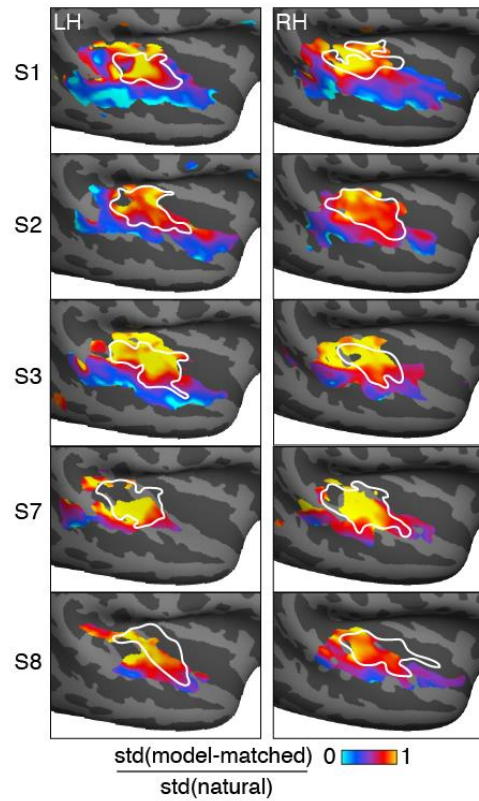


Figure S11. Plots showing component responses to sounds constrained by subsets of model features, as well as the full model (Figure 6 only shows results from the full model). (B) Normalized squared error between natural and model-matched sounds. (C) Ratio of the standard deviation of responses to model-matched and natural sounds. (D) Correlation of responses to natural and model-matched sounds. Format is the same as Figures 6C-6F.

A Ratio of response variation



B Correlation

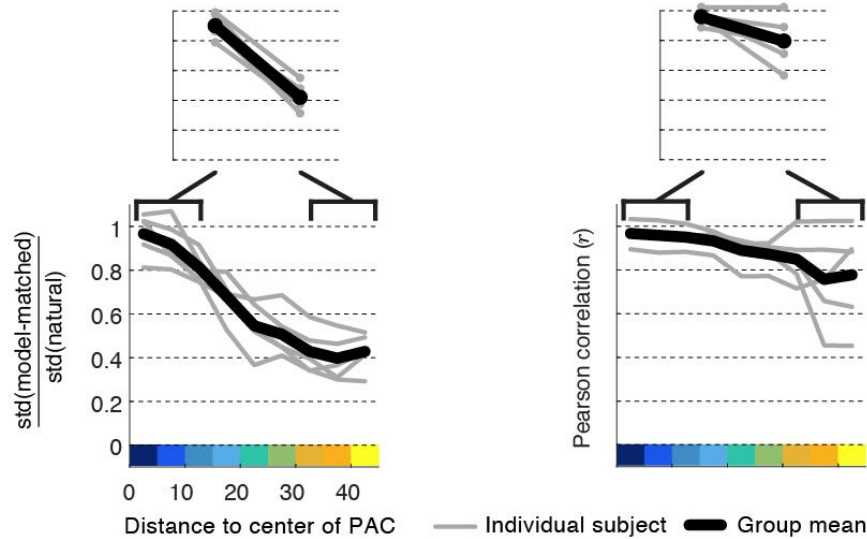
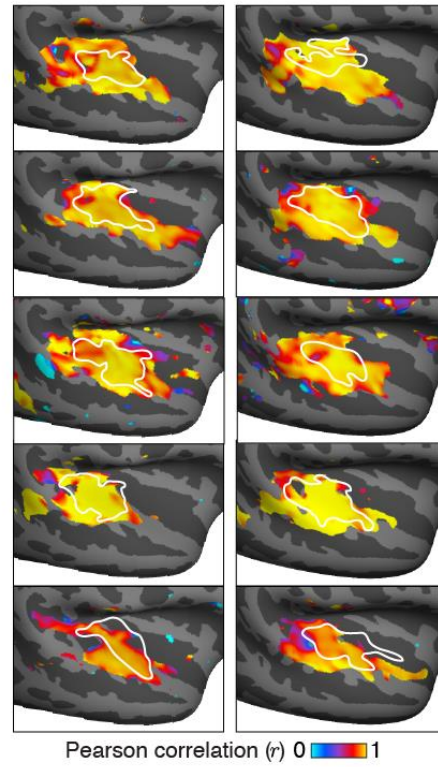


Figure S12. Examination of the nature of the divergent responses to natural and model-matched sounds. (A) Whole-brain maps plotting the variation in responses to natural vs. model-matched sounds, measured as the ratio of the standard deviation of responses to the two sound sets. Cool colors indicated less response variation for model-matched sounds. Distance-to-PAC summary analysis is plotted below. (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 ($NSE < 0.4$). For the correlation maps (panel B), we excluded voxels that did not have a reliable correlation to model-matched sounds ($r < 0.4$), as was the case in many non-primary voxels due to weak responses. For such voxels, it is difficult to estimate a reliable correlation, since the correlation is undefined as the variance of one variable goes to zero.

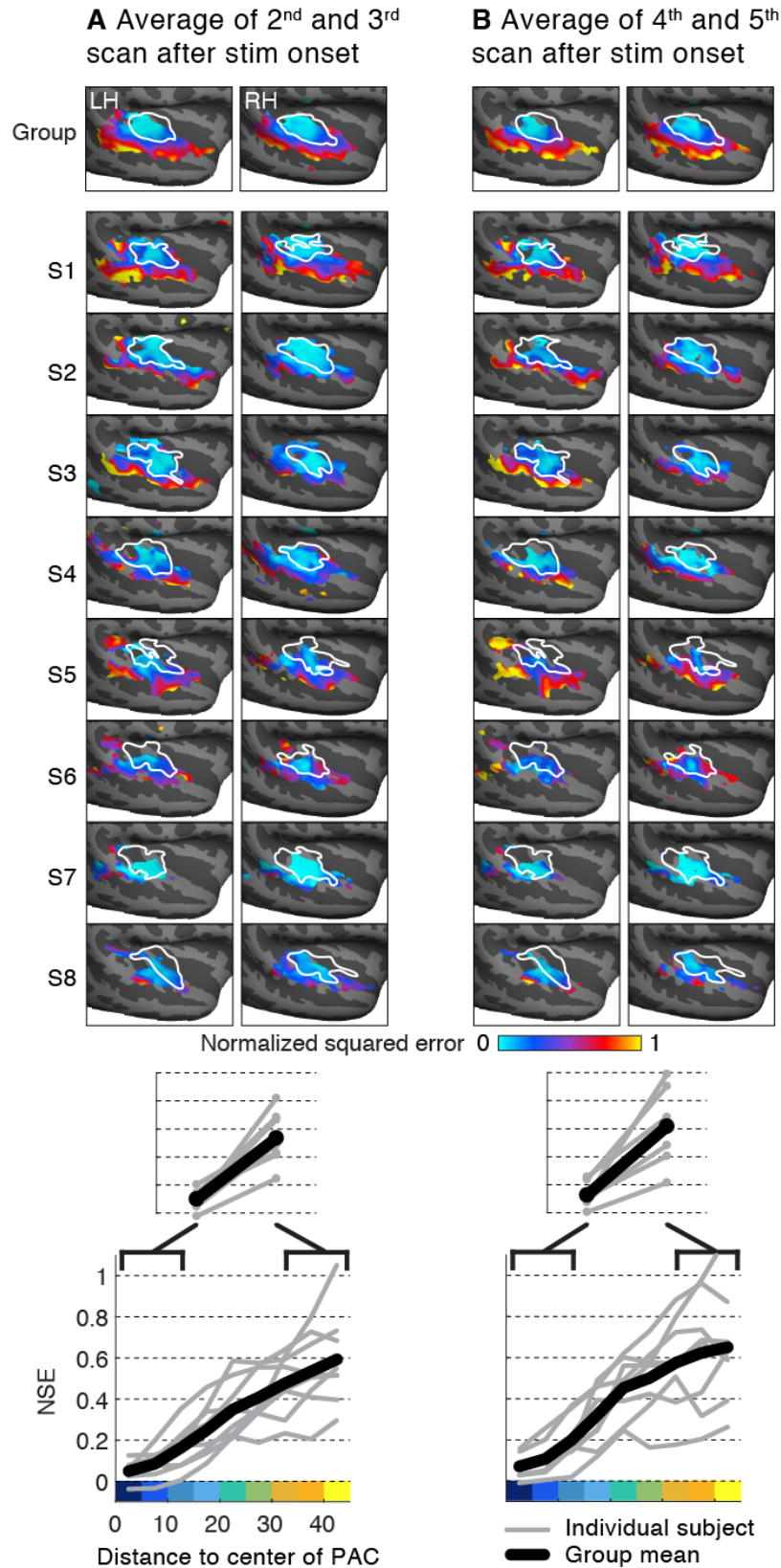
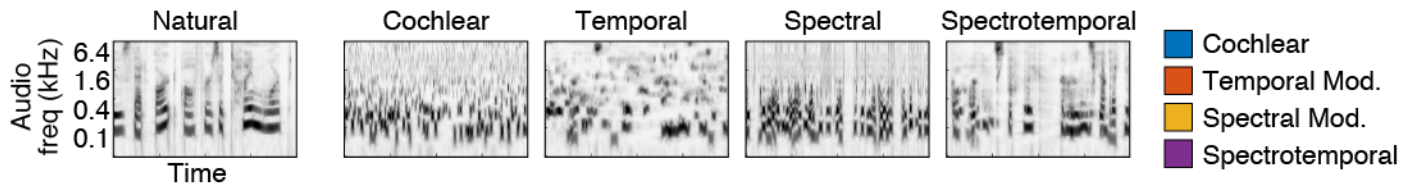
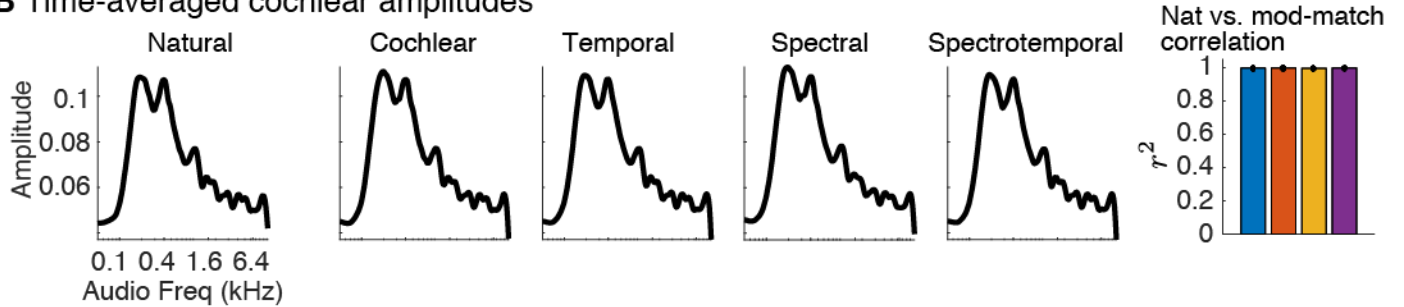


Figure S13. Effect of the fMRI averaging window on the dissimilarity of responses to natural and model matched sounds. Each stimulus was 10-seconds in duration, but was split-up into five 2-second segments (see Figure S2). After each segment a single scan acquisition was collected. Analyses in the main text were based on the average response of the 2nd through 5th acquisition 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 Figures 3C&D.

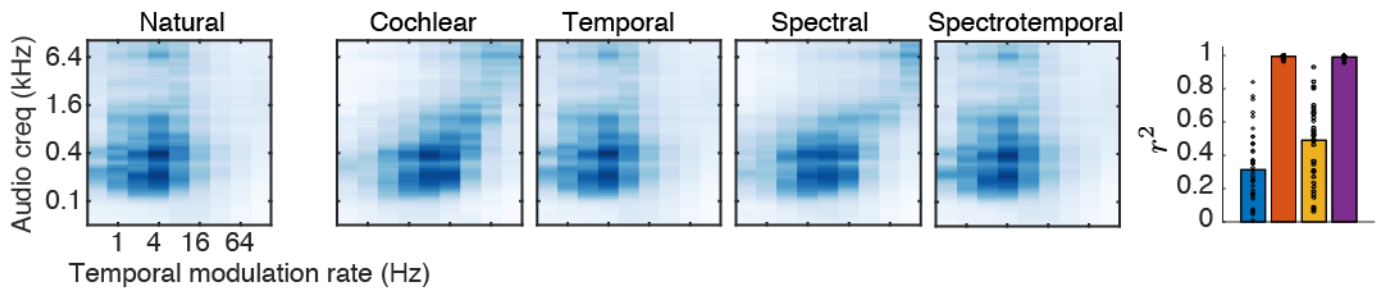
A Cochleograms for example model-matched stimuli



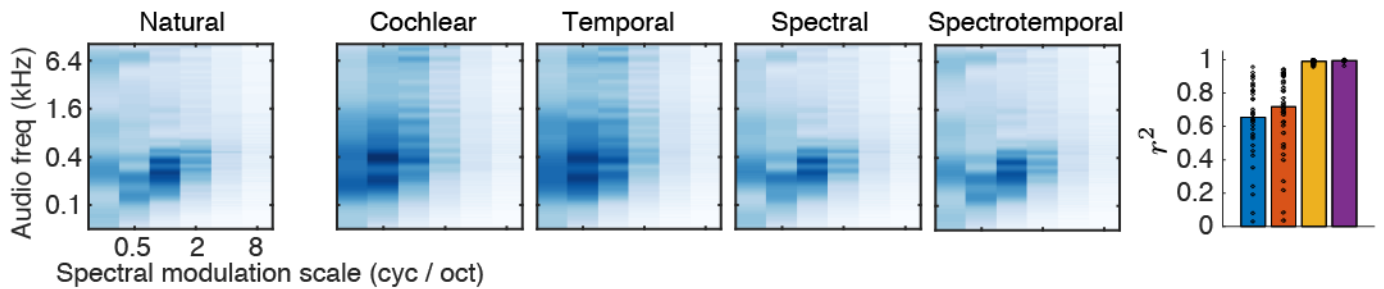
B Time-averaged cochlear amplitudes



C Time-averaged temporal modulation amplitude (standard deviation)



D Time-averaged spectral modulation amplitude (standard deviation)



E Time-averaged spectrotemporal modulation amplitude for example audio frequency (200 Hz)

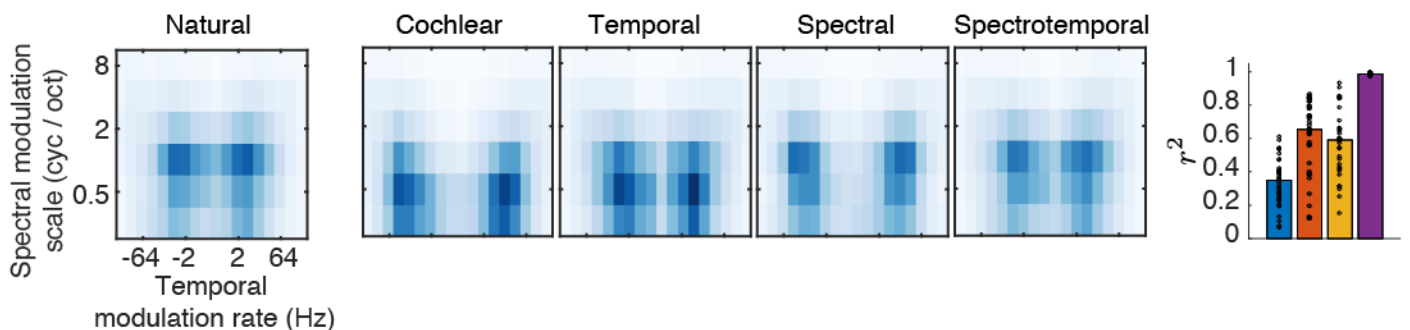
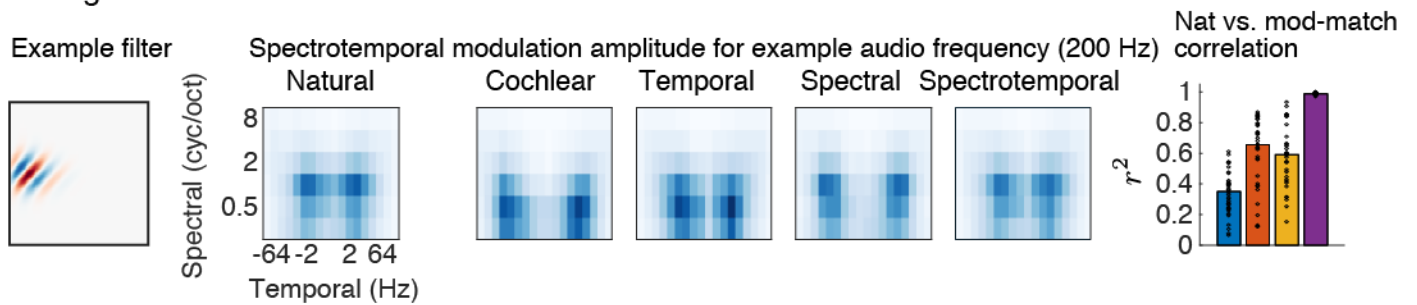


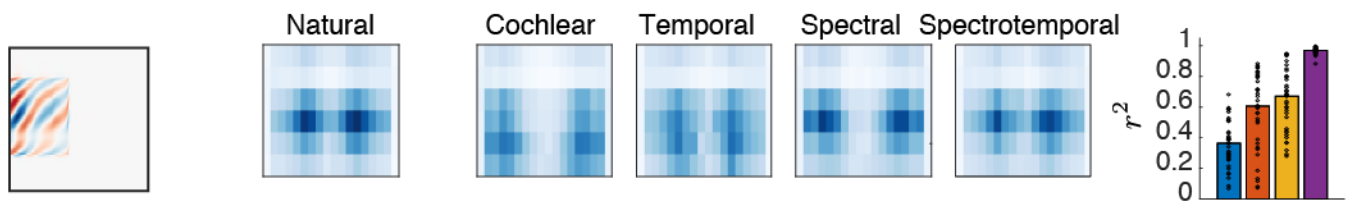
Figure S14. Validation of the model-matching synthesis procedure via comparison of time-averaged statistics for natural and model-matched sounds. (A) Cochleograms for a natural sound (a speech excerpt) and four corresponding model-matched sounds. (B-E) Each model was defined by a set of feature responses. This figure plots 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 side of the panel plots the correlation of the filter amplitudes across all model filters for corresponding natural and model-matched sounds. Each point 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 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 (200 Hz).

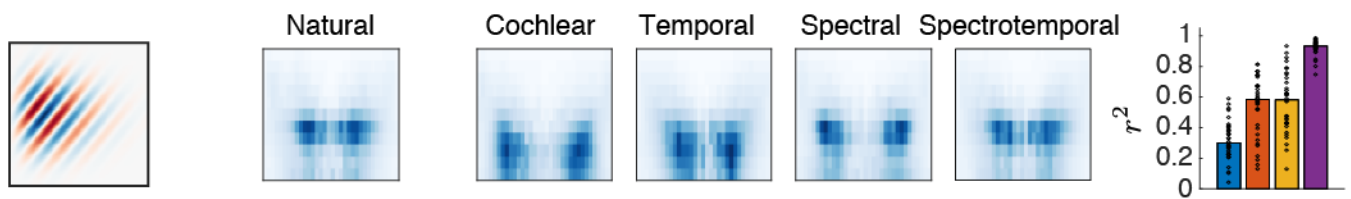
A Original model filters used to constrain the model-matched sounds



B Phase-randomized filters



C Half-bandwidth filters



D Random filters

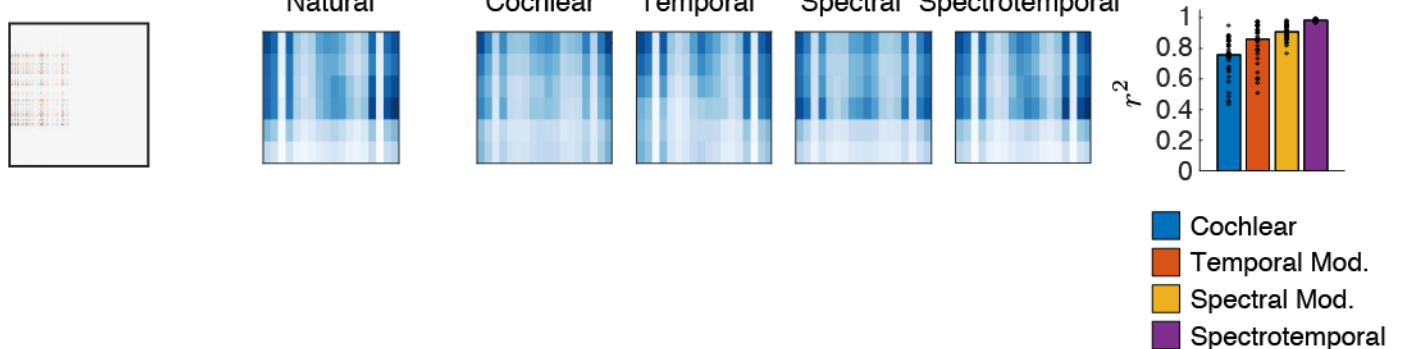


Figure S15. Comparison of how well the natural and synthesized (model-matched) sounds are matched when evaluated using spectrotemporal filters that differed from those used to generate the model-matched sounds. In each case, we plot an example filter from the model (left), the amplitude (standard deviation) of the filter responses as a function of the temporal rate and spectral scale for an example audio frequency (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 Figure S14E). (A) The original spectrotemporal filters from Chi et al. (2005) that were used to constrain the model-matched sounds (same as Figure S14E). (B) Spectrotemporal filters with randomized temporal and spectral phase. (C) A model with narrower bandwidths and more filters to compensate. (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, like those we observed in non-primary regions.

A Assuming the noise power of natural and model-matched sounds is equal

B Separately estimating the noise power of natural and model-matched sounds

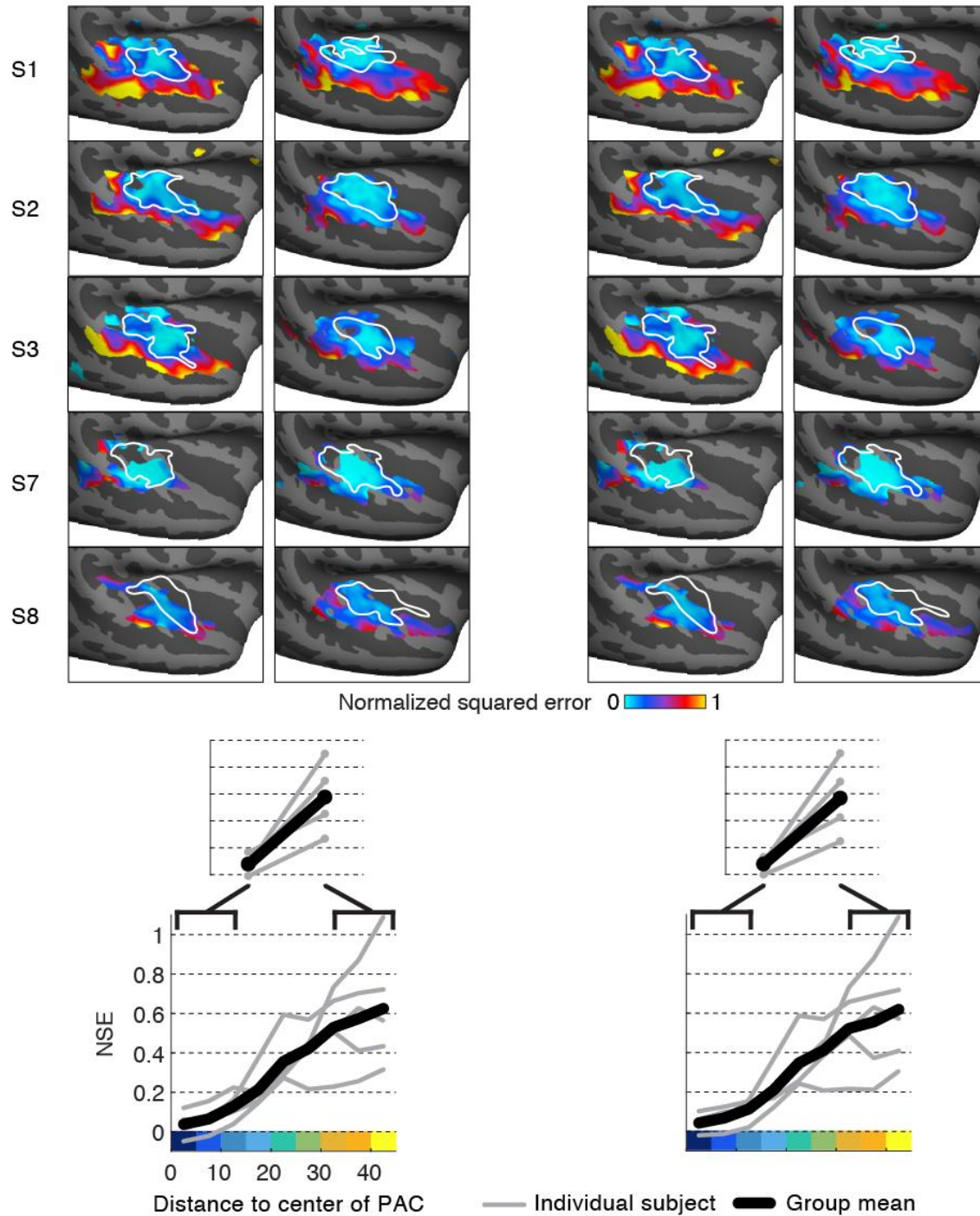


Figure S16. Comparison of noise correction using noise estimates exclusively from responses to natural sounds or from both natural and model-matched sounds. Noise-correction required estimating the power of the noise for natural and model-matched sounds. For Paradigm I, only responses to natural sounds 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 values computed by assuming the noise power for natural and model-matched sounds is equal, and 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 Figure 3C). Results are similar in both cases.

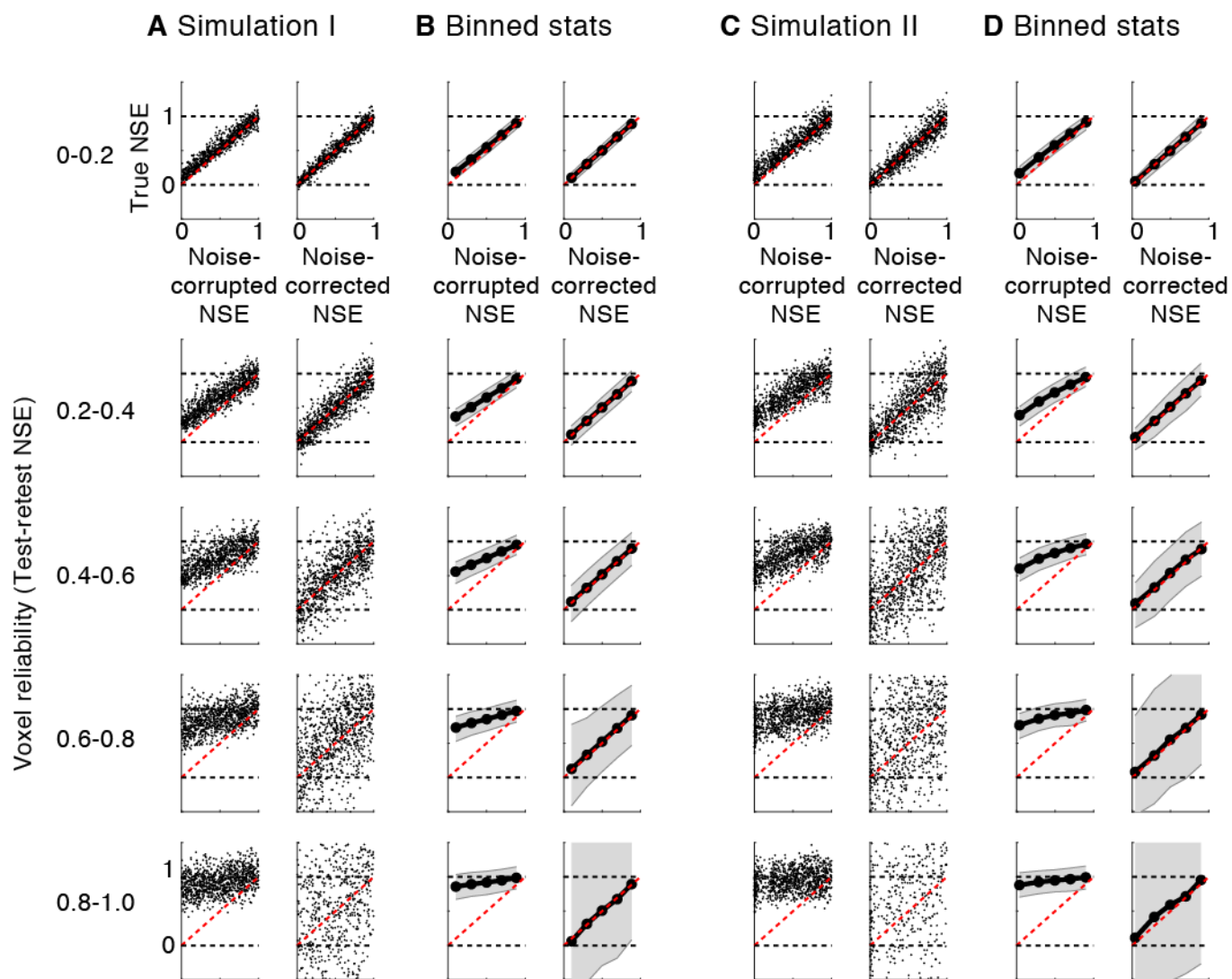


Figure S17. Results of noise-correction simulations. (A) Each dot corresponds to a single simulated voxel. The true NSE value is plotted against the noise-corrupted and noise-corrected NSE values. Results have been grouped by the reliability of the simulated voxel responses as measured by the test-retest NSE of the voxel responses (from high to low reliability going from top to bottom). (B) The median and standard deviation of the samples. (C&D) Same as for panels A&B, but for Simulation II (see *Evaluating the noise-corrected NSE with simulated data* in the Methods for details of the two simulations).

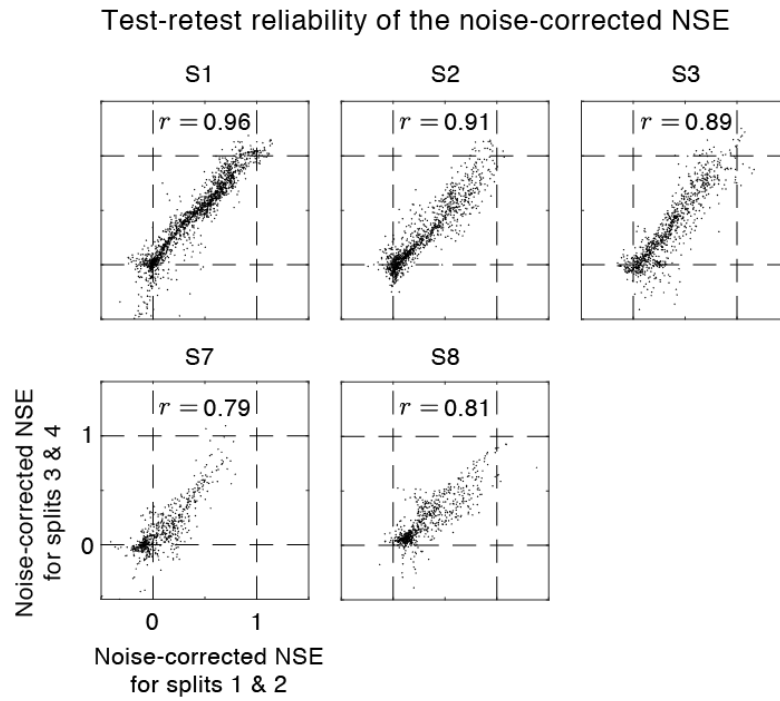


Figure S18. Test-retest reliability of the noise-corrected NSE (measured across splits of fMRI data). Each dot plots the noise-corrected NSE for a single voxel. The Spearman rank correlation is shown at the top of each plot for each subject. Results are shown for subjects scanned in Paradigm II, for which there was sufficient data to compute two separate estimates of the noise-corrected NSE.

Subject	Hemisphere	Duplications along HG
S1	Left	Partial Duplication
S1	Right	Single Gyrus
S2	Left	Single Gyrus
S2	Right	Single Gyrus
S3	Left	Partial Duplication
S3	Right	Single Gyrus
S4	Left	Partial Duplication
S4	Right	Single Gyrus
S5	Left	Partial Duplication
S5	Right	Partial Duplication
S6	Left	Partial Duplication
S6	Right	Complete Duplication
S7	Left	Single Gyrus
S7	Right	Complete Duplication
S8	Left	Partial Gyrus
S8	Right	Complete Duplication

Table S1. The presence/absence of HG duplications 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 et al. (2011)⁵²