Neuron

Article

Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition

Highlights

- Voxel decomposition infers canonical components of responses to natural sounds
- Decomposition reveals speech and music selectivity in distinct non-primary regions
- Music selectivity is diluted in raw voxel responses due to component overlap
- Organization of primary regions reflects tuning for frequency, modulation, and pitch

Authors

Sam Norman-Haignere,
Nancy G. Kanwisher,
Josh H. McDermott

Correspondence

snormanhaignere@gmail.com

In Brief

Norman-Haignere et al. introduce voxel decomposition: a method that infers putative neural populations ("components") from fMRI responses to natural stimuli. This method reveals distinct cortical pathways selective for music and speech, despite being unconstrained by prior functional hypotheses.

Norman-Haignere et al., 2015, Neuron *88*, 1281–1296
December 16, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.neuron.2015.11.035

Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition

Sam Norman-Haignere,[1,*] Nancy G. Kanwisher,[1,2,3] and Josh H. McDermott[1,3]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]McGovern Institute for Brain Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Co-senior author
*Correspondence: snormanhaignere@gmail.com

http://dx.doi.org/10.1016/j.neuron.2015.11.035

SUMMARY

The organization of human auditory cortex remains unresolved, due in part to the small stimulus sets common to fMRI studies and the overlap of neural populations within voxels. To address these challenges, we measured fMRI responses to 165 natural sounds and inferred canonical response profiles ("components") whose weighted combinations explained voxel responses throughout auditory cortex. This analysis revealed six components, each with interpretable response characteristics despite being unconstrained by prior functional hypotheses. Four components embodied selectivity for particular acoustic features (frequency, spectrotemporal modulation, pitch). Two others exhibited pronounced selectivity for music and speech, respectively, and were not explainable by standard acoustic features. Anatomically, music and speech selectivity concentrated in distinct regions of non-primary auditory cortex. However, music selectivity was weak in raw voxel responses, and its detection required a decomposition method. Voxel decomposition identifies primary dimensions of response variation across natural sounds, revealing distinct cortical pathways for music and speech.

INTRODUCTION

Just by listening, humans can discern a vast array of information about the objects and events in the environment around them. This ability to derive information from sound is instantiated in a cascade of neuronal processing stages extending from the cochlea to the auditory cortex. Although much is known about the transduction and subcortical processing of sound, cortical representations of sound are less well understood. Prior work has revealed tuning in and around primary auditory cortex for acoustic features such as frequency (Da Costa et al., 2011; Humphries et al., 2010), temporal and spectral modulations (Barton et al., 2012; Chi et al., 2005; Santoro et al., 2014; Schönwiesner and Zatorre, 2009), spatial cues (Rauschecker and Tian, 2000; Stecker et al., 2005), and pitch (Bendor and Wang, 2005; Norman-Haignere et al., 2013; Patterson et al., 2002). The tuning properties of non-primary regions are less clear. Although many studies have reported selectivity for vocal sounds (Belin et al., 2000; Petkov et al., 2008) and speech (Mesgarani et al., 2014; Overath et al., 2015; Scott et al., 2000), the cortical representation of environmental sounds (Engel et al., 2009; Giordano et al., 2013) and of music (Abrams et al., 2011; Angulo-Perkins et al., 2014; Fedorenko et al., 2012; Koelsch et al., 2005; Leaver and Rauschecker, 2010; Rogalsky et al., 2011; Tierney et al., 2013) is poorly understood. Moreover, debate continues about the extent to which the processing of music, speech, and other natural sounds relies on shared versus distinct neuronal mechanisms (Peretz et al., 2015; Zatorre et al., 2002) and the extent to which these mechanisms are organized hierarchically (Chevillet et al., 2011; Hickok and Poeppel, 2007; Staeren et al., 2009).

This paper was motivated by two limitations of many neuroimaging studies (including our own) that have plausibly hindered the understanding of human auditory cortical organization. First, responses are typically measured to only a small number of stimulus dimensions chosen to test particular hypotheses. Because there are many dimensions to which neurons could be tuned, it is difficult to test the specificity of tuning and to know whether the dimensions tested are those most important to the cortical response. Second, the spatial resolution of fMRI is coarse: each voxel represents the aggregate response of hundreds of thousands of neurons. If different neural populations spatially overlap, their response will be difficult to isolate using standard voxel-wise analyses.

To overcome these limitations, we developed an alternative method for inferring neuronal stimulus selectivity and its anatomical organization from fMRI data. Our approach tries to explain the response of each voxel to a large collection of natural sounds as the weighted sum of a small number of response profiles ("components"), each potentially reflecting the tuning properties of a different neuronal sub-population. This method infers response dimensions from structure in the data, rather than testing particular features hypothesized to drive neural responses. And unlike standard voxel-wise analyses, our method can isolate responses from overlapping neural populations, because multiple response profiles are used to model each voxel. When applied to auditory cortex, voxel decomposition identifies a small number of interpretable response dimensions and reveals their anatomical organization in the cortex.

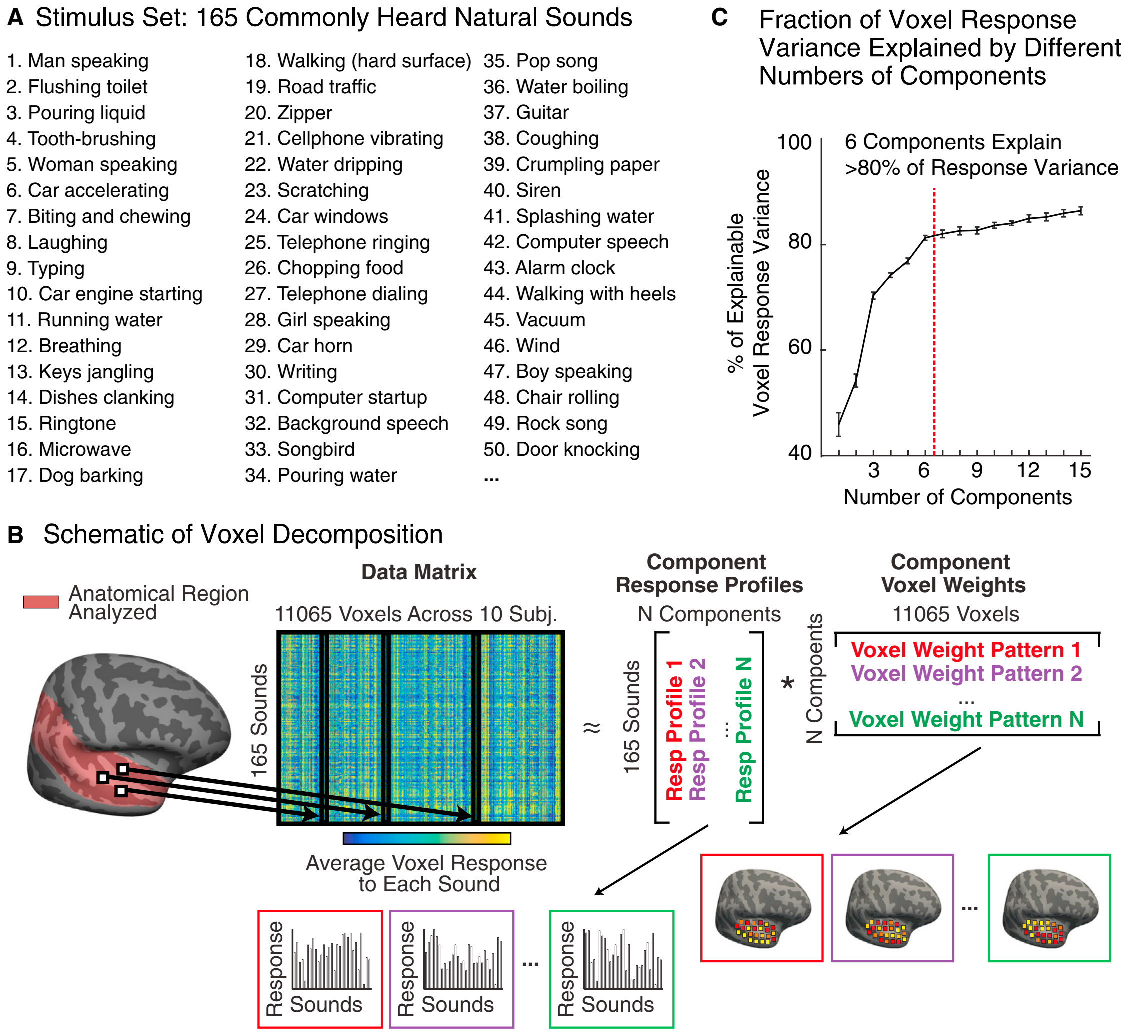


Figure 1. Voxel Decomposition Analysis

(A) Cortical responses to 165 commonly heard natural sounds were measured in human auditory cortex using fMRI. Fifty of the 165 sounds are listed, ordered by the frequency with which they were judged to be heard in daily life.

(B) The average response of each voxel to each sound was represented as a matrix (165 sounds × 11,065 voxels across all ten subjects). Each column contains the response of a single voxel to all 165 sounds. Each voxel's response was modeled as the weighted sum of a set of canonical "response profiles." This decomposition can be expressed as a factorization of the data matrix into a response matrix and a weight matrix. Response profiles and weights were inferred using statistical criteria alone, without using any information about the sounds or anatomical positions of the voxels.

(C) The proportion of voxel response variance explained by different numbers of components (see also Figure S1). The figure plots the median variance-explained across voxels, calculated separately for each subject and then averaged across the ten subjects from Experiment I; error bars plot one standard error of the mean across subjects. Variance estimates are noise-corrected (see Supplemental Experimental Procedures). Six components were sufficient to account for more than 80% of the noise-corrected variance.

RESULTS

Experiment I: Modeling Voxel Responses to Commonly Heard Natural Sounds

We measured the average response of voxels throughout auditory cortex to a diverse collection of 165 natural sounds (Figure 1A). The sound set included many of the most frequently heard and recognizable sounds that humans regularly encounter. We modeled each voxel's response as the weighted combination of a set of "components" (Figure 1B). Each component was defined by a response profile across the 165 sounds and a vector of weights across the voxels, specifying the contribution of that response profile to each voxel. Notably, no information about either the specific sounds or the anatomical positions of voxels was used to infer components. Thus, any consistent structure that emerges from the analysis must be driven by structure in the data and not by prior hypotheses about specific functional selectivities or their anatomical distribution.

Voxel Decomposition

The 165-dimensional response vectors from all voxels in all subjects were concatenated to form the data matrix (11,065 voxels across all ten subjects). To infer components, we searched for matrix factorizations that could approximate the data matrix as the product of two smaller matrices: a "response" matrix and a "weight" matrix (see Figure 1B). The response matrix expresses the response of each inferred component to each sound (165 sounds × N components), and the weight matrix expresses the contribution of each component to each voxel (N components × 11,065 voxels).

For a given number of components, the factorization is not unique and must be constrained by additional criteria. We constrained the factorization with assumptions about the distribution of component weights across voxels. We took advantage of the fact that summing independent random variables tends to produce a quantity that is closer to Gaussian-distributed. Thus, if voxel responses are a weighted sum of components with non-Gaussian weight distributions across voxels, the components should be identifiable as those whose weight distributions deviate most from Gaussianity. We searched for components with non-Gaussian weight distributions using two different algorithms. The first algorithm, a variant of independent components analysis (Hyvärinen, 1999), quantified deviations from Gaussianity using a non-parametric measure of non-Gaussianity ("negentropy"). The second algorithm used a non-Gaussian prior on the distribution of voxel weights (the Gamma distribution) and searched for response profiles that maximized the likelihood of the data given this prior. Both methods recovered components with non-Gaussian voxel weights that explained most of the reliable voxel response variance, providing empirical support for the assumption that the components underlying the data are distributed in a non-Gaussian manner. The specific response profiles and voxel weights inferred by each method were very similar, indicating that the results are robust to the specific statistical criterion used. We focus our discussion on the results of the first method because it is faster, more standard, and does not depend on a specific parameterization of the data.

The only free parameter in the analysis is the number of components. We found that six components were sufficient to explain more than 80% of the replicable voxel response variance (Figure 1C). Moreover, cross-validated prediction accuracy was best using just six components, indicating that components beyond the sixth were primarily driven by fMRI noise that did not replicate across scans (Figure S1). We focused on these first six components in all subsequent analyses.

We first describe the anatomical distribution of each component, obtained by projecting its voxel weights back into anatomical coordinates. We then describe the acoustic and semantic features of sounds that explained the response profile of each component. We refer to the components using numbers that reflect how much of their response could be accounted for by standard acoustic measures (1 being the most and 6 being the least, as explained below).

Component Voxel Weights Plotted in Anatomical Coordinates

We examined the component anatomy using group maps of the voxel weights. Maps were computed by aligning each subject to a standardized anatomical template, averaging the voxel weights for each component across subjects, and transforming this average weight into a measure of statistical significance (see Supplemental Experimental Procedures). For comparison, we identified tonotopic gradients using responses to pure tones. A group tonotopic map exhibited the two mirror-symmetric gradients widely observed in primary auditory cortex (Figure 2A) (Humphries et al., 2010). Figure 2B plots component weight maps with outlines of high- and low-frequency primary fields overlaid (see Figure S2 for weight maps from the parametric model). Tonotopic maps and voxel weights from individual subjects were generally consistent with the group results (Figure S3). As a summary, Figure 2C plots outlines of the regions with highest weight for each component.

Although no anatomical information was used to infer components, voxel weights for each component were significantly correlated across subjects ($p < 0.001$; permutation test). The weights systematically varied in their overlap and proximity to primary auditory cortex (as defined tonotopically). Components 1 and 2 primarily explained responses in low- and high-frequency tonotopic fields of primary auditory cortex (PAC), respectively. Components 3 and 4 were localized to distinct regions near the border of PAC: concentrated anteriorly and posteriorly, respectively. Components 5 and 6 concentrated in distinct non-primary regions: Component 5 in the superior temporal gyrus, lateral to PAC, and Component 6 in the planum polare, anterior to PAC, as well as in the left planum temporale, posterior to PAC.

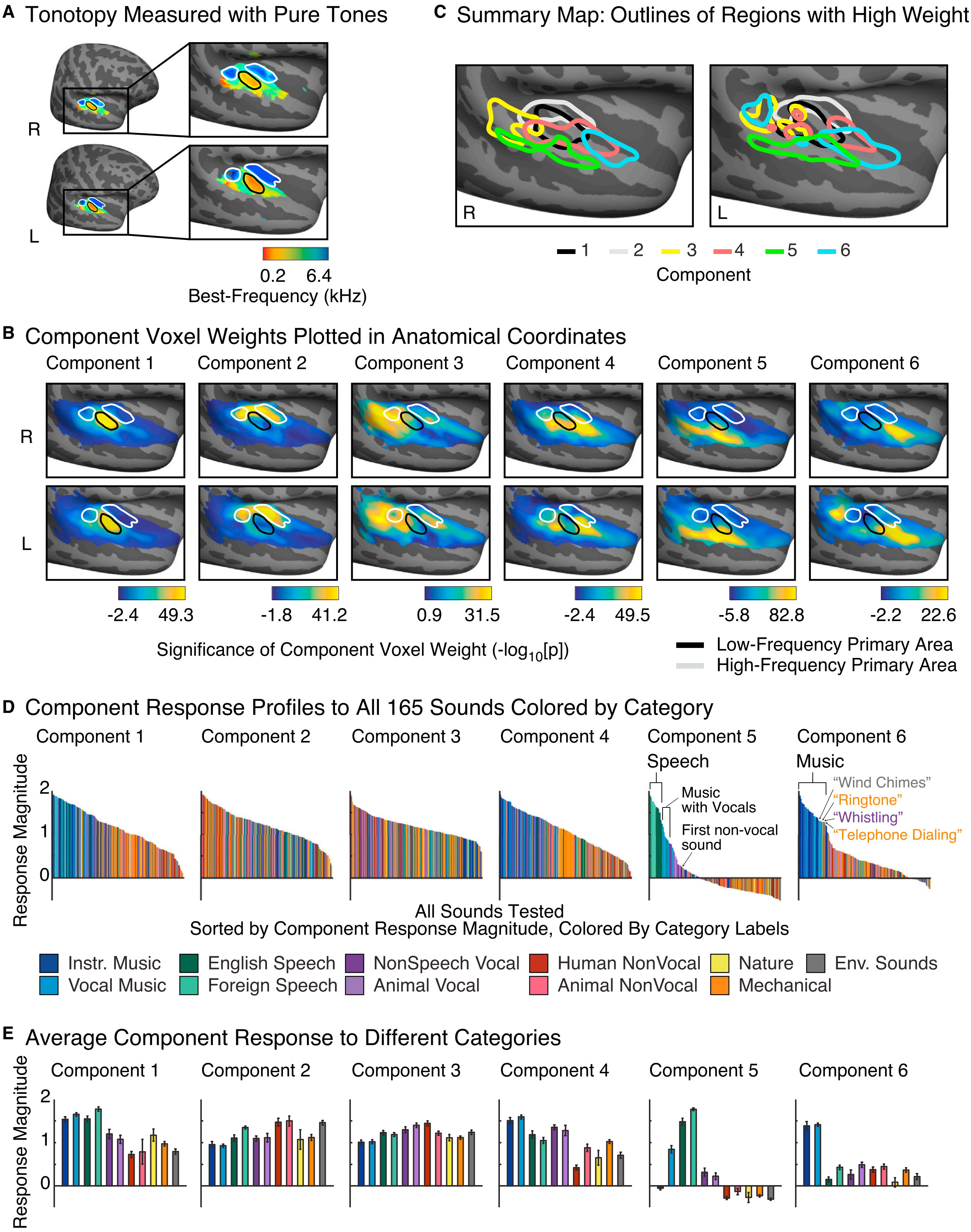
All of the components had a largely bilateral distribution; there were no significant hemispheric differences in the average weight for any of the components (Figure S4). There was a non-significant trend for greater weights in the left hemisphere of Component 6 ($t(9) = 2.21$; $p = 0.055$), consistent with the left-lateralized posterior region evident in the group map.

Component Response Profiles and Selectivity for Sound Categories

Figure 2D plots the full response profile of each inferred component to each of the 165 tested sounds. Sounds are colored based on their membership in one of 11 different categories. These profiles were reliable across independent fMRI scans (see Experimental Procedures; test-retest correlation: r = 0.94, 0.88, 0.70, 0.93, 0.98, 0.92 for Components 1–6, respectively; Figure S5A). The response profiles were also relatively robust to the exact sounds tested (Figure S5B): for randomly subsampled sets of 100 sounds, the profiles inferred were highly correlated with those inferred using all 165 sounds (median correlation > 0.95 across subsampled sound sets for all six components).

Figure 2E plots the average response of each component to sounds with the same category label (assigned based on an online survey; see Experimental Procedures). Components 1–4 responded substantially to all of the sound categories. In contrast, Components 5 and 6 responded selectively to sounds categorized as speech and music, respectively. Category labels accounted for more than 80% of the explainable response variance in these two components.

For Component 5, all of the sounds that produced a high response were categorized as "English speech" or "foreign



(legend on next page)

speech," with the next-highest response category being vocal music (which also had speech content due to lyrics). The response to non-speech vocalizations (human or animal) was higher than the response to non-vocal sounds, but substantially lower than the response to speech. Notably, responses to foreign speech were at least as high as responses to English speech, even though all of the participants were native English speakers (this remained true after excluding responses to foreign languages that subjects had studied for at least 1 year). Component 5 thus responded selectively to sounds with speech structure, regardless of whether the speech conveyed linguistic meaning.

Component 6, in contrast, responded primarily to sounds categorized as music: of the 30 sounds with the highest response, all but two were categorized as musical sounds by participants. Even the two exceptions were melodic: "wind chimes" and "ringtone" (categorized as "environmental" and a "mechanical" sounds, respectively). Other non-musical sounds produced a low response, even those with pitch (e.g., speech).

The anatomical distribution of these components (Figure 2B) suggests that speech- and music-selective responses are concentrated in distinct regions of non-primary auditory cortex, with speech selectivity lateral to primary auditory cortex and music selectivity anterior and posterior to primary auditory cortex. We emphasize that these components were determined by statistical criteria alone—no information about sound category or anatomical position contributed to their discovery. These results provide evidence that auditory cortex contains distinct anatomical pathways for the analysis of music and speech.

Response Correlations with Acoustic Measures

We next explored the acoustic sensitivity of each component, both to better understand their response properties and to test whether the selectivity of Components 5 and 6 for speech and music could be explained by standard acoustic features. First, we visualized the acoustic structure of the sounds that produced the highest and lowest response for each component by plotting their "cochleograms"—time-frequency decompositions, similar to spectrograms, intended to summarize the cochlea's representation of sound (Figure 3A). We then computed the correlation of each component's response profile with acoustic measures of frequency and spectrotemporal modulation for each sound (Figures 3B and 3C).

These analyses revealed that some of the components could be largely explained by standard acoustic features. Component 1 produced a high response for sounds with substantial low-frequency energy (Figures 3A and 3B; $p < 0.001$, permutation test), consistent with the anatomical distribution of its voxel weights, which concentrated in the low-frequency field of PAC (Figure 2B). Conversely, Component 2 responded preferentially to sounds with high-frequency energy ($p < 0.001$) and overlapped the high-frequency fields of PAC. This result demonstrates that our method can infer a well-established feature of auditory cortical organization.

Components 3 and 4 were primarily selective for patterns of spectrotemporal modulation in the cochleograms for each sound. The sounds eliciting the highest response in Component 3 were composed of broadband events that were rapidly modulated in time, evident as vertical streaks in the cochleograms. In contrast, the sounds eliciting the highest response in Component 4 all contained pitch, evident in the cochleograms as horizontal stripes, or spectral modulations, reflecting harmonics. The contrast between these two components is suggestive of a tradeoff in sensitivity to spectral versus temporal modulations (Singh and Theunissen, 2003). Accordingly, the response profile of Component 3 correlated most with measures of fast temporal modulation and coarse-scale spectral modulation ($p < 0.01$), while that of Component 4 correlated with measures of fine spectral modulation and slow temporal modulation ($p < 0.001$) (Figure 3C). We also observed significant modulation tuning in Components 1 and 2 (for fine spectral and rapid temporal modulations, respectively; $p < 0.001$), beyond that explained by their frequency tuning (frequency measures were partialled out before computing the modulation correlations). We note that although components 1-2 and 3-4 appear to have opposite tuning properties, their response profiles were not strongly anti-correlated, and they were thus identifiable as distinct components.

Prior studies have argued that the right and left hemispheres are differentially specialized for spectral and temporal resolution, respectively (Zatorre et al., 2002). Contrary to this hypothesis, Components 1–4 exhibited qualitatively similar patterns of voxel weights in the two hemispheres (Figure 2B), with no significant hemispheric differences when tested individually. However, the small biases present were in the expected direction (Figure S4), with a right-hemisphere bias for Components 1 and 4 and a left-hemisphere bias for Components 2 and 3. When these laterality differences (Right-Left) were pooled and directly compared ($[C1 + C4] - [C2 + C3]$), a significant difference

Figure 2. Component Voxel Weights and Response Profiles

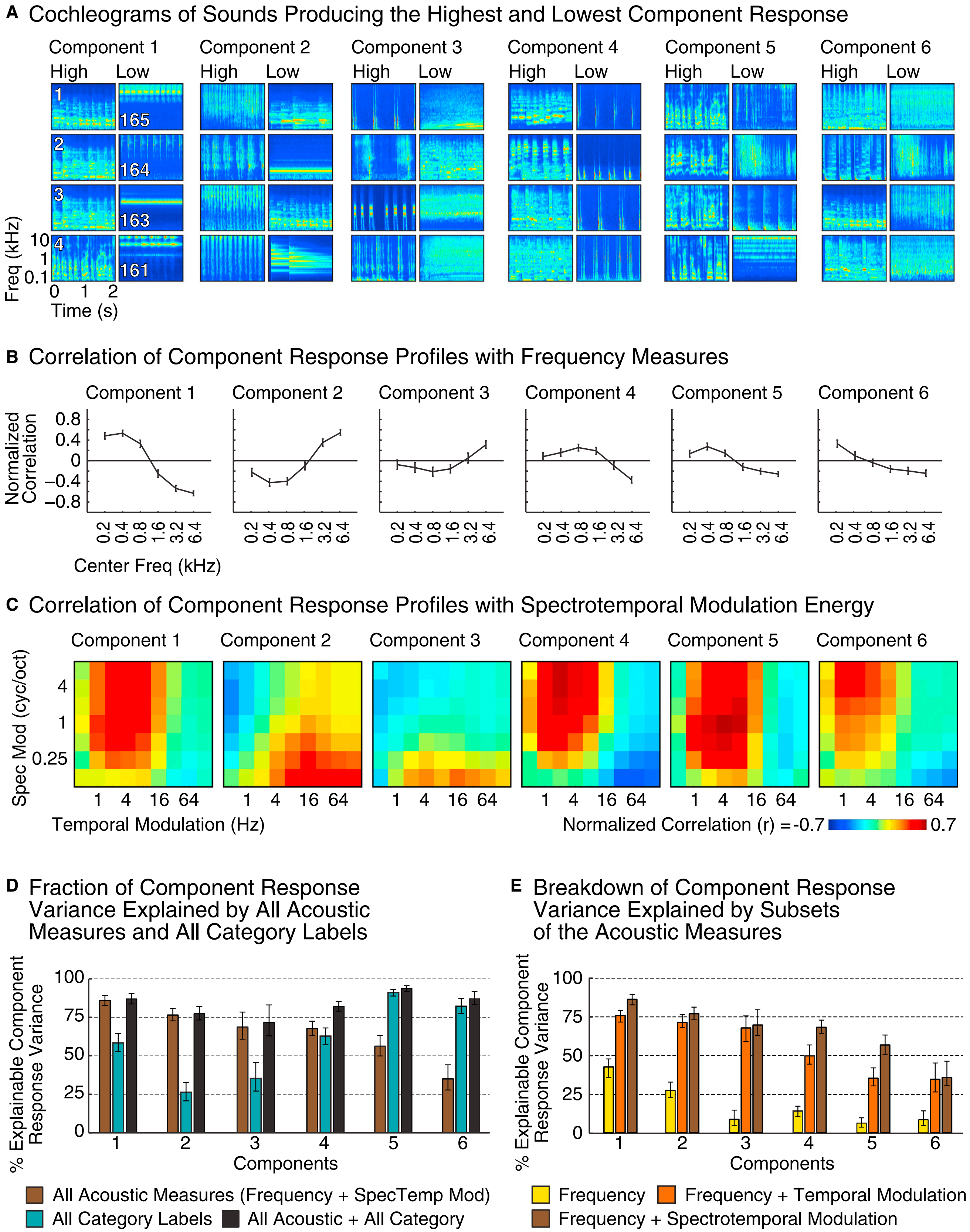
(A) Tonotopy measured using responses to pure tones. High- and low-frequency regions of primary auditory cortex are outlined with white and black outlines, respectively.

(B) Component voxel weights, averaged across subjects aligned to a standardized anatomical template, and transformed to a measure of significance via a permutation test across the sound set. Each map plots logarithmically transformed p values ($-\log_{10}[p]$), signed such that positive values indicate positive weights and negative values indicate negative weights. Color scales span the central 95% of the p value distribution for each component. Outlines of high- and low-frequency regions within primary auditory cortex are overlaid. See Figure S2B for weight maps inferred using the Parametric Model, Figure S3 for individual subject weight maps, and Figure S4 for a quantification of hemispheric differences.

(C) Summary map showing outlines of the 10% of voxels with the highest weight for each component.

(D) Response profiles for the inferred components. Each figure plots the response magnitude (arbitrary units) of each component to all 165 sounds tested. Sounds are ordered by response magnitude and colored based on their membership in one of 11 different categories, assigned based on the judgments of human listeners. Components 5 and 6 responded selectively to sounds categorized as speech and music, respectively. See Figure S2C for response profiles inferred using the Parametric Model and Figure S5 for measures of response profile reliability.

(E) Component responses averaged across sounds from the same category. Error bars plot one standard error of the mean across sounds from a category, computed using bootstrapping (10,000 samples).



A Cochleograms of Sounds Producing the Highest and Lowest Component Response

B Correlation of Component Response Profiles with Frequency Measures

C Correlation of Component Response Profiles with Spectrotemporal Modulation Energy

D Fraction of Component Response Variance Explained by All Acoustic Measures and All Category Labels

E Breakdown of Component Response Variance Explained by Subsets of the Acoustic Measures

(legend on next page)

emerged (t(9) = 2.47; p < 0.05). These results are consistent with the presence of hemispheric biases in spectral and temporal modulation sensitivity, but show that this bias is quite small relative to within-hemisphere differences.

Collectively, measures of frequency and modulation energy accounted for much of the response variance in Components 1–4 (Figure 3D; 86%, 76%, 68%, and 67%, respectively; see Figure 3E for the variance explained by subsets of acoustic measures). Category labels explained little to no additional variance for these components. In contrast, for Components 5 and 6, category labels explained substantially more variance than the acoustic features (p < 0.001), and when combined, acoustic features explained little additional variance beyond that explained by the categories. Thus, the selectivity of Components 5 and 6 for speech and music sounds cannot be explained by standard acoustic features.

Experiment II: Speech and Music Scrambling

Music and speech are both notable for having distinct and recognizable structure over relatively long timescales. One approach to probing sensitivity to temporal structure is to reorder short sound segments so that local but not global structure is preserved (Abrams et al., 2011). A recent study introduced "quilting" for this purpose—a method for reordering sound segments while minimizing acoustic artifacts (Overath et al., 2015)—and demonstrated that regions in the superior temporal gyrus respond preferentially to intact compared with quilt-scrambled speech. We used the same procedure to provide an additional test of the selectivity of our components.

We measured responses to intact and scrambled speech and music in the same subjects scanned in Experiment I. As a result, we could use the component voxel weights from Experiment I to infer the response of each component to the new stimulus conditions from Experiment II (see Experimental Procedures). For Components 1–4, there was little difference in the response to intact and scrambled sounds for either category (Figures 4A and 4B). In contrast, Component 5 responded more to intact than scrambled speech (t(7) = 7.24, p < 0.001) and Component 6 responded more to intact than scrambled music (t(7) = 6.05, p < 0.001), producing a three-way interaction between category (speech, music), scrambling, and components (F(1,5) = 7.37, p < 0.001). This result provides further evidence that Components 5 and 6 respond selectively to speech and music structure, respectively.

Searching for Music-Selective Responses with Standard Methods

There are few prior reports of highly selective responses to musical sounds (Angulo-Perkins et al., 2014; Leaver and Rauschecker, 2010). One possible explanation is that prior studies have tested for music selectivity in raw voxel responses. If music-selective neural populations overlap within voxels with other neural populations, the music selectivity of raw voxel responses could be diluted. Component analysis should be less vulnerable to such overlap because voxels are modeled as the weighted sum of multiple components. To test this possibility, we directly compared the response of the music-selective component (Component 6) with the response of the voxels most selective for music (Figure 5) (see Experimental Procedures). We found that the selectivity of these voxels for musical structure was notably weaker than that observed for the music-selective component across a number of metrics. First, the response profiles to the sound set were graded for the voxels but closer to binary for the component (i.e., high for music, low for non-music) (Figure 5A). Second, acoustic features and category labels explained similar amounts of response variance in music-selective voxels, unlike the component, in which category labels explained substantially more variance than acoustic features (Figure 5B). Third, although music-selective voxels responded slightly less to scrambled music (Figure 5C; t(7) = 4.82, p < 0.01), the effect was much larger in Component 6, producing a significant interaction between the effect of scrambling (intact versus scrambled) and the type of response being measured (component versus voxel) (p < 0.01). The ability to decompose voxel responses into their underlying components was thus critical to identifying neural selectivity for music.

We observed similar but less pronounced trends when comparing speech-selective voxels with the speech-selective component (Component 5): speech-selective voxels exhibited robust selectivity for speech sounds (Figure S6) that could not be accounted for by standard acoustic features. This finding suggests that speech selectivity is more anatomically segregated than music selectivity and thus easier to identify in raw voxel responses.

Selectivity of Voxels for Individual Components

The lack of clear music selectivity in raw voxels suggests that at least some components spatially overlap. We performed two analyses to quantify the extent of overlap between components. First, we assessed the selectivity of voxels for individual components (Figure 6A): for each voxel the weight for a single component was normalized by the sum of the absolute values of the weights for all six components. Normalized weights near 1 indicate voxels that weight strongly on a single component. Figure 6B plots normalized weights averaged across the top N % of voxels with the most significant weight along each individual component (varying N; permutation test, see Experimental Procedures). Independent data were used to select voxels in

Figure 3. Component Correlations with Acoustic Measures

(A) Cochleograms of the four sounds producing the highest and lowest response in each component. Cochleograms plot an estimate of cochlear response magnitudes for a sound as a function of time and frequency.

(B) Correlation of component response profiles with energy in different frequency bands.

(C) Correlation of component response profiles with spectrotemporal modulation energy in the cochleograms for each sound.

(D) Total amount of component response variation explained by (1) all acoustic measures, (2) all category labels, and (3) the combination of acoustic measures and category labels. For Components 1–4, category labels explained little additional variance beyond that explained by acoustic features. For Components 5 and 6, category labels explained most of the response variance, and acoustic features accounted for little additional variance.

(E) Breakdown of the component response variation explained by subsets of the acoustic measures.

Correlation coefficients and measures of explained variance were noise-corrected (see Supplemental Experimental Procedures). Error bars in all panels plot standard errors across the sound set (via bootstrap).

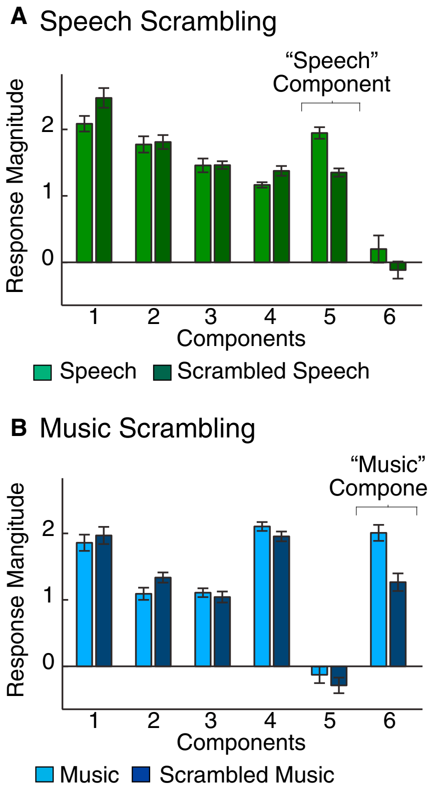


Figure 4. Experiment II: Testing for Category-Selective Responses via Scrambling

(A) Component responses (arbitrary units) to intact and temporally scrambled speech (via "quilting", see Experimental Procedures).

(B) Component responses to intact and scrambled music.

Error bars in both panels plot one standard error of the mean across subjects.

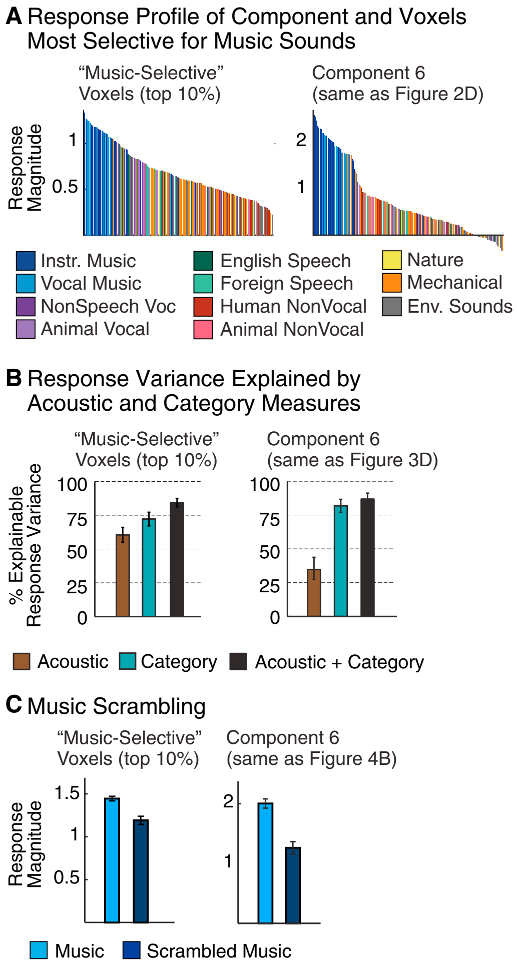


Figure 5. Analyses of Music Selectivity in Raw Voxels

(A) Left: the average response profile (in units of percent signal change) of voxels with the most significant response preference for sounds categorized as music (i.e., music > non-music). Sounds are ordered by response magnitude and colored by category. Right: the response profile of Component 6 (arbitrary units), which responded selectively to music.

(B) The amount of response variance explainable by acoustic features, category labels, and their combination. Error bars plot standard errors across the sound set (via bootstrap).

(C) Response to intact and scrambled music. Error bars plot one standard error of the mean across subjects.

See Figure S6 for analogous plots of speech selectivity measured in raw voxels.

individual subjects and measure their component weights to avoid statistical bias/circularity. As a summary, inset pie charts show normalized weights averaged across the top 10% of voxels.

The highest normalized weights were observed for Component 5 (speech selective) in the superior temporal gyrus (Figure 6A), consistent with the robust speech selectivity we observed in raw voxels (Figure S6). The top 10% of voxels with the most significant weight for Component 5 had an average normalized weight of 0.70 (Figure 6B), and thus most of their response was explained by Component 5 alone. By contrast, there were no voxels with similarly high normalized weights for Component 6 (music selective), consistent with the weak music selectivity observed in raw voxels (Figure 5). The top 10% of voxels for Component 6 (average normalized weight of 0.49) also had substantial weight from Component 4 (normalized weight of 0.20; Figure 6B), which responded preferentially to sounds with pitch. This finding is consistent with the anatomical distribution of these components, both of which overlapped a region anterior to primary auditory cortex (Figures 2B and 2C).

Testing Assumptions of Non-Gaussianity

Our voxel component analysis relied on assumptions about the distribution of weights across voxels to constrain the factorization of the data matrix. The key assumption of our approach is that these weight distributions are non-Gaussian. This assumption raises two questions: first, does the assumption hold for the voxel responses we analyzed, and second, what properties of cortical responses might give rise to non-Gaussian voxel weights?

To evaluate whether the non-Gaussian assumption was warranted for our dataset, we relied on the fact that linear combinations of Gaussian variables remain Gaussian. As a consequence, our method would only have been able to infer components with non-Gaussian voxel weights if the components that generated

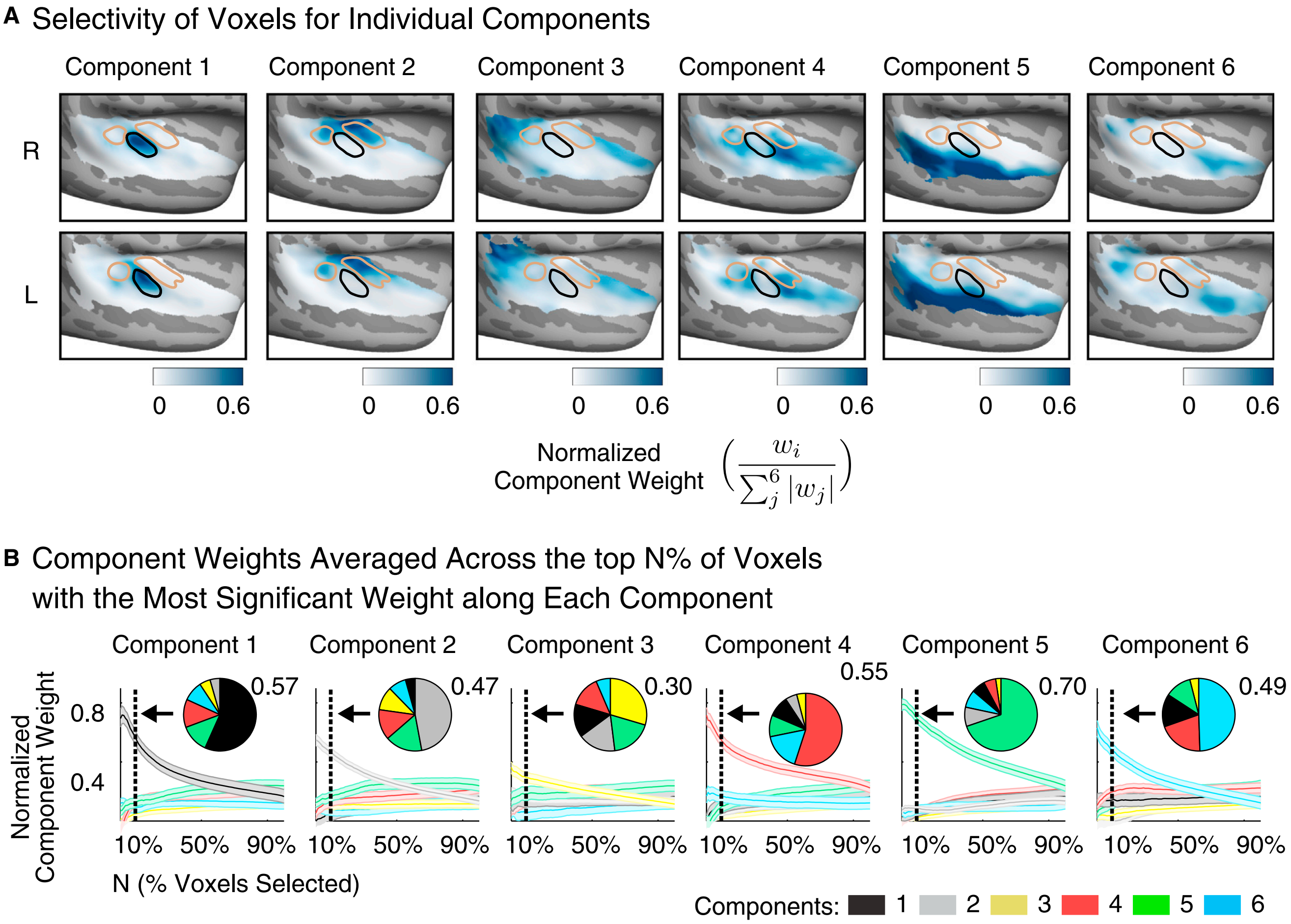


Figure 6. Selectivity of Voxels for Individual Components

(A) Group maps plotting component weights for each voxel, normalized by the sum of the absolute values of the weights for all six components. These normalized weights provide an estimate of the selectivity of each voxel for individual components.

(B) Component weights averaged across the top N% of voxels with the most significant weight along each component. Averaged component weights were subsequently normalized using the same equation shown in (A). Data used to measure weights were independent from those used to compute significance. Error bars plot one standard error of the mean, computed via bootstrapping across the sound set. As a summary, inset pie charts show normalized weights averaged across the top 10% of voxels.

the data also had non-Gaussian weights. We thus tested whether the voxel weights for the inferred components were significantly non-Gaussian (evaluated in independent data).

For all six components, the distribution of weights was significantly more skewed and kurtotic (sparse) than the Gaussian distribution (Figure 7A). As a result, a modified Gaussian distribution with flexible skew and sparsity (the four-parameter "Johnson" distribution) provided a significantly better fit to the weight distributions than the Gaussian (Figure 7B) (as measured by the log-likelihood of left-out data; $p < 0.01$ in all cases, via bootstrapping). These results show that all of the components inferred by our analysis are indeed non-Gaussian by virtue of being skewed and sparse, validating a key assumption underlying our approach (see also Figure S7).

Why would the distribution of neural selectivities in the brain be skewed and sparse? In practice, we found that the anatomical distributions of the component weights were spatially clustered. If neurons with similar response properties are spatially clustered in the brain, they should contribute substantially to only a small fraction of voxels, producing skewed and sparse weight distributions. Skew and sparsity may thus be useful statistical signatures for identifying components from fMRI responses, due to anatomical clustering of neurons with similar response selectivities.

DISCUSSION

Our findings reveal components of neuronal stimulus selectivity that collectively explain fMRI responses to natural sounds throughout human auditory cortex. Each component has a distinct response profile across natural sounds and a distinct spatial distribution across the cortex. Four components reflected selectivity for standard acoustic dimensions (Figure 3), such as frequency, pitch, and spectrotemporal modulation. Two other components were highly selective for speech and music (Figures

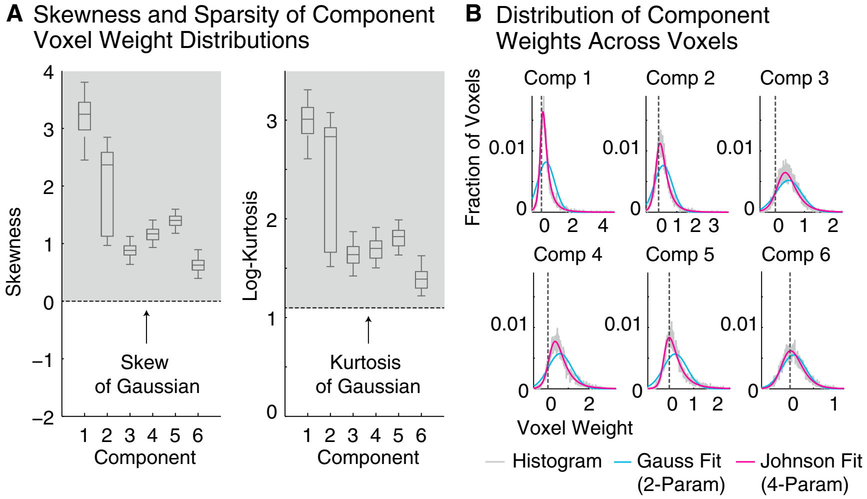


Figure 7. Statistical Properties of Component Voxel Weights

(A) Skewness and log-kurtosis (a measure of sparsity) for each component. All components were skewed and sparse relative to a Gaussian. Box-and-whisker plots show the central 50% (boxes) and central 96% (whiskers) of the distribution for each statistic (via bootstrapping across subjects).

(B) Histograms of voxel weights for each component (gray). Weight distributions were fit using a Gaussian distribution (blue) as well as a modified Gaussian ("Johnson") distribution with flexible skew and kurtosis (magenta). For all components, the Johnson distribution fit the measured voxel weights significantly better than the Gaussian. Fits were evaluated using left-out data. See Figure S7 for another test of non-Gaussianity.

2D and 2E). The response of these two components could not be explained by standard acoustic measures, and their specificity for speech and music was confirmed with hypothesis-driven experiments that probed sensitivity to category-specific temporal structure (Figure 4). The selective responses we observed for music have little precedent (Angulo-Perkins et al., 2014; Leaver and Rauschecker, 2010), and our analyses suggest an explanation: the music-selective component spatially overlapped with other components (Figure 6). As a result, music selectivity was not clearly evident in raw voxel responses, which are the focus of most fMRI analyses (Figure 5). Anatomically, the acoustically driven components (Components 1–4) concentrated in and around primary auditory cortex, whereas speech and music-selective components concentrated in distinct non-primary regions (Figure 2B). This pattern suggests that representations of music and speech diverge in non-primary areas of human auditory cortex.

Our findings were enabled by a novel approach for inferring neural response dimensions (Figure 1). Our method searches the space of possible response profiles to natural stimuli for those that best explain voxel responses. The method is blind to the properties of each sound and the anatomical position of each voxel, but the components it infers can be examined post hoc to reveal tuning properties and functional organization. The method revealed both established properties of auditory cortical organization, such as tonotopy (Da Costa et al., 2011; Humphries et al., 2010), as well as novel properties not evident with standard methods.

Voxel Decomposition

Our method falls into a family of recent computational approaches that seek to uncover functional organization from responses to large sets of naturalistic stimuli. One prior approach has been to model voxel responses to natural stimuli using candidate sets of stimulus features ("encoding models"; Huth et al., 2012; Mitchell et al., 2008; Moerel et al., 2013). Such models can provide insights into the computations underlying neural activity, but require a prior hypothesis about the stimulus features encoded in voxel responses. Our approach is complementary: it searches for canonical response profiles to the stimulus set that collectively explain the response of many voxels without requiring prior hypotheses about the stimulus features that underlie their response (Vul et al., 2012). While there is no guarantee that voxel responses will be explained by a small number of response profiles or that the profiles will be interpretable, we found that auditory voxels could be explained by six components that each reflected selectivity for particular acoustic or semantic properties.

An additional benefit of our approach is its ability to express voxel responses as the combination of distinct underlying components, potentially related to neural sub-populations. We used linear decomposition techniques to infer components because the mapping between hemodynamic activity and the underlying neural response is thought to be approximately linear (Boynton et al., 1996). Such techniques have previously been used to analyze fMRI time courses (Beckmann and Smith, 2004), typically to reveal large-scale brain systems based on "resting state" activity (Mantini et al., 2007). In contrast, our method decomposes stimulus-driven voxel responses to natural stimuli to reveal functional organization within a sensory system.

The non-parametric algorithm we used to recover components is closely related to standard algorithms for "independent component analysis" (Bell and Sejnowski, 1995; Hyvärinen, 1999) and "sparse coding" (Olshausen and Field, 1997), both of which rely on measures of non-Gaussianity to infer structure in data. Notably, we found that all of the components inferred by the non-parametric algorithm had skewed and sparse distributions (Figure 7A). This finding does not reflect an assumption of the method, because our algorithm could in principle find any non-Gaussian distribution, including those less sparse than a Gaussian. Similar results were obtained using a parametric model that explicitly assumed a skewed and sparse prior on the voxel weights (Figure S2), providing evidence that the results are robust to the specific statistical criterion used.

Although six components were sufficient to capture most of the replicable variation in our experiment (Figure 1C; Figure S1), this result does not imply that auditory cortical responses are spanned by only six dimensions. Instead, the number of components detectable by our analysis is likely to reflect three factors: the resolution of fMRI, the amount of noise in fMRI measurements, and the variation in our stimulus set along different neural response dimensions. Thus, the dimensions inferred likely reflect dominant sources of response variation across commonly heard natural sounds.

Selectivity for Music

Despite longstanding interest in the brain basis of music (Abrams et al., 2011; Fedorenko et al., 2012; Koelsch et al., 2005; Rogalsky et al., 2011; Tierney et al., 2013), there is little precedent for neural responses specific to music (Angulo-Perkins et al., 2014; Leaver and Rauschecker, 2010). One reason is the small number of conditions tested in most fMRI experiments, which limits the ability to distinguish responses to music from responses to other acoustic features (e.g., pitch). Our results suggest a second reason: voxel responses underestimate neuronal selectivity if different neural populations overlap at the scale of voxels, since each voxel reflects the pooled response of hundreds of thousands of neurons. We found that the music-selective component exhibited consistently higher selectivity than did the most music-selective voxels (Figure 5), due to overlap with other components that have different tuning properties (Figure 6). Voxel decomposition was thus critical to isolating music selectivity. The anatomical distribution of the music-selective component our method revealed was nonetheless consistent with prior neuroimaging studies that have implicated anterior regions of auditory cortex in music processing (Angulo-Perkins et al., 2014; Fedorenko et al., 2012; Leaver and Rauschecker, 2010; Tierney et al., 2013) and with prior neuropsychology studies that have reported selective deficits in music perception after focal lesions (Peretz et al., 1994).

Selectivity for Speech

Our analysis also revealed a component that responded selectively to speech (Component 5), whose anatomical distribution was consistent with prior studies (e.g., Hickok and Poeppel, 2007; Scott et al., 2000). The response properties and anatomy of this component are consistent with a recent study that reported larger responses to intact compared with temporally scrambled foreign speech in the superior temporal gyrus (Overath et al., 2015). Our findings extend this prior work by demonstrating that: (1) speech-selective regions are highly selective, responding much less to over 100 other non-speech sounds, and (2) speech-selective regions in the mid-STG show little to no response preference for linguistically meaningful utterances, in contrast with putatively downstream regions in lateral temporal and frontal cortex (Fedorenko et al., 2011; Friederici, 2012). This component may thus reflect an intermediate processing stage that encodes speech-specific structure (e.g., phonemes and syllables), independent of linguistic intelligibility.

The anatomy of this component also resembles that of putative "voice-selective" areas identified in prior studies (Belin et al., 2000). Notably, the component responded substantially more to speech sounds than to non-speech vocal sounds (e.g., crying, laughing) (Fecteau et al., 2004), suggesting that speech structure is the primary driver of its response. However, our results do not reveal the specific speech features or properties that drive its response and do not preclude the coding of vocal identity.

Selectivity for Acoustic Features

Four components had response profiles that could be largely explained by standard acoustic features. Two of these components (1 and 2) reflected tonotopy, one of the most widely cited organizing dimensions of the auditory system. Consistent with prior reports (Da Costa et al., 2011; Humphries et al., 2010), the tonotopic gradient we observed was organized in a V-shaped pattern surrounding Heschl's Gyrus. We also observed tonotopic gradients beyond primary auditory cortex (Figure S3), but these were weaker than those in primary areas.

Component responses were also tuned to spectrotemporal modulation. The distinct tuning properties of different components were suggestive of a tradeoff in selectivity for spectral and temporal modulation (Rodríguez et al., 2010; Singh and Theunissen, 2003). Components 1 and 4 responded preferentially to fine spectral modulation and slow temporal modulation (characteristic of sounds with pitch), while Components 2 and 3 responded preferentially to coarse spectral modulation and rapid temporal modulation. Anatomically, the components selective for fine spectral modulation clustered near anterior regions of Heschl's Gyrus, whereas those selective for fine temporal modulation clustered in more posterior-medial regions of Heschl's gyrus and the planum temporale. On average the components sensitive to fine spectral modulations (1 and 4) were slightly more right-lateralized than the components sensitive to rapid temporal modulations (2 and 3), consistent with a well-known hypothesis of hemispheric specialization (Zatorre et al., 2002). However, all components exhibited much greater variation within hemispheres than across hemispheres. These results are consistent with a prior study that measured modulation tuning using natural sounds (Santoro et al., 2014).

One of the acoustically responsive components (4) was functionally and anatomically similar to previously identified pitch-responsive regions (Norman-Haignere et al., 2013; Patterson et al., 2002; Penagos et al., 2004). These regions respond primarily to "resolved harmonics," the dominant cue to human pitch perception, and are localized to anterolateral regions of auditory cortex, partially overlapping low-frequency but not high-frequency tonotopic areas.

Implications for the Functional Organization of Auditory Cortex

A key question animating debates on auditory functional organization is the extent to which the cortex is organized hierarchically (Chevillet et al., 2011; Hickok and Poeppel, 2007; Staeren et al., 2009). Many prior studies have reported increases in response complexity in non-primary areas relative to that in primary auditory cortex (PAC) (Chevillet et al., 2011; Obleser et al., 2007; Petkov et al., 2008), potentially reflecting the abstraction of behaviorally relevant features from combinations of simpler responses. Consistent with this idea, simple acoustic features predicted the response of components in and around primary auditory cortex (Components 1–4), while components overlapping non-primary areas (Components 5 and 6) responded selectively to sound categories and could not be explained by frequency and modulation statistics.

Models of hierarchical processing have often posited the existence of distinct ‘‘streams’’ within non-primary areas (Lomber and Malhotra, 2008; Rauschecker and Scott, 2009). For example, regions ventral to PAC have been implicated in the recognition of spectrotemporal patterns (Hickok and Poeppel, 2007; Lomber and Malhotra, 2008), while regions dorsal to PAC have been implicated in spatial computations (Miller and Recanzone, 2009; Rauschecker and Tian, 2000) and processes related to speech production (Dhanjal et al., 2008). Although our findings do not speak to the locus of spatial processing (because sound location was not varied in our stimulus set), they suggest an alternative type of organization based on selectivity for important sound categories (Leaver and Rauschecker, 2010), with speech encoded lateral to PAC (reflected by Component 5) and music encoded anterior/posterior to PAC (reflected by Component 6). Our results speak less definitively to the representation of other natural sounds. But the posterior distribution of Component 3, which responded to a wide range of sound categories, is consistent with a third processing stream for the analysis of environmental sounds.

Conclusions and Future Directions

The organization we observed was inferred without any prior functional or anatomical hypotheses, suggesting that organization based on speech and music is a dominant feature of cortical responses to natural sounds. These findings raise a number of further questions. Is the functional organization revealed by our method present from birth? Do other species have homologous organization? What sub-structure exists within speech- and music-selective cortex? Voxel decomposition provides a natural means to answer these questions, as well as analogous questions in other sensory systems.

EXPERIMENTAL PROCEDURES

Experiment I: Measuring Voxel Responses to Commonly Heard Natural Sounds

Participants

Ten individuals (4 male, 6 female, all right-handed, ages 19–27) completed two scan sessions (each ~1.5 hr); eight subjects completed a third session. Subjects were non-musicians (no formal training in the 5 years preceding the scan), native English speakers, and had self-reported normal hearing. Three other subjects were excluded due to excessive motion or sporadic task performance. The decision to exclude these subjects was made before analyzing their data to avoid potential bias. The study was approved by MIT's human subjects review committee (COUHES); all participants gave informed consent.

Stimuli

We determined from pilot experiments that we could measure reliable responses to 165 sounds in a single scan session. To generate our stimulus set, we began with a set of 280 everyday sounds for which we could find a recognizable, 2-second recording. Using an online experiment (via Amazon's Mechanical Turk), we excluded sounds that were difficult to recognize (below 80% accuracy on a ten-way multiple choice task; 55–60 participants for each sound), yielding 238 sounds. We then selected a subset of 160 sounds that were rated as most frequently heard in everyday life (in a second Mechanical Turk study; 38–40 ratings per sound). Five additional ‘‘foreign speech’’ sounds were included (‘‘German,’’ ‘‘French,’’ ‘‘Italian,’’ ‘‘Russian,’’ ‘‘Hindi’’) to distinguish responses to acoustic speech structure from responses to linguistic structure.

Procedure

Sounds were presented using a ‘‘block design’’ that we found produced reliable voxel responses in pilot experiments. Each block included five repetitions of the same 2-second sound. After each 2-second sound, a single fMRI volume was collected (‘‘sparse sampling’’). Each scan acquisition lasted 1 second, and stimuli were presented during a 2.4 s interval between scans. Because of the large number of sounds tested, each scan session included only a single block per sound. Despite the small number of block repetitions, the inferred components were highly reliable (Figure S5A).

Blocks were grouped into 11 ‘‘runs,’’ each with 15 stimulus blocks and 4 blocks of silence with no sounds. Silence blocks were the same duration as the stimulus blocks and were spaced evenly throughout the run.

To encourage subjects to attend equally to all of the sounds, subjects performed a task in which they detected a change in sound level. In each block, one of the five sounds was 7 dB lower than the others. Subjects were instructed to a press a button when they heard the quieter sound (never the first sound in the block). The magnitude of the level change (7 dB) was selected to produce good performance in attentive participants given the intervening fMRI noise. Sounds were presented through MRI-compatible earphones (Sensimetrics S14) at 75 dB SPL (68 dB for the quieter sounds).

Data acquisition and preprocessing used standard procedures (see Supplemental Experimental Procedures). We estimated the average response of each voxel to each stimulus block (five repetitions of the same sound) by averaging the response of the second through fifth scans after the onset of each block (the first scan was excluded to account for hemodynamic delay). Results were similar using a GLM instead of signal averaging to estimate voxel responses. Signal-averaged responses were converted to percent signal change by subtracting and dividing by each voxel's response to blocks of silence. These PSC values were subsequently downsampled to a 2 mm isotropic grid (on the FreeSurfer-flattened cortical surface).

Voxel Selection

For the decomposition analysis, we selected voxels with a consistent response to the sounds from a large anatomical constraint region encompassing the superior temporal and posterior parietal cortex (Figure 1B). We used two criteria: (1) a significant response to sounds compared with silence ($p < 0.001$) and (2) a reliable response pattern to the 165 sounds across scans 1 and 2 (note that component reliability was quantified using independent data from scan 3; see Supplemental Experimental Procedures). The reliability measure we used is shown in Equation 1. This measure differs from a correlation in assigning high values to voxels with a consistent response to the sound set, even if the response does not vary greatly across sounds. Such responses are found in many voxels in primary auditory cortex, and using the correlation across scans to select voxels would cause many of these voxels to be excluded:

$$r = 1 - \frac{\|\mathbf{v}_1 - \mathrm{proj}_{\mathbf{v}_2}\mathbf{v}_1\|}{\|\mathbf{v}_1\|} \qquad \text{(Equation 1)}$$

$$\mathrm{proj}_{\mathbf{v}_2}\mathbf{v}_1 = \mathbf{v}_2\left(\frac{\mathbf{v}_2^T}{\|\mathbf{v}_2\|}\mathbf{v}_1\right) \qquad \text{(Equation 2)}$$

where $\mathbf{v}_1$ and $\mathbf{v}_2$ indicate the response vector of a single voxel to the 165 sounds measured in two different scans, and $\| \, \|$ is the L2 norm. The numerator in the second term of Equation 1 is the magnitude of the residual left in $\mathbf{v}_1$ after projecting out the response shared by $\mathbf{v}_2$. This ‘‘residual magnitude’’ is divided by its maximum possible value (the magnitude of $\mathbf{v}_1$). The reliability measure is thus bounded between 0 and 1.

For the component analysis, we included voxels with a reliability of 0.3 or higher, which amounted to 64% of sound-responsive voxels. Although our results were robust to the exact setting of this parameter, restricting the analysis to reliable voxels improved the reliability of the inferred components, helping to compensate for the relatively small amount of data collected per sound.

Experiment II: Measuring Voxel Responses to Scrambled Music and Speech

Participants

A subset of eight subjects from Experiment I participated in Experiment II (4 male, 4 female, all right-handed, ages 22–28).

Stimuli

The intact speech sounds were 2 s excerpts of German utterances from eight different speakers (4 male, 4 female). We used foreign speech to isolate responses to acoustic speech structure, independent of linguistic meaning (Overath et al., 2015). Two of the subjects tested had studied German in school, and for one of these subjects, we used Russian utterances instead of German utterances. The other subject was tested with German because the Russian stimuli were not available at the time of the scan. The inclusion or exclusion of their data did not change the results. The intact music stimuli were 2-second excerpts from eight different "big band" musical recordings.

Speech and music stimuli were scrambled using the "quilting" algorithm described by Overath et al. (2015). Briefly, the algorithm divides a source signal into non-overlapping 30 ms segments. These segments are then re-ordered with the constraint that segment-to-segment cochleogram changes are matched to those of the original recordings. The reordered segments are concatenated using pitch-synchronous-overlap-and-add (PSOLA) so as to avoid boundary artifacts.

Procedure

Stimuli were presented in a block design with five stimuli from the same condition presented in series, with fMRI scan acquisitions interleaved (as in Experiment I). Subjects performed a "1-back" task to help maintain their attention on the sounds: in each block, four sounds were unique (i.e., different 2-second excerpts from the same condition), and one sound was an exact repetition of the sound that came before it. Subjects were instructed to press a button after the repeated sound.

Each "run" included 2 blocks per condition. The number of runs was determined by the amount of time available in each scanning session. Five subjects completed three runs, two subjects completed four runs, and one subject completed two runs. All other methods details were the same as Experiment I.

Voxel Decomposition Methods

Overview of Decomposition

We approximated the data matrix, **D** (165 sounds × 11,065 voxels), as the product of a response matrix, **R** (165 sounds × N components), and a weight matrix, **W** (N components × 11,065 voxels):

$$\mathbf{D} \approx \mathbf{RW} \qquad \text{(Equation 3)}$$

We used two methods to factorize the data matrix: a "non-parametric" algorithm that searches for maximally non-Gaussian weights (quantified using a measure of entropy) and a parametric model that maximizes the likelihood of the data matrix given a non-Gaussian prior on voxel weights. The two methods produced qualitatively similar results. The main text presents results from the non-parametric algorithm, which we describe first. A MATLAB implementation of both algorithms is available on the authors' websites, along with all of the stimuli.

Non-Parametric Decomposition Algorithm

The non-parametric algorithm is similar to ICA algorithms that search for components with non-Gaussian distributions by minimizing the entropy of the weight distribution (because the Gaussian distribution has highest entropy for a fixed variance). The key difference between our method and standard algorithms (e.g., "FastICA") is that we directly estimated entropy via a histogram method (Moddemeijer, 1989), rather than using a "contrast function" designed to approximate entropy. For example, many ICA algorithms use kurtosis as a metric for non-Gaussianity, which is useful if the latent distributions are non-Gaussian due to their sparsity, but not if the non-Gaussianity results from skew. Directly estimating negentropy makes it possible to detect any source of non-Gaussianity. Our approach was enabled by the large number of voxels (>10,000), which made it possible to robustly estimate entropy using a histogram.

The algorithm had two main steps. First, the data matrix was reduced in dimensionality and whitened using PCA. Second, the whitened and reduced data matrix was rotated to maximize negentropy (J), defined as the difference in entropy between the empirical distribution and a Gaussian distribution of the same variance:

$$J(\boldsymbol{y}) = H(\boldsymbol{y}_{gauss}) - H(\boldsymbol{y}) \qquad \text{(Equation 4)}$$

The first step was implemented using singular value decomposition, which approximates the data matrix using the top N principal components with highest variance:

$$\mathbf{D} \approx \mathbf{USV} \qquad \text{(Equation 5)}$$

where **U** is the response matrix for the top N principal components with highest variance (165 sounds × N components), **V** is the weight matrix for these components (N components × 11,065 voxels), and **S** is a diagonal matrix of singular values (N × N). The number of components, N, was chosen by measuring the variance explained by different numbers of components and the accuracy of components in predicting voxel responses in left-out data (see Supplemental Experimental Procedures).

In the second step, we found a rotation of the principal component weight matrix (V from Equation 5 above) that maximized the negentropy summed across components (Hyvärinen, 1999):

$$\widehat{\mathbf{T}} = \underset{\mathbf{T}}{\operatorname{argmax}} \sum_{c=1}^{N} J(\mathbf{W}[c,:]), \quad \text{where } \mathbf{W} = \mathbf{TV} \qquad \text{(Equation 6)}$$

where W is the rotated weight matrix (N × 11,065), T is an orthonormal rotation matrix (N × N), and $\mathbf{W}[c,:]$ is the cth row of W. We estimated negentropy using a histogram-based method (Moddemeijer, 1989) applied to the voxel weight vector for each component ($\mathbf{W}[c,:]$).

We optimized this function by iteratively selecting pairs of components and finding the rotation that maximized their negentropy (using grid-search over all possible rotations; see Figure S7). This pairwise optimization was repeated until no rotation could further increase the negentropy. All pairwise rotations were then combined into a single rotation matrix ($\widehat{\mathbf{T}}$), which we used to compute the response profiles (**R**) and voxel weights (W):

$$\mathbf{R} = \mathbf{US}\widehat{\mathbf{T}}^{-1} \qquad \text{(Equation 7)}$$

$$\mathbf{W} = \widehat{\mathbf{T}}\mathbf{V} \qquad \text{(Equation 8)}$$

Parametric Decomposition Model

The non-parametric algorithm just described, like many ICA algorithms, constrained the voxel weights to be uncorrelated, a necessary condition for independence (Hyvärinen, 1999). Although this constraint greatly simplifies the algorithm, it could conceivably bias the results if the neural components that generated the data have voxel weights that are correlated. To address this issue, we repeated all our analyses using a second algorithm that did not constrain the weights to be uncorrelated. The algorithm placed a non-Gaussian prior (the Gamma distribution) on the distribution of voxel weights and searched for response profiles that maximized the likelihood of the data, integrating across all possible weights. For computational tractability, the prior on voxel weights was factorial. However, the posterior distribution over voxel weights, given data, was not constrained to be independent or uncorrelated, and could thus reflect statistical dependencies between the component weights.

This second approach is closely related to sparse coding algorithms (Olshausen and Field, 1997), which infer basis functions (components) assuming a sparse prior on the component weights. Such methods typically assume a fixed prior for all components. This assumption seemed suboptimal for our purposes because the components inferred using the non-parametric algorithm varied in skew/sparsity (Figure 7A). Instead, we developed an alternative approach, which inferred a separate prior distribution for each component, potentially accommodating different degrees of sparsity in different neural sub-populations.

Our approach was inspired by a method developed by Liang et al., 2014 to factorize spectrograms. The Liang et al. method was a useful starting point because it allows the prior distribution on weights to vary across components (see Figure S2A). Like Liang et al., we used a single-parameter Gamma distribution to model latent variables (the weights in our case) because it can fit many non-negative distributions depending on the shape parameter. Unlike Liang et al., we modeled measurement noise with a Gaussian distribution rather than a Gamma because the Gaussian fit our empirical noise estimates better. We also used a different algorithm to optimize the model (stochastic Expectation-Maximization) (Dempster et al., 1977; Wei and

Tanner, 1990), which we found to be more accurate when tested on simulated data. The mathematical details of the model and the optimization algorithm used to infer components are described in Supplemental Experimental Procedures.

Analyses of Component Response Properties and Anatomy

Component Voxel Weight Anatomy

We averaged voxel weights across subjects in standardized anatomical coordinates (FreeSurfer's FsAverage template) (Figure 2B). Voxel weights were smoothed with a 5 mm FWHM kernel on the cortical surface prior to averaging. Voxels without a reliable response pattern to the sound set, after averaging across the ten subjects tested, were excluded. The inclusion criteria were the same as that used to select voxels from individual subjects. We transformed these average weight maps into a map of statistical significance using a permutation test across the sound set (Nichols and Holmes, 2002) (see Supplemental Experimental Procedures).

To verify that the weight maps were more similar across subjects than would be expected by chance, we measured the average correlation between weight maps across all pairs of subjects for the same component. We compared this correlation with a null distribution generated by randomly permuting the correspondence between components across subjects (10,000 permutations).

To test for laterality effects, we compared the average voxel weight for each component in the left and right hemisphere of each subject (Figure S4) using a paired t test across subjects.

Sound Category Assignments

In an online experiment, Mechanical Turk participants chose the category that best described each sound, and we assigned each sound to its most frequently chosen category (30–33 participants per sound) (Figures 2D and 2E). Category assignments were highly reliable (split-half kappa = 0.93).

Acoustic Features

Cochleograms were measured using a bank of band-pass filters (McDermott and Simoncelli, 2011), similar to a gammatone filter bank (Slaney, 1998) (Figure 3A). There were 120 filters spaced equally on an $\mathrm{ERB_N}$ scale between 20 Hz and 10 kHz (87.5% overlap, half-cycle cosine frequency response). Each filter was intended to model the response of a different point along the basilar membrane. Acoustic measurements were computed from the envelopes of these filter responses (the magnitude of the analytic signal, raised to the 0.3 power to model cochlear compression).

Because voxels were represented by their average response to each sound, we used summary acoustic measures, averaged across the duration of each sound, to predict component response profiles. For each feature, we correlated a vector of acoustic measures with the response profile of each component. To estimate the variance explained by sets of acoustic features, we regressed sets of feature vectors against the response profile of each component (see Supplemental Experimental Procedures). Both correlations and variance-explained estimates were corrected for noise in fMRI measurements (see Supplemental Experimental Procedures).

As a measure of audio frequency, we averaged cochlear envelopes over the 2-second duration of each sound. Because the frequency tuning of voxels is broad relative to cochlear filters (e.g., Humphries et al., 2010), we summed these frequency measures within six octave-spaced frequency ranges (centered on 200, 400, 800, 1,600, 3,200, and 6,400 Hz). The frequency ranges were non-overlapping, and the lowest and highest bands were lowpass and highpass, respectively. We measured the amount of energy in each frequency band for each sound, after subtracting the mean for each sound across the six bands. This demeaned vector was then correlated with the response profile for each component (Figure 3B).

We used a spectrotemporal modulation filter bank (Chi et al., 2005) to measure the energy at different temporal "rates" (in Hz) and spectral "scales" (in cycles per octave) for each sound. The filter bank crossed nine octave-spaced rates (0.5–128 Hz) with seven octave-spaced scales (0.125–8 cyc/oct). Each filter was complex-valued (real and imaginary parts were in quadrature phase). Cochleograms were zero-padded (2 seconds) prior to convolution with each filter. For each rate/scale, we correlated the average magnitude of the filter response for each sound with the component response profiles (Figure 3C) after partialling out correlations with the audio frequency measures just described. We averaged the magnitude of "negative" and "positive" temporal rates (i.e., left and right quadrants of the 2D Fourier Transform), because their pattern of correlations was very similar. Temporal modulation was computed from the same model (Chi et al., 2005) using filters modulated in time, but not frequency.

We used a permutation test to assess whether the correlation values across a set of acoustic measures differed significantly (Figures 3B and 3C). As in a one-way ANOVA, the variance of the correlation across a set of acoustic measures was compared with that for a null distribution (here computed by permuting the mapping between acoustic features and response profiles).

Measuring Component Responses to Scrambled Speech and Music

We used the pseudoinverse of the component voxel weights from Experiment I ($\mathbf{W}_{\mathbf{ExpI}}$) to estimate the response of each component to the stimulus conditions from Experiment II ($\mathbf{R}_{\mathbf{ExpII}}$) (Figure 4):

$$\mathbf{R}_{\mathrm{ExpII}} = \mathbf{D}_{\mathrm{ExpII}} \mathbf{W}_{\mathrm{ExpI}}^{T} \left(\mathbf{W}_{\mathrm{ExpI}} \mathbf{W}_{\mathrm{ExpI}}^{T} \right)^{-1} \qquad \text{(Equation 9)}$$

where $\mathbf{D}_{\mathbf{ExpII}}$ is a matrix containing the response of each voxel to each condition from Experiment II. We measured component responses separately for each of the eight subjects and used ANOVAs and t tests to evaluate significance.

Identifying Music- and Speech-Selective Voxels

We identified music-selective voxels by contrasting responses to music and non-music sounds (Figure 5) using regression with a binary category vector on data from scan 1. To control for correlations with acoustic measures, we included our acoustic feature vectors (see above) as nuisance regressors. We then selected the 10% of voxels from each subject with the most significant regression weight for the music versus non-music contrast, measured using ordinary least-squares. Similar results were obtained using different thresholds (5% or 15%). Voxel responses were then measured using data from scans 2 and 3. The same analysis was used to identify speech voxels, by contrasting responses to speech and non-speech sounds (Figure S6).

Component Overlap within Voxels

To calculate the normalized voxel weights plotted in Figure 6A, we standardized the response profiles to have the same units by setting the variance of each profile to 1. Both the response profiles and voxels were demeaned so that the overall response of each voxel to the sound set would not affect its relative selectivity for different components. We then regressed the component response profiles against the voxel responses and averaged these regression weights across subjects (in standardized anatomical coordinates). Finally, the regression weights for each component were normalized by the sum of the absolute weights for all six components (separately for each voxel):

$$\frac{\omega_i}{\sum_{j=1}^{6} \left| \omega_j \right|}. \qquad \text{(Equation 10)}$$

We note that variability in the anatomical distribution of components across subjects could lead to lower selectivity values; to mitigate this concern, we also quantified selectivity in voxels from individual subjects (Figure 6B). Specifically, we (1) ranked voxels from each subject by their weight along a single component, (2) selected the top N% of voxels from this list (varying N), (3) averaged component weights (for all six components) across the selected voxels and across subjects (in that order), and (4) normalized these average weights using Equation 10. Error bars were computed via bootstrapping across the sound set (Efron and Efron, 1982).

To avoid statistical bias/circularity in this procedure, the data used to select voxels was independent of that used to measure their component weights. Data from the first two scans of each subject was used to infer components and select voxels with high weight for a single component. We selected voxels using a measure of the significance of their weights (p values from the permutation test described above). Data from a third, independent scan was then used to estimate component weights in the selected voxels.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and seven figures and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2015.11.035.

ACKNOWLEDGMENTS

We thank Matt Hoffman and Polina Golland for helpful discussions and Ed Vul, Elias Issa, Alex Kell, and Simon Kornblith for comments on an earlier draft of this paper. This study was supported by the National Science Foundation (Research Fellowship to S.N.-H.), the National Eye Institute (EY13455 to N.G.K.), and the McDonnell Foundation (Scholar Award to J.H.M.).

Received: June 1, 2015
Revised: October 3, 2015
Accepted: November 23, 2015
Published: December 16, 2015

REFERENCES

Abrams, D.A., Bhatara, A., Ryali, S., Balaban, E., Levitin, D.J., and Menon, V. (2011). Decoding temporal structure in music and speech relies on shared brain resources but elicits different fine-scale spatial patterns. Cereb. Cortex *21*, 1507–1518.

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

Barton, B., Venezia, J.H., Saberi, K., Hickok, G., and Brewer, A.A. (2012). Orthogonal acoustic dimensions define auditory field maps in human cortex. Proc. Natl. Acad. Sci. USA *109*, 20738–20743.

Beckmann, C.F., and Smith, S.M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans. Med. Imaging *23*, 137–152.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. Nature *403*, 309–312.

Bell, A.J., and Sejnowski, T.J. (1995). An information-maximization approach to blind separation and blind deconvolution. Neural Comput. *7*, 1129–1159.

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

Boynton, G.M., Engel, S.A., Glover, G.H., and Heeger, D.J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. J. Neurosci. *16*, 4207–4221.

Chevillet, M., Riesenhuber, M., and Rauschecker, J.P. (2011). Functional correlates of the anterolateral processing hierarchy in human auditory cortex. J. Neurosci. *31*, 9345–9352.

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

Da Costa, S., van der Zwaag, W., Marques, J.P., Frackowiak, R.S.J., Clarke, S., and Saenz, M. (2011). Human primary auditory cortex follows the shape of Heschl's gyrus. J. Neurosci. *31*, 14067–14075.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Methodol. *39*, 1–38.

Dhanjal, N.S., Handunnetthi, L., Patel, M.C., and Wise, R.J. (2008). Perceptual systems controlling speech production. J. Neurosci. *28*, 9969–9975.

Efron, B., and Efron, B. (1982). The jackknife, the bootstrap and other resampling plans (SIAM).

Engel, L.R., Frum, C., Puce, A., Walker, N.A., and Lewis, J.W. (2009). Different categories of living and non-living sound-sources activate distinct cortical networks. Neuroimage *47*, 1778–1791.

Fecteau, S., Armony, J.L., Joanette, Y., and Belin, P. (2004). Is voice processing species-specific in human auditory cortex? An fMRI study. Neuroimage *23*, 840–848.

Fedorenko, E., Behr, M.K., and Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. Proc. Natl. Acad. Sci. USA *108*, 16428–16433.

Fedorenko, E., McDermott, J.H., Norman-Haignere, S., and Kanwisher, N. (2012). Sensitivity to musical structure in the human brain. J. Neurophysiol. *108*, 3289–3300.

Friederici, A.D. (2012). The cortical language circuit: from auditory perception to sentence comprehension. Trends Cogn. Sci. *16*, 262–268.

Giordano, B.L., McAdams, S., Zatorre, R.J., Kriegeskorte, N., and Belin, P. (2013). Abstract encoding of auditory objects in cortical activity patterns. Cereb. Cortex *23*, 2025–2037.

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nat. Rev. Neurosci. *8*, 393–402.

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

Huth, A.G., Nishimoto, S., Vu, A.T., and Gallant, J.L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron *76*, 1210–1224.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Netw. *10*, 626–634.

Koelsch, S., Fritz, T., Schulze, K., Alsop, D., and Schlaug, G. (2005). Adults and children processing music: an fMRI study. Neuroimage *25*, 1068–1076.

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

Liang, D., Hoffman, M.D., and Mysore, G.J. (2014). A generative product-of-filters model of audio. arXiv, arXiv:1312.5857. http://arxiv.org/abs/1312.5857.

Lomber, S.G., and Malhotra, S. (2008). Double dissociation of 'what' and 'where' processing in auditory cortex. Nat. Neurosci. *11*, 609–616.

Mantini, D., Perrucci, M.G., Del Gratta, C., Romani, G.L., and Corbetta, M. (2007). Electrophysiological signatures of resting state networks in the human brain. Proc. Natl. Acad. Sci. USA *104*, 13170–13175.

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

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

Miller, L.M., and Recanzone, G.H. (2009). Populations of auditory cortical neurons can accurately encode acoustic space across stimulus intensity. Proc. Natl. Acad. Sci. USA *106*, 5931–5935.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.-M., Malave, V.L., Mason, R.A., and Just, M.A. (2008). Predicting human brain activity associated with the meanings of nouns. Science *320*, 1191–1195.

Moddemeijer, R. (1989). On estimation of entropy and mutual information of continuous distributions. Signal Processing *16*, 233–248.

Moerel, M., De Martino, F., Santoro, R., Ugurbil, K., Goebel, R., Yacoub, E., and Formisano, E. (2013). Processing of natural sounds: characterization of multipeak spectral tuning in human auditory cortex. J. Neurosci. *33*, 11888–11898.

Nichols, T.E., and Holmes, A.P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. *15*, 1–25.

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

Obleser, J., Zimmermann, J., Van Meter, J., and Rauschecker, J.P. (2007). Multiple stages of auditory speech perception reflected in event-related FMRI. Cereb. Cortex *17*, 2251–2257.

Olshausen, B.A., and Field, D.J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? Vision Res. *37*, 3311–3325.

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

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

Penagos, H., Melcher, J.R., and Oxenham, A.J. (2004). A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. J. Neurosci. *24*, 6810–6815.

Peretz, I., Kolinsky, R., Tramo, M., Labrecque, R., Hublet, C., Demeurisse, G., and Belleville, S. (1994). Functional dissociations following bilateral lesions of auditory cortex. Brain *117*, 1283–1301.

Peretz, I., Vuvan, D., Lagrois, M.-É., and Armony, J.L. (2015). Neural overlap in processing music and speech. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140090.

Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., and Logothetis, N.K. (2008). A voice region in the monkey brain. Nat. Neurosci. *11*, 367–374.

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

Rauschecker, J.P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proc. Natl. Acad. Sci. USA *97*, 11800–11806.

Rodríguez, F.A., Read, H.L., and Escabí, M.A. (2010). Spectral and temporal modulation tradeoff in the inferior colliculus. J. Neurophysiol. *103*, 887–903.

Rogalsky, C., Rong, F., Saberi, K., and Hickok, G. (2011). Functional anatomy of language and music perception: temporal and structural factors investigated using functional magnetic resonance imaging. J. Neurosci. *31*, 3843–3852.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., and Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput. Biol. *10*, e1003412.

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

Scott, S.K., Blank, C.C., Rosen, S., and Wise, R.J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. Brain *123*, 2400–2406.

Singh, N.C., and Theunissen, F.E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. J. Acoust. Soc. Am. *114*, 3394–3411.

Slaney, M. (1998). Auditory toolbox. Technical Report #1998-010. Interval Research Corporation. https://engineering.purdue.edu/~malcolm/interval/1998-010.

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

Stecker, G.C., Harrington, I.A., and Middlebrooks, J.C. (2005). Location coding by opponent neural populations in the auditory cortex. PLoS Biol. *3*, e78.

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

Vul, E., Lashkari, D., Hsieh, P.-J., Golland, P., and Kanwisher, N. (2012). Data-driven functional clustering reveals dominance of face, place, and body selectivity in the ventral visual pathway. J. Neurophysiol. *108*, 2306–2322.

Wei, G.C., and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J. Am. Stat. Assoc. *85*, 699–704.

Zatorre, R.J., Belin, P., and Penhune, V.B. (2002). Structure and function of auditory cortex: music and speech. Trends Cogn. Sci. *6*, 37–46.

SUPPLEMENTAL FIGURES

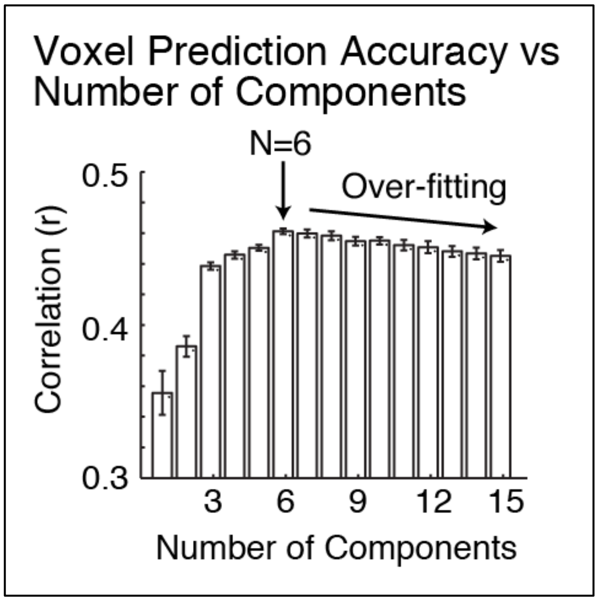


Figure S1. Voxel Prediction Accuracy vs. Number of Components

Related to Main Figure 1C

Accuracy of the component model in predicting voxel responses measured from left-out data not used to fit the model, as a function of the number of components used (see Supplemental Methods). The figure plots the median correlation between the measured and predicted response across voxels (averaged across subjects). Components driven by reliable variance will improve prediction accuracy, while components driven by noise will degrade the performance, due to over-fitting. Best performance was achieved using a model with 6 components. Error bars plot one standard error of the mean across subjects.

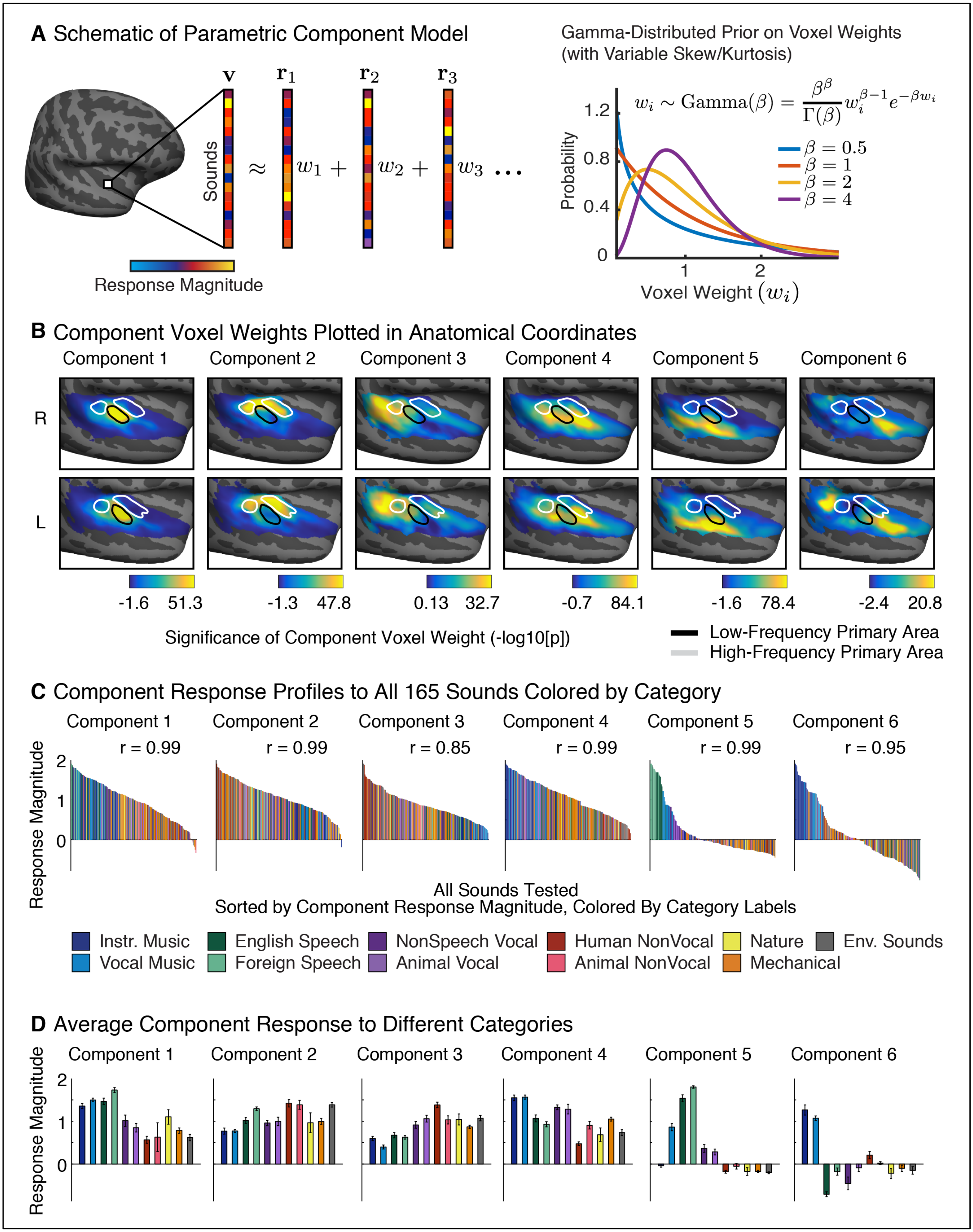

A Schematic of Parametric Component Model
$\mathbf{v}$
$\mathbf{r}_1$
$\mathbf{r}_2$
$\mathbf{r}_3$
Sounds
$\approx$
$w_1 +$
$w_2 +$
$w_3 \ldots$
Response Magnitude
Gamma-Distributed Prior on Voxel Weights
(with Variable Skew/Kurtosis)
$w_i \sim \text{Gamma}(\beta) = \frac{\beta^\beta}{\Gamma(\beta)} w_i^{\beta-1} e^{-\beta w_i}$
1.2
0.8
0.4
0
Probability
$\beta = 0.5$
$\beta = 1$
$\beta = 2$
$\beta = 4$
1
2
Voxel Weight (w_i)
B Component Voxel Weights Plotted in Anatomical Coordinates
Component 1
Component 2
Component 3
Component 4
Component 5
Component 6
R
L
-1.6
51.3
-1.3
47.8
0.13
32.7
-0.7
84.1
-1.6
78.4
-2.4
20.8
Significance of Component Voxel Weight (-log10[p])
Low-Frequency Primary Area
High-Frequency Primary Area
C Component Response Profiles to All 165 Sounds Colored by Category
Component 1
Component 2
Component 3
Component 4
Component 5
Component 6
r = 0.99
r = 0.99
r = 0.85
r = 0.99
r = 0.99
r = 0.95
2
1
0
Response Magnitude
All Sounds Tested
Sorted by Component Response Magnitude, Colored By Category Labels
Instr. Music
Vocal Music
English Speech
Foreign Speech
NonSpeech Vocal
Animal Vocal
Human NonVocal
Animal NonVocal
Nature
Mechanical
Env. Sounds
D Average Component Response to Different Categories
Component 1
Component 2
Component 3
Component 4
Component 5
Component 6
2
1
0
Response Magnitude

Figure S2. Parametric Component Model

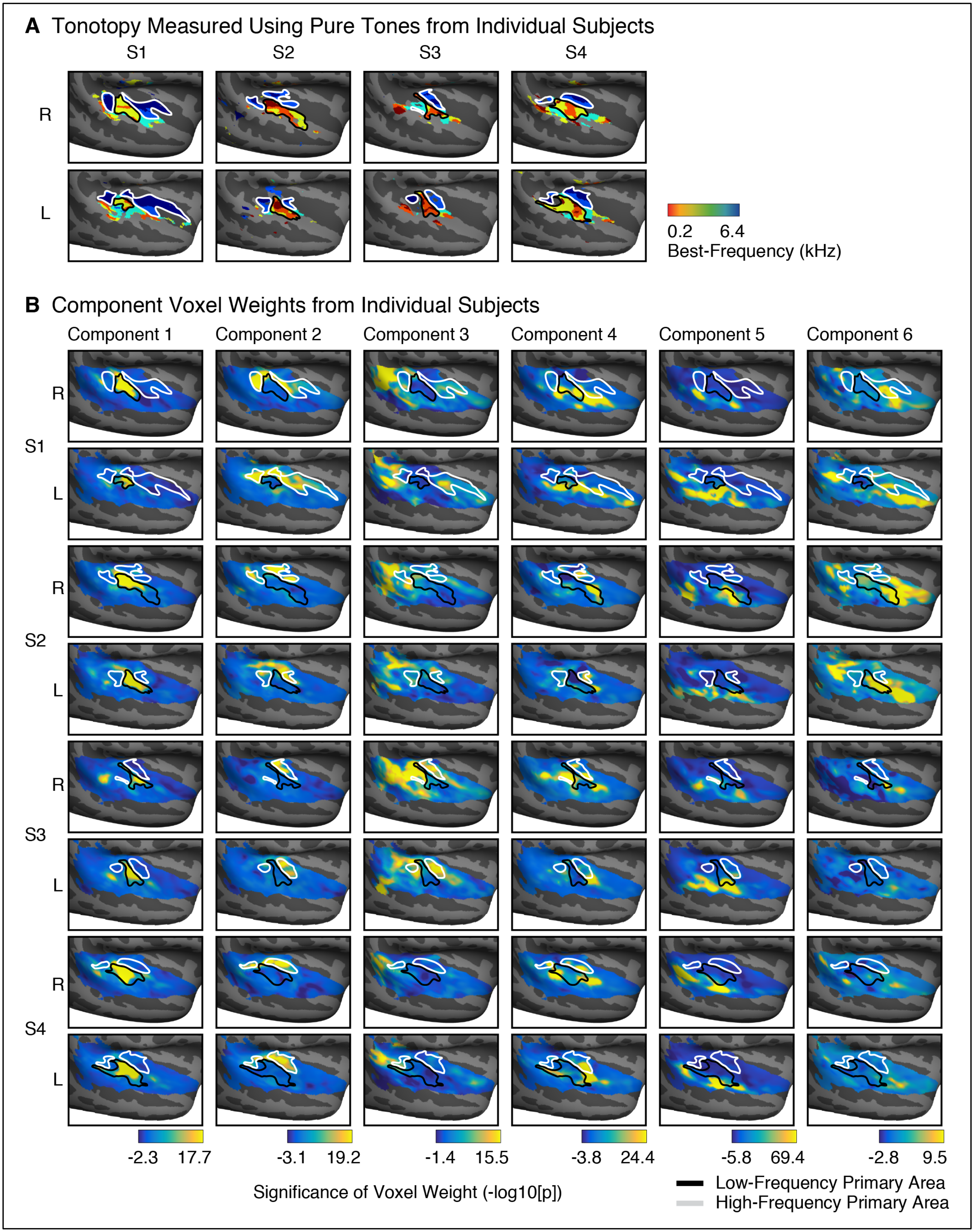
Related to Main Figure 2

(A) Model schematic: each voxel was modeled as the weighted sum of a set of response profiles ($\mathbf{r}_1$, $\mathbf{r}_2$, $\mathbf{r}_3$, …) with a Gamma-distributed prior on the voxel weights (w_1, w_2, w_3, …). The Gamma distribution constrains the weights to be positive and can model distributions with variable skewness/sparsity depending on the shape parameter (β). Because of the positivity constraint, the weights could be interpreted as reflecting the proportion of different neuronal populations present in each voxel. Components were discovered by finding response profiles and shape parameters that maximized the likelihood of the data, integrating across all possible voxel weights.

(B) Component voxel weights averaged across subjects after aligning their brains to a standardized anatomical template (same format as Figure 2B).

(C) Response profiles discovered using the parametric algorithm (same format as Figure 2D). The correlation coefficient for the best-matching profile from the non-parametric algorithm is shown. Each component discovered by the parametric algorithm was similar in both its voxel weights and response profile to a single, unique component from the non-parametric algorithm.

(D) Component responses averaged across sounds with the same category assignment (same format as Figure 2E).



A Tonotopy Measured Using Pure Tones from Individual Subjects

B Component Voxel Weights from Individual Subjects

Figure S3. Component Voxel Weights from Individual Subjects

Related to Main Figure 2B

(A) Tonotopic maps, measured with pure tones, from 4 individual subjects that participated in an extra scan session to more robustly measure tonotopy in their individual brains. Colors indicate which of six different frequency ranges best drove each voxel's response. Each subject exhibited two mirror-symmetric maps, characteristic of primary auditory cortex. High- and low-frequency regions of primary auditory cortex are outlined with white and black outlines, respectively.

(B) Component voxel weight maps from these same four subjects, with outlines of high- and low-frequency primary regions overlaid. Maps plot a measure of significance for each component and voxel (logarithmically transformed p-values, calculated via a permutation test). Color scales show the central 95% of the p-value distribution for each component.

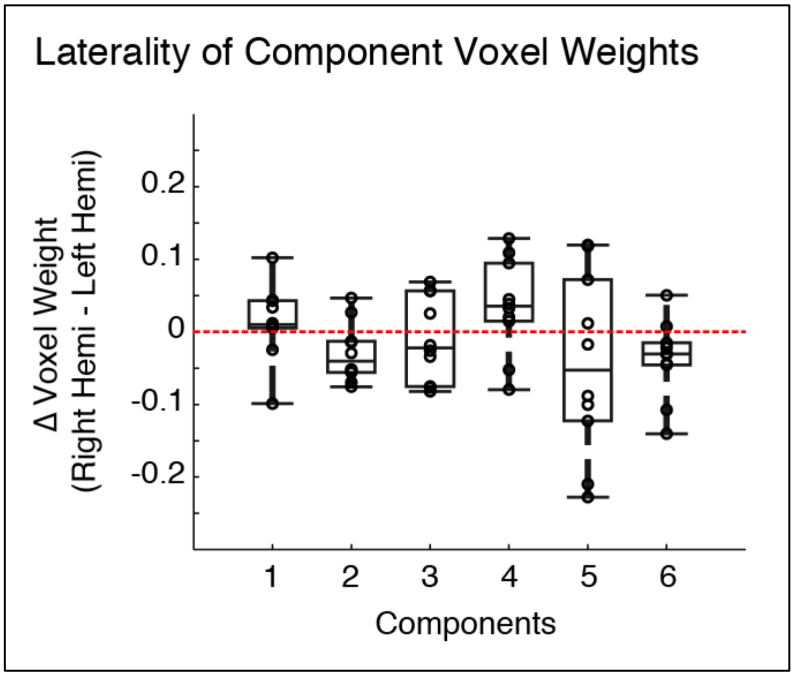


Figure S4. Laterality of Component Voxel Weights

Related to Main Figure 2B

The average difference in voxel weights between the right and left hemisphere for all six components. Circles correspond to individual subjects. Box plots show medians and the central 50% of the distribution for each component.

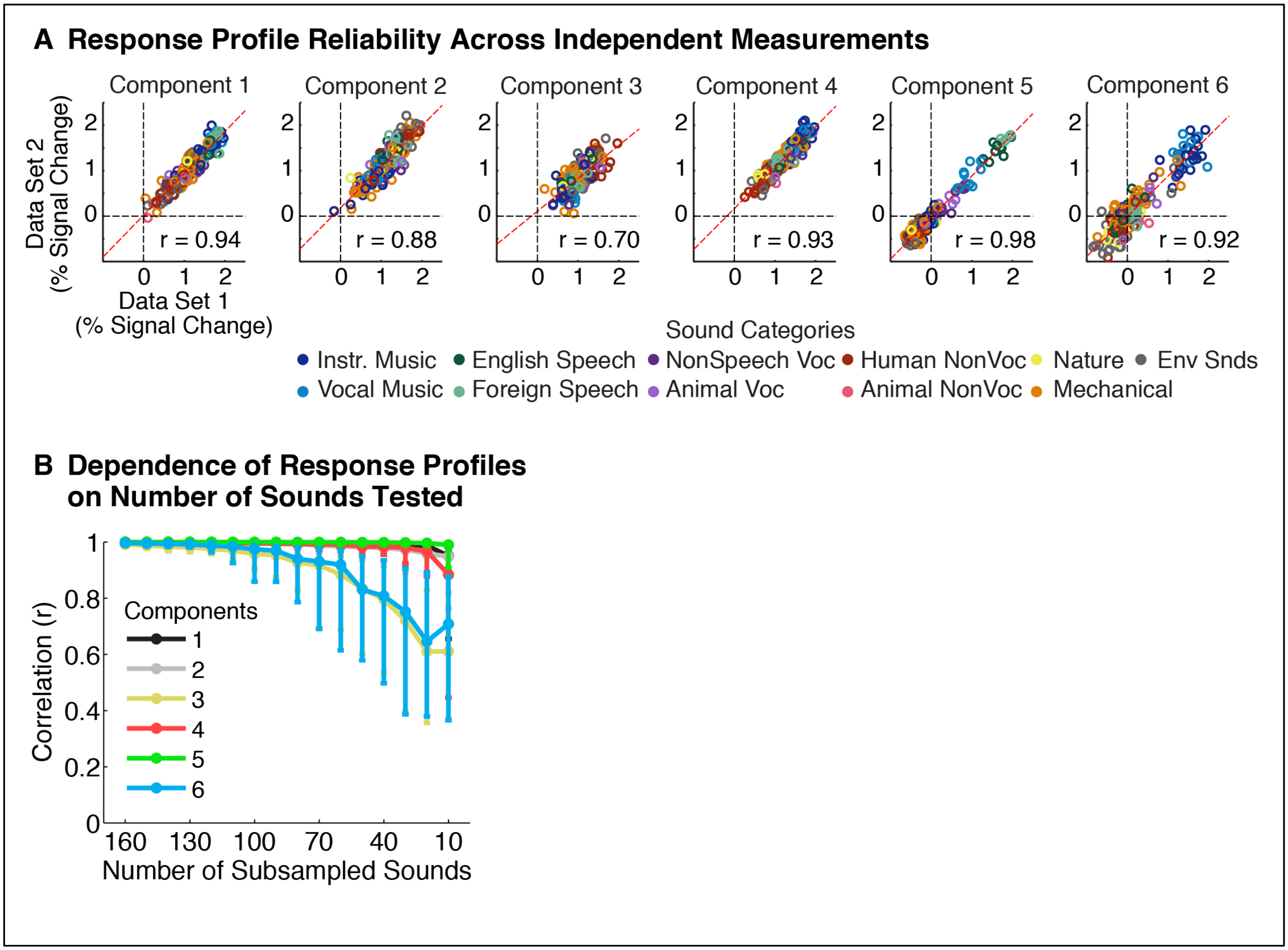


Figure S5. Component Response Profile Reliability

Related to Main Figure 2D

(A) Components were inferred using a subset of the data (scans 1 and 2), and their response profiles were re-estimated using the left-out data (scan 3) (see Supplemental Methods). Each circle plots the response of one component to a single sound, measured in each of the two data sets. The circles are colored based on the category of each sound. The test-retest correlation for each component is indicated.

(B) Components were inferred using a smaller sound set, randomly selected from the full 165-sound set. The components discovered from the reduced sound set were matched and correlated with those discovered using the full sound set. The figure plots the median and standard error of this correlation across all reduced sets of a given size.

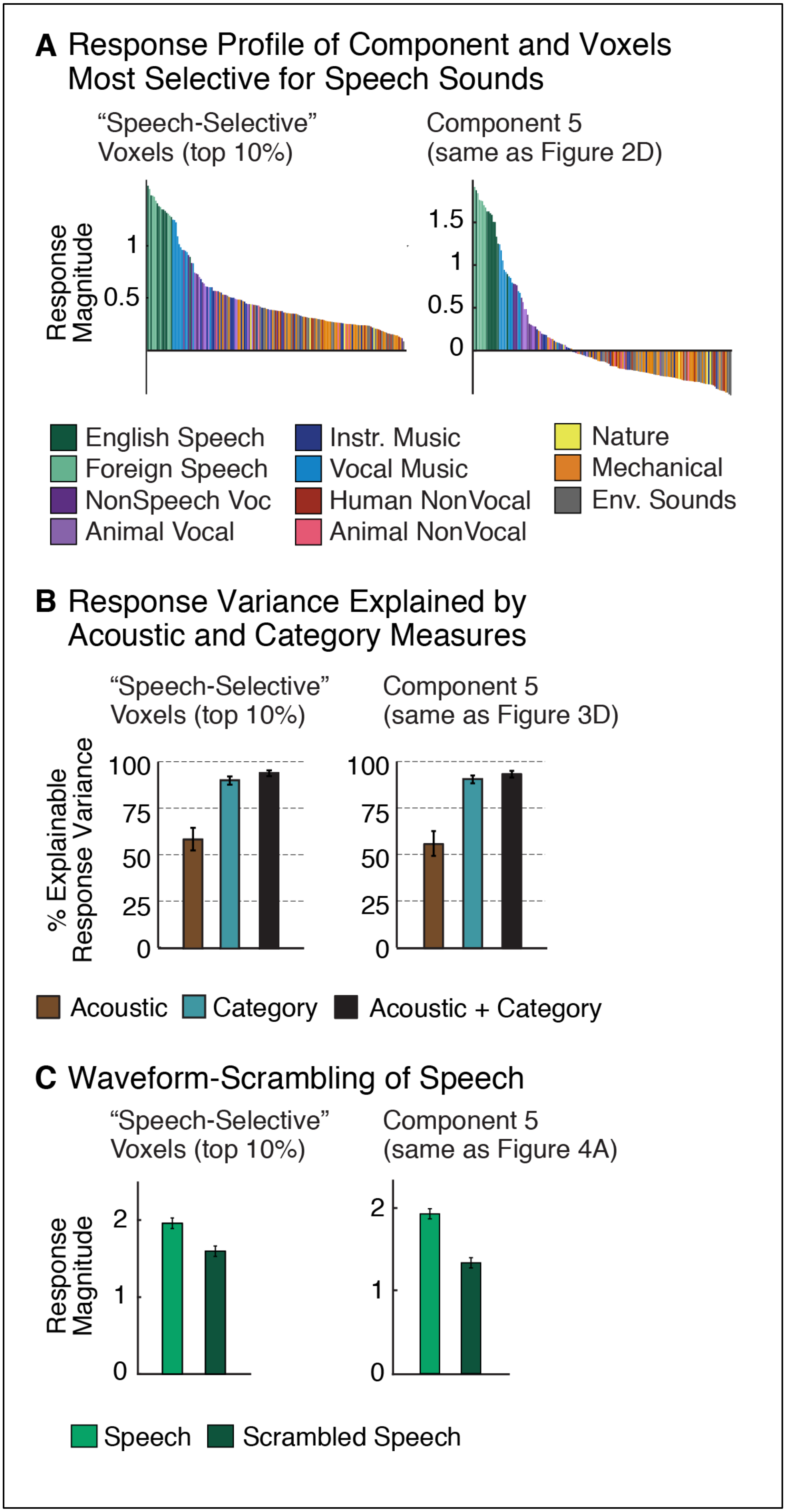


Figure S6. Analyses of Speech-Selectivity in Raw Voxels

Related to Main Figure 5.

(A) Left panel plots the average response profile of voxels with the most significant response preference for speech sounds. The response profile of Component 5, which responded selectively to speech sounds, is re-plotted for comparison (right panel).

(B) The amount of response variance explainable by acoustic features, category labels, and the combination of both acoustic and category measures for speech-selective voxels and Component 5. Both the speech-selective voxels and the Component showed robust selectivity for categories that could not be explained by acoustic features (in contrast with the pattern observed for music-selective voxels, see Figure 5B). Error bars plot standard errors across the sound set, estimated via bootstrap.

(C) The effect of audio scrambling on the response of the speech-selective voxels and Component 5. Effects of scrambling were stronger in the Component, but remained robust in speech-selective voxels. Error bars plot one standard error of the mean across subjects.

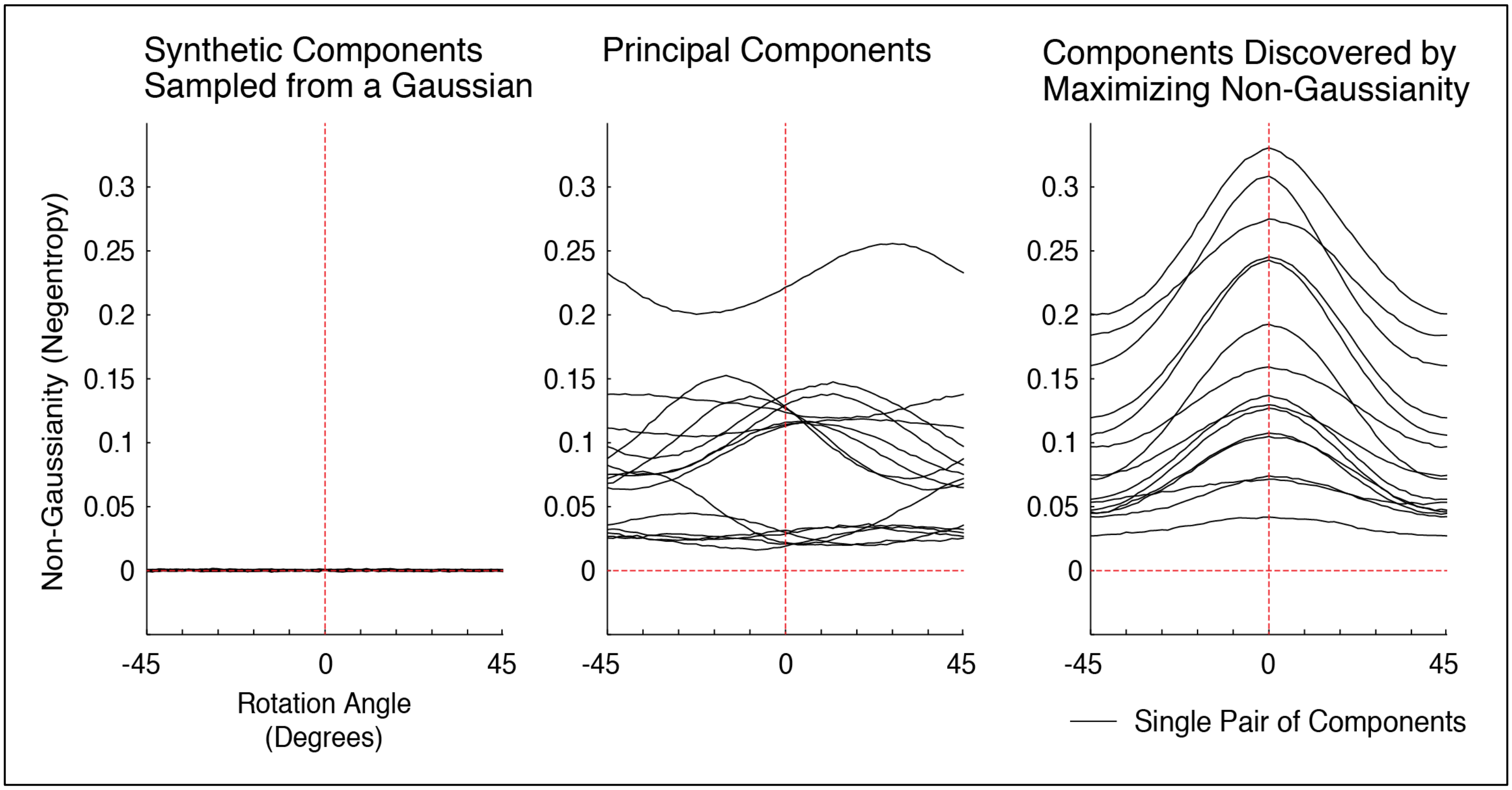


Figure S7. Testing Assumptions of Non-Gaussianity

Related to Main Figure 7

The algorithm used to discover components iteratively “rotated” pairs of principal components to maximize a measure of non-Gaussianity (“negentropy”). This approach is ineffective if the weights for the “true” latent components are Gaussian-distributed, because the Gaussian distribution is rotationally symmetric. The left panel illustrates this fact by plotting a measure of negentropy as a function of rotation for pairs of principal components measured from synthetic Gaussian data. In contrast, the principal components measured from the voxels were not rotationally symmetric (middle panel), and we could thus increase their negentropy via rotation. By iterating this process, our algorithm was able to discover a clear optimum, such that no additional rotation could increase the negentropy of the weights (right panel).

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Data Acquisition and Preprocessing

Data were collected on a 3T Siemens Trio scanner with a 32-channel head coil (at the Athinoula A. Martinos Imaging Center of the McGovern Institute for Brain Research at MIT). The functional volumes were designed to provide good spatial resolution in auditory cortex. Each functional volume (i.e. a single 3D image) 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).

Functional volumes were preprocessed using FSL software and custom MATLAB scripts. Volumes were motion-corrected, slice-time-corrected, skull-stripped, linearly detrended, and aligned to the anatomical volumes (using FLIRT and BBRegister; Greve and Fischl, 2009; Jenkinson and Smith, 2001). Volume data were then resampled to the reconstructed cortical surface computed by FreeSurfer (Dale et al., 1999), and smoothed using a 3mm FWHM kernel to improve SNR.

Measurement of Tonotopy

We measured tonotopy using responses to pure tones from one of six frequency ranges (center frequencies: 200, 400, 800, 1600, 3200, and 6400 Hz; Humphries et al., 2010; Norman-Haignere et al., 2013). 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). These best-frequency maps were averaged across subjects to form group maps. Voxels in which fewer than three subjects had frequency-modulated voxels were excluded from the group map.

Additional Details of the Non-Parametric Decomposition Algorithm

Assessing Convergence

The non-parametric algorithm is guaranteed to reach a local optimum, since it continues until no “rotation” can further improve the objective. To ensure the optimization procedure found the global optimum, we applied the algorithm 1000 times with different random initializations (random rotations of the principal component weight matrix, $\mathbf{V}_N$). We then correlated the response profiles of the best solution (highest negentropy) with the response profiles from all other initializations (after matching the response profiles via the ‘Hungarian’ algorithm; Kuhn, 1955). For the 500 solutions with highest negentropy, this correlation was very high (average $r > 0.99$), indicating that the best solution was likely a global optimum.

De-Meaning

As is standard in ICA algorithms (Hyvarinen, 1999), the rows of the data matrix were demeaned prior to applying the non-parametric algorithm: for each sound, the mean response across voxels was subtracted from the response of each voxel. This demeaning operation causes the rows of the inferred voxel weight matrix to also be zero mean, but does not change the response profile matrix. As a result, the voxel weights needed to explain the original non-demeaned data matrix can be recovered by applying the pseudoinverse of the response matrix:

$$\mathbf{W} = (\mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}\mathbf{R}^{\mathrm{T}}\mathbf{D} \tag{11}$$

where **R** is the inferred response profile matrix, **D** is the non-demeaned data matrix, and **W** is the component weight matrix.

In practice, we found it useful to demean voxels from each subject separately. Without this step, the algorithm discovered additional components that just reflected the difference or "offset" between the average response of voxels from a single subject and the average voxel response across all subjects. These "offset vectors" were generally not reliable across scan sessions, and were plausibly driven by correlated sources of noise across voxels (e.g. due to motion).

Determining the Sign of the Components

The "sign" of the response profiles and weights is not uniquely determined by the algorithm, since they can be flipped without changing the solution:

$$\mathbf{RW} = (-\mathbf{R})(-\mathbf{W}) \tag{12}$$

In practice, each inferred component could be oriented such that its average response and voxel weight were both positive. We used this convention in all of the Figures.

Determining the Number of Components

Voxel decomposition can in principle recover as many components as generated the data, but in practice is limited by the SNR of fMRI measurements. To determine the number of components to analyze, we measured (1) the amount of replicable variance accounted for by the components (Figure 1C) and (2) the accuracy of the components in predicting voxel responses from a left-out subject, not used to identify components (Figure S1). The first measure estimates the fraction of voxel response variation the components would explain if fMRI responses were perfectly reliable. The second measure, by contrast, is sensitive to the relative contribution of replicable vs. non-replicable sources in driving each component, since only components driven by replicable variance should improve prediction accuracy. We sought to find a set of N components that explained a large fraction of the explainable variance (measure 1) while maintaining good prediction accuracy (measure 2).

In the absence of noise, the amount of replicable variance (measure 1) can be computed by correlating the response of each voxel with its response projected onto the components. In the presence of noise, this correlation needs to be corrected by the reliability of the voxel and component-projected responses measured in independent scans. We did this as follows. First, we projected the response of each voxel, measured in two different scans ($\mathbf{v}^{\mathrm{scan1}}$ and $\mathbf{v}^{\mathrm{scan2}}$), onto component response profiles inferred using data from all other subjects (**R**):

$$\mathbf{v}^{\mathrm{scan1-proj}} = \mathbf{R}(\mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}\mathbf{R}^{\mathrm{T}}\mathbf{v}^{\mathrm{scan1}} \tag{13}$$

$$\mathbf{v}^{\text{scan2}-\text{proj}} = \mathbf{R}(\mathbf{R}^{\text{T}}\mathbf{R})^{-1}\mathbf{R}^{\text{T}}\mathbf{v}^{\text{scan2}} \tag{14}$$

We then correlated voxel responses from one scan with the component-projected responses from the other scan, and Z-averaged the two correlation values:

$$\rho_N^{(1)} = \text{Corr}(\mathbf{v}_N^{\text{scan1}-\text{proj}}, \mathbf{v}^{\text{scan2}}) \tag{15}$$

$$\rho_N^{(2)} = \text{Corr}(\mathbf{v}_N^{\text{scan2}-\text{proj}}, \mathbf{v}^{\text{scan1}}) \tag{16}$$

$$\begin{aligned} \rho_N &= Z(\rho_N^{(1)}, \rho_N^{(2)}) \\ &= \tanh\left[\frac{1}{2}\sum_{i=1}^{2} \tanh^{-1}\rho_N^{(i)}\right] \end{aligned} \tag{17}$$

Z-averaging reduces a small bias caused by directly averaging correlation coefficients (Silver and Dunlap, 1987). We noise-corrected this correlation measure by the reliability of the variables used to compute it (measure 1):

$$\rho_N^{\text{norm}} = \frac{\rho_N}{\sqrt{r_N^{(1)} r_N^{(2)}}} \tag{18}$$

$$r_N^{(1)} = \text{Corr}(\mathbf{v}_N^{\text{scan1}}, \mathbf{v}_N^{\text{scan2}}) \tag{19}$$

$$r_N^{(2)} = \text{Corr}(\mathbf{v}_N^{\text{scan1}-\text{proj}}, \mathbf{v}_N^{\text{scan2}-\text{proj}}) \tag{20}$$

Figure 1C plots the median of this correlation measure (equation 18) across voxels, squared to provide an estimate of explained variance.

Measure 2 is given by equation 17: the correlation between voxel responses and component-projected responses measured in different scans, not corrected for noise (Figure S1). Because the measure is not corrected, adding components does not monotonically increase prediction accuracy because higher-order components are eventually driven more by noise than replicable signal.

Additional Details of Parametric Decomposition Model

Model Specification

The model assigned a probability to each voxel's response, given a set of component response profiles and a Gamma-distributed prior on component voxel weights. In the equations below:

- Lower-case, bolded symbols denote vectors
- Upper-case bolded symbols denote matrices
- Unbolded symbols denote scalars

The Gamma prior on weights took the following form:

$$p(\mathbf{w}_i|\boldsymbol{\beta}) = \prod_{\mathrm{c}=1}^{N} \mathrm{Gamma}(w_{c,i}|\beta_c) = \prod_{\mathrm{c}=1}^{N} \frac{\beta_c^{\beta_c}}{\Gamma(\beta_c)} w_{c,i}^{\beta_c-1} e^{-\beta_c w_{c,i}} \tag{21}$$

where N is the number of components, w$_{c,i}$ is the weight for component *c* in voxel *i*, and β_{c} is the shape parameter of the Gamma distribution for component *c*.

Given a set of response profiles and weights, we modeled the likelihood of observing each voxel's response as a diagonal Gaussian, with mean centered on the weighted sum of the response profiles:

$$p(\mathbf{v}_{i,j}|\mathbf{R}, \mathbf{w}_i, \sigma_i^2) = \mathrm{Normal}(\mathbf{v}_{i,j}|\boldsymbol{\mu}_i = \mathbf{R}\mathbf{w}_i, \sigma_i^2\mathbf{I}) \tag{22}$$

where $\mathbf{v}_{i,j}$ denotes the response vector of voxel *i* measured in scan *j* and **R** is the response profile matrix [165 x N].

The variance ($\sigma_i{}^2$) for each voxel was set to its empirical variance across scans:

$$\sigma_i^2 = \frac{1}{2S}\binom{M_i}{2}^{-1} \sum_{k=1}^{M_i} \sum_{l=k+1}^{M_i} (\mathbf{v}_{i,k} - \mathbf{v}_{i,l})^T(\mathbf{v}_{i,k} - \mathbf{v}_{i,l}) \tag{23}$$

where M_i indicates the number of measurements/scans for voxel *i* (2 or 3 depending on the subject), and *S* the total number of stimuli (165).

The log-likelihood of the data integrating across all possible weights is then given by:

$$\log p(\{\mathbf{v}_{i,j}\}|\mathbf{R}, \boldsymbol{\beta}) = \sum_{i=1}^{V} \log \int p(\mathbf{w}_i|\boldsymbol{\beta}) \prod_{j=1}^{M} p(\mathbf{v}_{i,j}|\mathbf{R}, \mathbf{w}_i, \sigma_i^2) \mathrm{d}\mathbf{w}_i \tag{24}$$

where {$\mathbf{v}_{i,j}$} indicates the set of all voxel responses across all subjects and scans, and *V* is the total number of voxels. The response matrix (**R**) and shape parameters (β) were chosen to maximize this log-likelihood via the optimization procedure described below.

Model Optimization

The data log-likelihood (equation 24) cannot be computed in closed form because the prior (equation 21) and likelihood distributions (equation 22) are not conjugate (Murphy, 2012). We therefore optimized the model using a stochastic variant of the standard expectation-maximization (EM) algorithm (Dempster et al., 1977; Wei and Tanner, 1990). The EM algorithm takes advantage of the fact that the logarithm of the joint distribution over the data and latent parameters (equation 25 below) - in our case the voxel weights - is often easier to compute than the data log-likelihood (equation 24), which requires integrating across the latent parameters.

$$\log p(\{\mathbf{v}_{i,j}\}, \{\mathbf{w}_i\}|\mathbf{R}, \boldsymbol{\beta}) = \sum_{i=1}^{V} \log p(\mathbf{w}_i|\boldsymbol{\beta}) + \sum_{i=1}^{V}\sum_{j=1}^{M} \log p(\mathbf{v}_{i,j}|\mathbf{R}, \mathbf{w}_i, \sigma_i^2) \tag{25}$$

EM computes an expectation of the log-joint probability with respect to the posterior distribution over the latent parameters (voxel weights), and this expectation is iteratively maximized with respect to the hyper-parameters - in this case, the response profiles (**R**) and shape parameters (β):

$$\mathbf{R}_{new}, \boldsymbol{\beta}_{new} = \tag{26}$$

$$\underset{\mathbf{R},\boldsymbol{\beta}}{\arg\max}\, \mathbf{E}\big[\log p(\{\mathbf{v}_{i,j}\}\{\mathbf{w}_i\}|\mathbf{R},\boldsymbol{\beta}) \;\Big|\; p(\{\mathbf{w}_i\}|\{\mathbf{v}_{i,j}\},\mathbf{R}_{fixed},\boldsymbol{\beta}_{fixed})\big]$$

The posterior distribution over the voxel weights is computed with respect to a fixed set of hyper-parameters (**R***fixed*, β *fixed*), and the expectation is then maximized with respect to the hyper-parameters of the joint distribution (**R**, β). The posterior over voxel weights is then re-computed using the new hyper-parameters (**R***fixed* = **R***new*, β *fixed* = β *new*), and the process is repeated.

The expectation in equation 26 can be expanded using equations 21 and 22. It includes many terms, but only three quantities depend on the posterior weight distribution over which the expectation is computed: the first two moments of the voxel weights (E[w*c,i*] and E[w*l,i* w*m,i*]) and the expectation of the log-transformed voxel weights (E[log w*c,i*]):

$$\mathbf{E}\big[\ \log p(\{\mathbf{v}_{i,j}\}\{\mathbf{w}_i\}|\mathbf{R},\boldsymbol{\beta})\ \ |\ \ p(\{\mathbf{w}_i\}|\{\mathbf{v}_{i,j}\},\mathbf{R}_{fixed},\boldsymbol{\beta}_{fixed})\ \big] = \tag{27}$$

$$\sum_{i=1}^{V}\sum_{c=1}^{N}\beta_c\log\beta_c - \log\Gamma(\beta_c) + (\beta_c - 1)\mathbf{E}[\log w_{c,i}] - \beta_c\mathbf{E}[w_{c,i}]$$

$$-\sum_{i=1}^{V}\sum_{j=1}^{M}\frac{1}{2\sigma_i^2}\Big(\mathbf{v}_{i,j}^T\mathbf{v}_{i,j} - 2\mathbf{v}_{i,j}^T\mathbf{R}\mathbf{E}[\mathbf{w}_i] + \sum_{l,m=1}^{C}\mathbf{E}[w_{l,i}w_{m,i}]\sum_{k=1}^{S}R_{k,l}R_{k,m}\Big)$$

$$-\sum_{i=1}^{V}\sum_{j=1}^{M}\Big(\frac{S}{2}\log 2\pi + \frac{S}{2}\log\sigma_i^2\Big)$$

These three statistics also cannot be computed in closed form (because like the data log-likelihood, they require an intractable integral over voxel weights). We estimated them using “importance-weighted” samples from an approximating Gaussian distribution (Bishop and others, 2006). This was accomplished in five steps. First we log-transformed the voxel weights, so that the distribution being sampled from had support everywhere (unlike the un-transformed weights which were non-negative due to the Gamma prior):

$$\mathbf{z}_i = \log\mathbf{w}_i \tag{28}$$

$$p(\mathbf{z}_i) = e^{\mathbf{z}_i}p(\mathbf{w}_i = e^{\mathbf{z}_i}) \tag{29}$$

Second, we approximated the posterior distribution over log-weights with a Gaussian centered at the maximum of the distribution ($\mathbf{z}_i^{max}$, computed using Newton’s method) and covariance matrix set to:

$$\mathbf{\Sigma}_i = (-\mathbf{H}_i)^{-1} \tag{30}$$

$$\mathbf{H}_i = \frac{\partial \log p(\mathbf{z}_i = \mathbf{z}_i^{max} | \{\mathbf{v}_{i,j}\}, \mathbf{R}_{fixed}, \boldsymbol{\beta}_{fixed})}{\partial z_{m,i} \partial z_{n,i}} \tag{31}$$

where $\mathbf{H}_i$ is the Hessian of the log-posterior over log-weights at the maximum (i.e. the "Laplace approximation") (Murphy, 2012). Third, we sampled a set of N values from the approximating Gaussian ($\mathbf{z}_i^{(n)} \sim G_i$) and exponentiated the samples ($\mathbf{w}_i^{(n)} = e^{\mathbf{z}i(n)}$) to undo the effect of the log-transformation. Fourth, for each sample, we computed an "importance weight" ($q(\mathbf{z}_i^{(n)})$), proportional to the ratio of the true posterior and approximating Gaussian:

$$\begin{aligned} q(\mathbf{z}_i^{(n)}) &= \frac{p(\mathbf{z}_i, \{\mathbf{v}_{i,j}\} | \mathbf{R}_{fixed}, \boldsymbol{\beta}_{fixed})}{\mathrm{G}_i(\mathbf{z}_i^{(n)})} \\ &\propto \frac{p(\mathbf{z}_i | \{\mathbf{v}_{i,j}\}, \mathbf{R}_{fixed}, \boldsymbol{\beta}_{fixed})}{\mathrm{G}_i(\mathbf{z}_i^{(n)})} \end{aligned} \tag{32}$$

Fifth and finally, we used the sampled voxel weights ($\mathbf{w}_i^{(n)}$) and the importance weights ($q(\mathbf{z}_i^{(n)})$) to approximate the 3 required statistics:

$$\mathbf{E}[\mathbf{w}_i] \approx \frac{\sum_{n=1}^{N} q(\mathbf{z}_i^{(n)}) \mathbf{w}_i^{(n)}}{\sum_{l=1}^{N} q(\mathbf{z}_i^{(l)})} \tag{33}$$

$$\mathbf{E}[\mathbf{w}_i \mathbf{w}_i^T] \approx \frac{\sum_{n=1}^{N} q(\mathbf{z}_i^{(n)}) \mathbf{w}_i^{(n)} {\mathbf{w}_i^{(n)}}^T}{\sum_{l=1}^{N} q(\mathbf{z}_i^{(l)})} \tag{34}$$

$$\mathbf{E}[\log \mathbf{w}_i] \approx \frac{\sum_{n=1}^{N} q(\mathbf{z}_i^{(n)}) \log \mathbf{w}_i^{(n)}}{\sum_{l=1}^{N} q(\mathbf{z}_i^{(l)})} \tag{35}$$

As the number of samples (N) increases, these sums converge to the true statistics of the posterior (Wei and Tanner, 1990).

Using our estimates of these 3 statistics, we maximized the objective in equation 26 with respect to the response matrix ($\mathbf{R}$) and shape parameters (β). The maximum-likelihood solution for the response matrix was computed in closed form using weighted least squares:

$$\mathbf{R} = \mathbf{D}\hat{\mathbf{W}}\mathbf{X}^{-1} \tag{36}$$

$$\hat{\mathbf{W}}[c, i] = \frac{M_i}{\sigma_i^2} \mathbf{E}[w_{c,i}] \tag{37}$$

$$\mathbf{X} = \sum_i \frac{M_i}{\sigma_i^2} \mathbf{E}[\mathbf{w_i} {\mathbf{w_i}}^T] \tag{38}$$

The optimization with respect to the shape parameters was performed using MATLAB's implementation of BFGS (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970), a quasi-Newton method.

Subject Offsets

As in the non-parametric method, we found it was useful to subtract a subject-specific "offset" vector from the response of each voxel (see "De-Meaning" section above). We used our model to infer an optimal offset vector ($\mathbf{o}_s$), one per subject, that maximized the likelihood of the data (using weighted least-squares). The voxel responses ($\mathbf{v}_{i,j}$) in all other equations were then replaced with "offset" voxel responses:

$$\mathbf{v}_{i,j}^{offset} = \mathbf{v}_{i,j} - \mathbf{o}_{s(i)} \tag{39}$$

where s(i) denotes the subject for voxel *i*.

Assessing and Improving Global Convergence

The EM algorithm is guaranteed to converge to a local, but not a global optimum. In practice, we found that applying the EM algorithm in an iterative manner improved global convergence. First, we initialized the component response profiles with the response of randomly selected voxels projected onto the first N principal components (ensuring that the response profiles started near regions of high response variance). The initial values of the shape parameters had little effect on the optimization and were fixed (β_c = 1). Subject offset vectors were initialized to the average response difference (or 'offset') between the voxels of a single subject and the voxels of all ten subjects. Second, the algorithm was run for 10 EM iterations, using 100 samples to approximate the posterior statistics (equations 33-35). Third, two of the response profiles were randomly re-initialized (using two more randomly selected voxels), and another 10 iterations were run. Fourth, we compared likelihood estimates (described below) for the solutions found before and after re-initialization, and kept the solution with highest likelihood. We repeated steps 3-4, randomly re-initializing response profiles for all pairs of components ten times. The resulting solution was then further refined using 200 EM iterations with 1000 samples per iteration.

To evaluate convergence, this entire process was repeated 200 times. We then correlated the response profiles for the solution with highest estimated likelihood with the response profiles for all other solutions (after matching them using the Hungarian algorithm). Of the top 100 solutions with the highest likelihood, the average correlation was 0.98, indicating that the algorithm converged to a stable solution across different initializations.

Likelihood Estimates

We estimated the likelihood of the data given parameters in two steps. First, we approximated the posterior distribution over log-transformed weights with a Gaussian (as described above). Second, we used importance-weighted samples from the Gaussian to directly approximate the log-likelihood of the data:

$$\begin{aligned}\log p(\{\mathbf{v}_{i,j}\}|\mathbf{R},\boldsymbol{\beta}) &= \sum_{i=1}^{V} \log \int p(\mathbf{w}_i|\boldsymbol{\beta}) \prod_{j=1}^{M} p(\mathbf{v}_{i,j}|\mathbf{R},\mathbf{w}_i,\sigma_i^2)\mathrm{d}\mathbf{w} \\ &\approx \sum_{i=1}^{V} \log \frac{1}{N} \sum_{n}^{N} \frac{1}{\mathrm{G}_i(\mathbf{w}_i^{(n)})} p(\mathbf{w}_i^{(n)}|\boldsymbol{\beta}) \prod_{j=1}^{M} p(\mathbf{v}_{i,j}|\mathbf{R},\mathbf{w}_i^{(n)},\sigma_i^2)\end{aligned} \tag{40}$$

where G_i is the approximating Gaussian for voxel *i*, and $w_i^{(n)}$ is a sample from that Gaussian. We used 1000 samples per voxel to approximate the integral. Although stochastic, the log-likelihood estimates were highly stable across independent sets of samples.

Additional Analyses of Component Response Properties and Anatomy

Statistical Significance of Weight Maps

We computed significance for the component voxel weights via a permutation test (Figures 2B, S2, & S3). Specifically, we computed a null distribution for each component by permuting/shuffling its response profile 10,000 times and re-computing the component weights for all voxels. To avoid changing the correlation between response profiles of different components, we permuted response variation unique to that component (i.e. the residual after removing shared variance). Results were similar permuting the raw profile. We fit the null distribution for each component and voxel with a Gaussian, and calculated the likelihood of obtaining the observed component weight (based on the un-permuted profile), given a sample from this Gaussian.

Variance Explained by Acoustic Features and Category Labels

We estimated the variance explained by different sets of acoustic features by regressing them against the response profile of each component (Figures 3D&E). Each set of features (audio frequency, temporal modulation, and spectrotemporal modulation) was defined by a 165 x N matrix, with one vector per feature: six for audio frequency, nine for temporal modulation and 49 for spectrotemporal modulation (7 scales x 9 rates). Because the spectrotemporal matrix was relatively high dimensional, and its features highly correlated, we reduced its dimensionality by selecting the top 15 principal components (accounting for 95% of the total variation). For the temporal and spectrotemporal feature matrices we included the mean energy vector across frequency as an additional predictor, because variation in mean energy was driven by modulation (due to RMS normalization of stimuli in conjunction with power compression).

We regressed category judgments against the response profile of each component to measure the variance they explained. Category judgments were represented by a matrix (165 x 11) containing the proportion of subjects that assigned each category to each sound (this matrix was reliable across participants; split-half correlation of 0.98). To measure the variance explained by acoustic features and categories, we concatenated the acoustic and category feature matrices.

To avoid over-fitting, we predicted the response to each sound using regression weights estimated using all other sounds. We correlated the resulting prediction vector with the response profile of each component, normalized by the reliability of the measures (see below), and squared it to estimate variance explained. Error bars on these estimates were computed via bootstrapping: sampling with replacement across the sound set (10,000 samples), and re-computing the correlation between the acoustic feature predictions and the component response profile. Statistical significance was determined using a null distribution obtained by permuting the rows of the feature matrices and re-computing the correlation with the component profile (10,000 permutations).

Correlation Normalization to Correct for Measurement Noise

For the acoustic correlation values plotted in Figures 3B and 3C, we noise-corrected the correlation between acoustic feature vectors and component response profiles by the test-retest reliability of the profiles across scans:

$$\rho = \frac{Z(\mathrm{Corr}(\mathbf{s}, \mathbf{r}_1), \mathrm{Corr}(\mathbf{s}, \mathbf{r}_2))}{\sqrt{\mathrm{Corr}(\mathbf{r}_1, \mathbf{r}_2)}} \qquad (41)$$

$$Z(\rho_1, \rho_2) = \tanh\Big[\frac{1}{2}\sum_{i=1}^{2} \tanh^{-1}\rho_i\Big] \qquad (42)$$

where **r₁** and **r₂** indicate estimates of each component's response vector measured in two different scans, and **s** is a vector of stimulus features. Z-averaging was again used to reduce a small bias caused by directly averaging correlation coefficients (Silver and Dunlap, 1987). $\mathbf{r}_1$ and $\mathbf{r}_2$ were computed by projecting the voxel responses from the first scan, $\mathbf{D}_1$, onto the component response profile matrix, **R**, and then using the resulting voxel weights, $\mathbf{W}_1$, to re-estimate the response profiles from voxel responses measured in scans 2 and 3 ($\mathbf{D}_2$ and $\mathbf{D}_3$):

$$\mathbf{W}_1 = (\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T\mathbf{D}_1 \qquad (43)$$

$$\mathbf{R}_1 = \mathbf{D}_2\mathbf{W}_1^T(\mathbf{W}_1\mathbf{W}_1^T)^{-1} \qquad (44)$$

$$\mathbf{R}_2 = \mathbf{D}_3\mathbf{W}_1^T(\mathbf{W}_1\mathbf{W}_1^T)^{-1} \qquad (45)$$

Note that these estimates are not fully independent, since the response profile matrix **R** was computed from all of the data. However, the effect of any non-independence will be to make the normalized correlations smaller (because the test-retest correlation will be higher), and our measures thus provide a conservative estimate of the correlation between stimulus predictors and component response profiles. We adopted this method because the component analysis is more reliable with three scans worth of data compared with a single scan, producing a more robust **R** matrix.

For the regression analyses used to estimate explained variance (Figures 3D&E, 5B and S6B), we corrected for the reliability of both the component response profiles and the prediction vectors (necessary because the predictions depend on the response profiles, and thus are subject to effects of fMRI noise):

$$\rho = \frac{\tanh\left[\frac{1}{4}\sum_{i,j=1}^{2}\tanh^{-1}[\mathrm{Corr}(\mathbf{p}_{\mathrm{i}},\mathbf{r}_{\mathrm{j}})]\right]}{\sqrt{\mathrm{Corr}(\mathbf{r}_1,\mathbf{r}_2)\mathrm{Corr}(\mathbf{p}_1,\mathbf{p}_2)}} \tag{46}$$

In these equations, $\mathbf{p_1}$ and $\mathbf{p_2}$ indicate prediction vectors estimated by regressing feature matrices against the two response profiles, $\mathbf{r_1}$ and $\mathbf{r_2}$. We used the square of this normalized correlation as a measure of explained variance.

Component Response Profile Reliability Across Scans

We tested the reliability of each response profile by inferring components using data from the first two scans of each subject, and then re-estimating their response profiles using data from a third scan (Figure S5A). The response profiles were re-estimated by multiplying the voxel responses measured in scan 3 ($\mathbf{D}_3$) by the pseudoinverse of the component weights from scans 1 and 2 ($\mathbf{W}_{12}$):

$$\mathbf{D}_3\mathbf{W}_{12}^T(\mathbf{W}_{12}\mathbf{W}_{12}^T)^{-1} \tag{47}$$

Sensitivity of Component Response Profiles to the Sounds Tested

We investigated the sensitivity of the discovered response profiles to the specific sounds tested by re-running the analysis on subsets of sounds (Figure S5B). Each subset contained M unique, randomly chosen sounds (M varied from 10 to 160 sounds, in steps of 10). For each subset, we used the non-parametric algorithm to infer six components that best modeled the reduced data matrix (formed from the reduced sound set). We then compared the response profiles inferred from the reduced sound set to those discovered using all 165 sounds, by matching (via the Hungarian algorithm) and correlating their response profiles (using just the sounds from the reduced set). This process was repeated 200 times per set size (with different subsampled sound sets). Figure S5B plots the median correlation value for each component across the 200 samples, as a function of the set size.

Testing Assumptions of Non-Gaussianity

We tested whether the inferred voxel weights were more skewed (s_c) and kurtotic (k_c) than would be expected from a Gaussian distribution (Figure 7A):

$$s_c = \frac{1}{N\sigma_c^3}\sum_{i=1}^{N}(w_{i,c} - \mu_c)^3 \tag{48}$$

$$k_c = \frac{1}{N\sigma_c^4}\sum_{i=1}^{N}(w_{i,c} - \mu_c)^4 \tag{49}$$

where

$$\mu_c = \frac{1}{N}\sum_{i=1}^{N} w_{i,c} \tag{50}$$

$$\sigma_c = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(w_{i,c} - \mu_c)^2} \tag{51}$$

$w_{i,c}$ indicates the weight for component *c* in voxel *i*, and N is the total number of voxels. Voxel weights were also fit with two parametric distributions: a Gaussian distribution and non-Gaussian 'Johnson' distribution (Figure 7B), obtained by transforming a Gaussian-distributed random variable (*g*) via the hyperbolic sine function (Johnson, 1949):

$$j \sim \sigma \sinh(\frac{1}{b}(g - a)) + \mu \tag{52}$$

We also directly compared the non-Gaussianity (via negentropy) of principled components with the non-Gaussianity of components inferred by our non-parametric algorithm, which rotated principle components to maximize non-Gaussianity (Figure S7). If the underlying components are Gaussian, then the voxel weights for each principal component would also be Gaussian, and would remain so following any rotation (because whitened Gaussians are rotationally symmetric) (Murphy, 2012).

For all of the analyses of non-Gaussianity, we used independent data to infer components (scans 1 and 2) and measure their statistical properties (scan 3). Bootstrapping across subjects was used to assess significance.

SUPPLEMENTAL REFERENCES

Bishop, C.M., others, 2006. Pattern recognition and machine learning. springer New York.

Broyden, C.G., 1970. The convergence of a class of double-rank minimization algorithms 1. general considerations. IMA J. Appl. Math. 6, 76–90.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis: I. segmentation and surface reconstruction. NeuroImage 9, 179–194. doi:10.1006/nimg.1998.0395

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Methodol. 1–38.

Fletcher, R., 1970. A new approach to variable metric algorithms. Comput. J. 13, 317–322.

Goldfarb, D., 1970. A family of variable-metric methods derived by variational means. Math. Comput. 24, 23–26.

Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48, 63.

Humphries, C., Liebenthal, E., Binder, J.R., 2010. Tonotopic organization of human auditory cortex. NeuroImage 50, 1202–1211. doi:10.1016/j.neuroimage.2010.01.046

Hyvarinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. Neural Netw. IEEE Trans. On 10, 626–634.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5, 143–156.

Johnson, N.L., 1949. Systems of frequency curves generated by methods of translation. Biometrika 149–176.

Kuhn, H.W., 1955. The Hungarian method for the assignment problem. Nav. Res. Logist. Q. 2, 83–97.

Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.

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

Shanno, D.F., 1970. Conditioning of quasi-Newton methods for function minimization. Math. Comput. 24, 647–656.

Silver, N.C., Dunlap, W.P., 1987. Averaging correlation coefficients: should Fisher's z transformation be used? J. Appl. Psychol. 72, 146.

Wei, G.C., Tanner, M.A., 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J. Am. Stat. Assoc. 85, 699–704.