SUPPLEMENTAL FIGURES

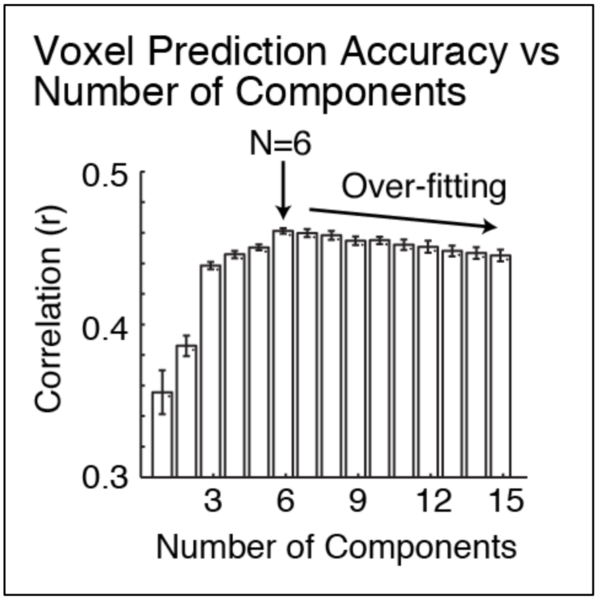
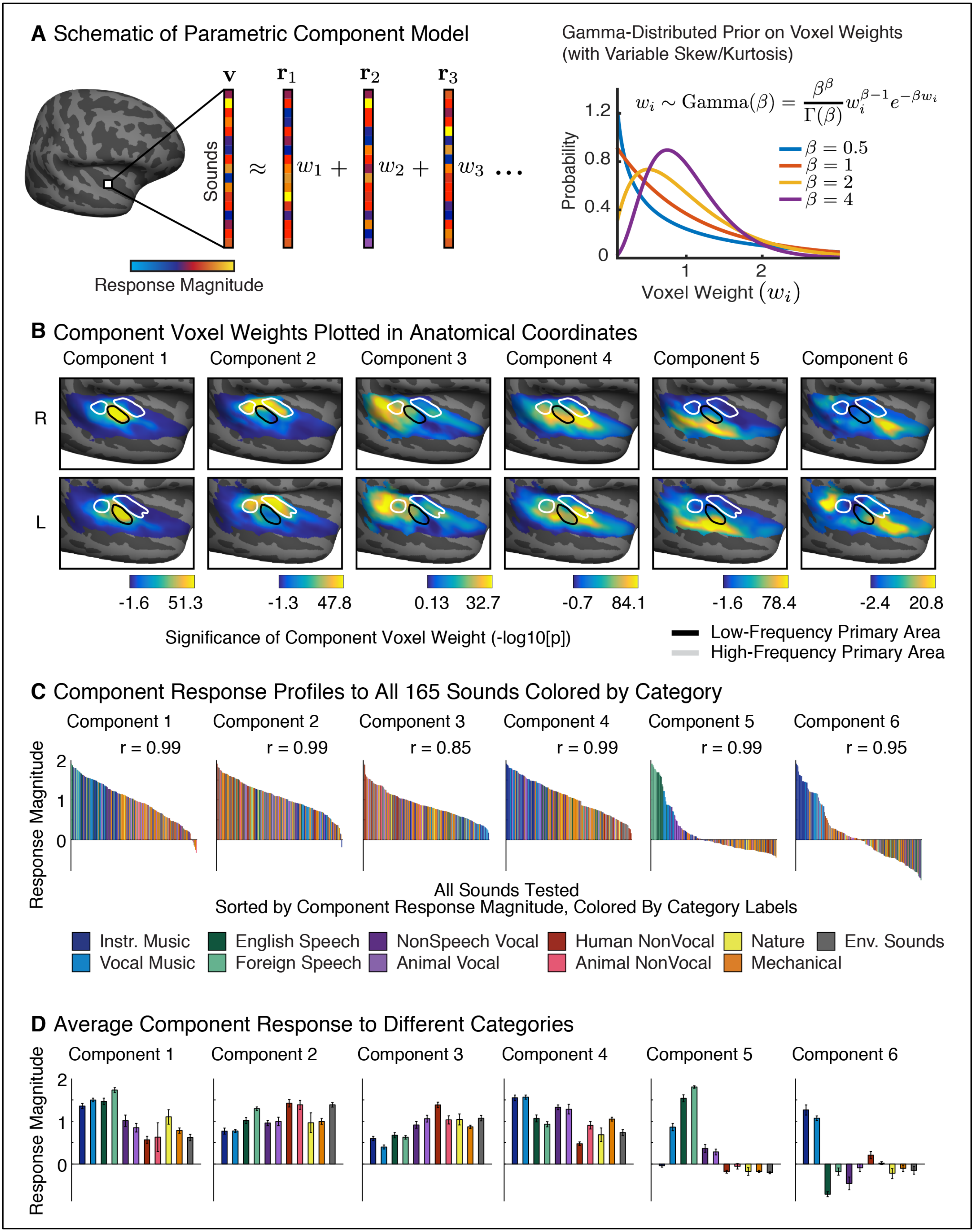


Figure S1. Voxel Prediction Accuracy vs. Number of Components

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

Gamma-Distributed Prior on Voxel Weights (with Variable Skew/Kurtosis)

B Component Voxel Weights Plotted in Anatomical Coordinates

C Component Response Profiles to All 165 Sounds Colored by Category

D Average Component Response to Different Categories

Figure S2. Parametric Component Model

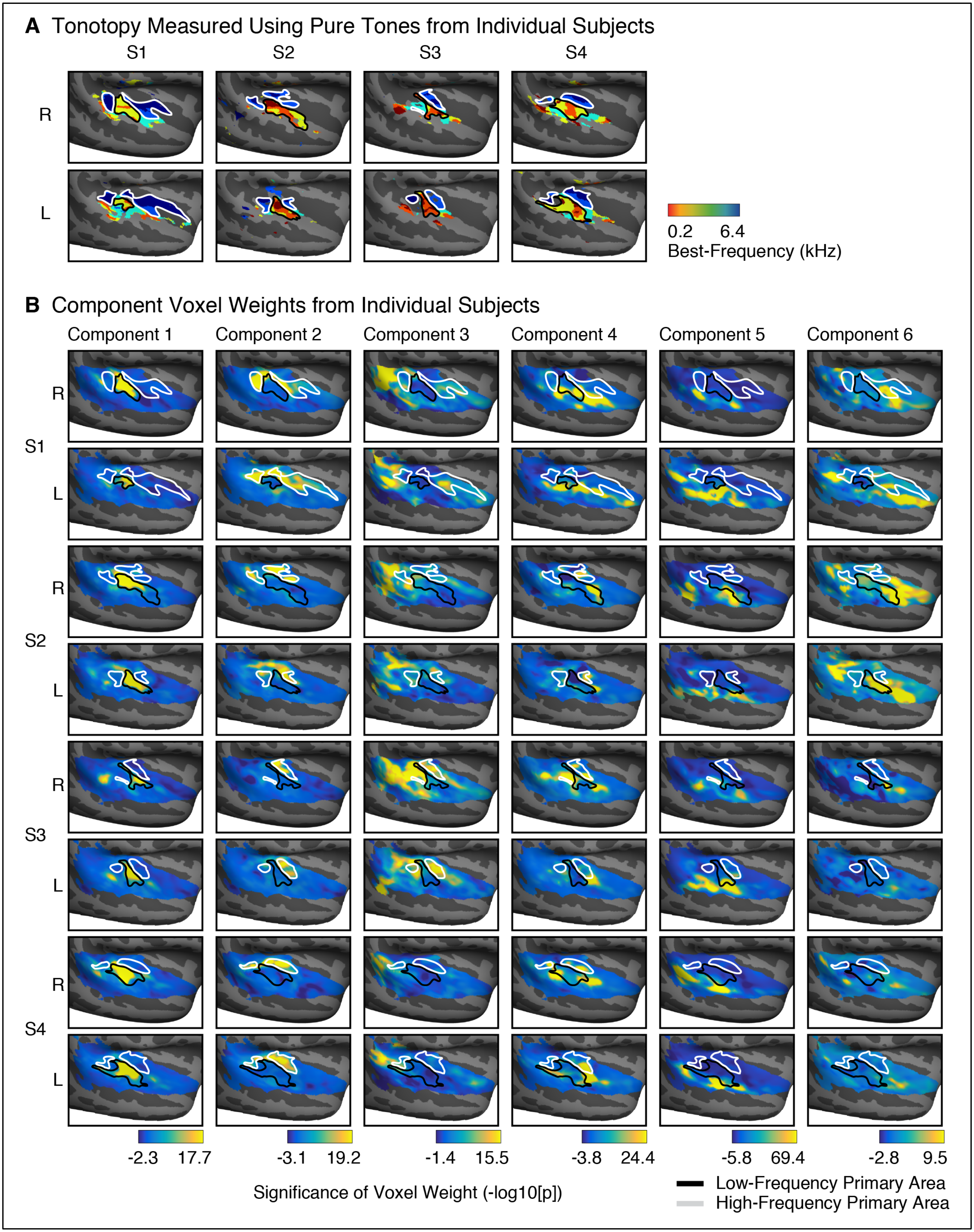
Related to Main Figure 2

(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.

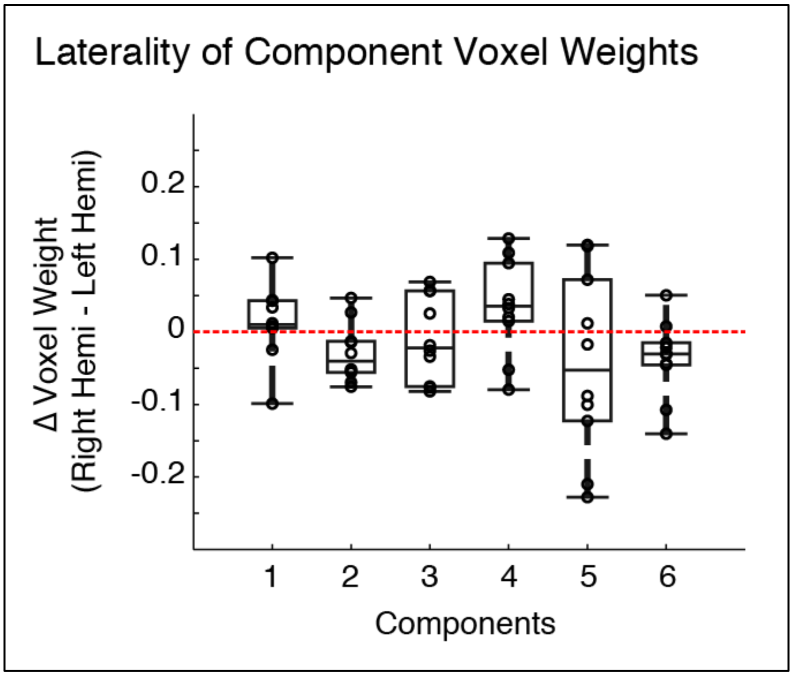
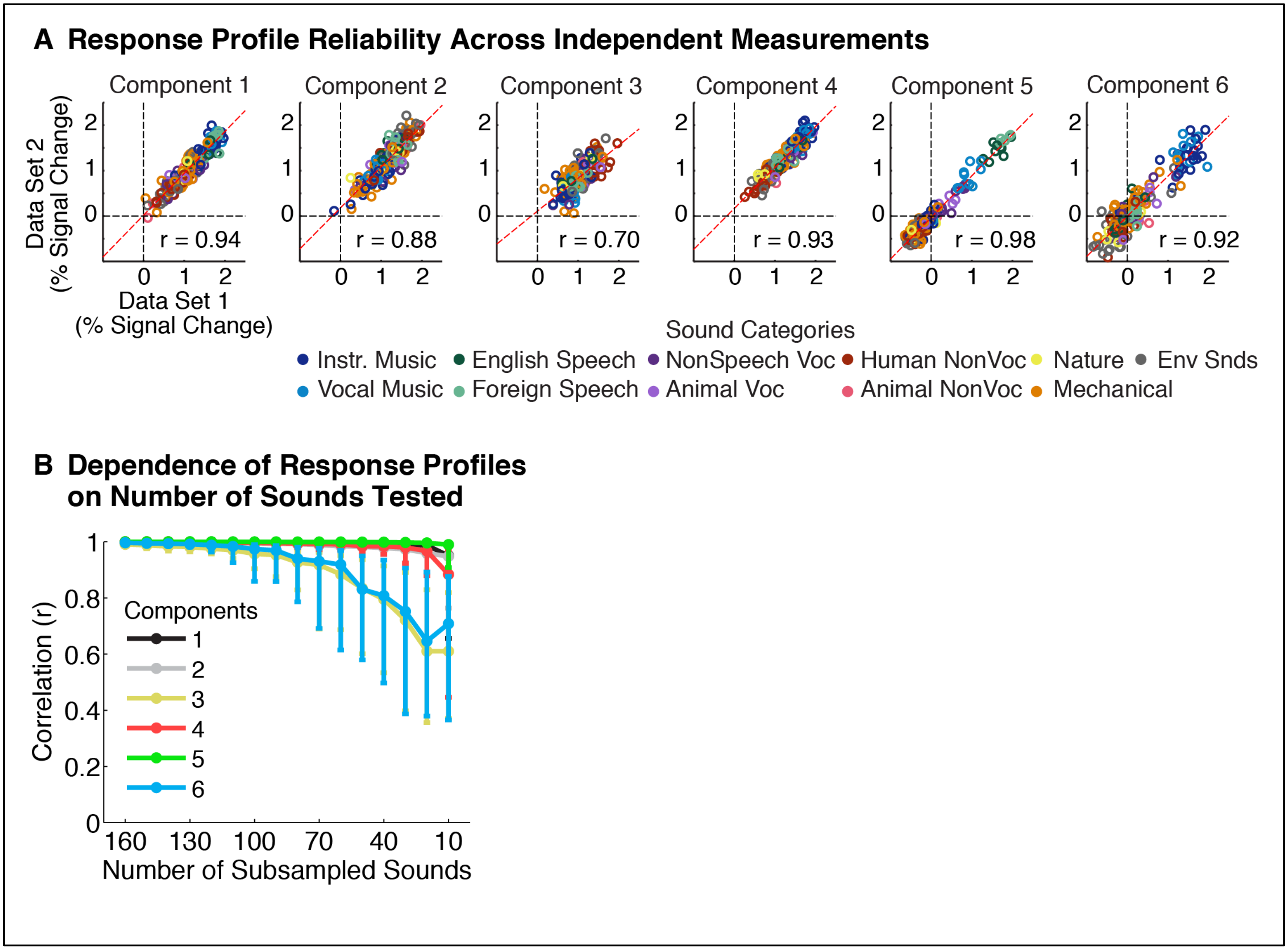


Figure S4. Laterality of Component Voxel Weights

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.



A Response Profile Reliability Across Independent Measurements

B Dependence of Response Profiles on Number of Sounds Tested

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.

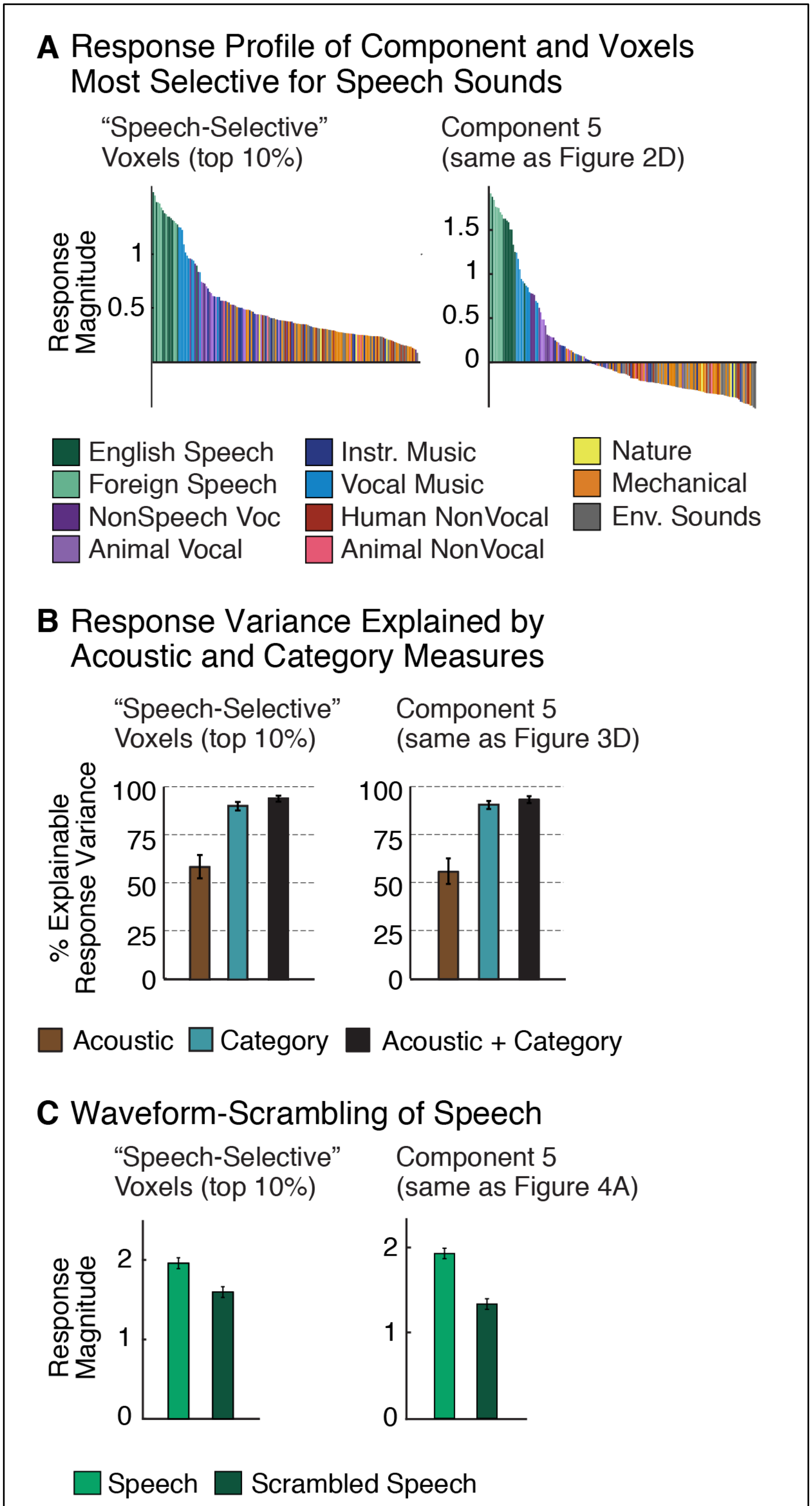


Figure S6. Analyses of Speech-Selectivity in Raw Voxels

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.

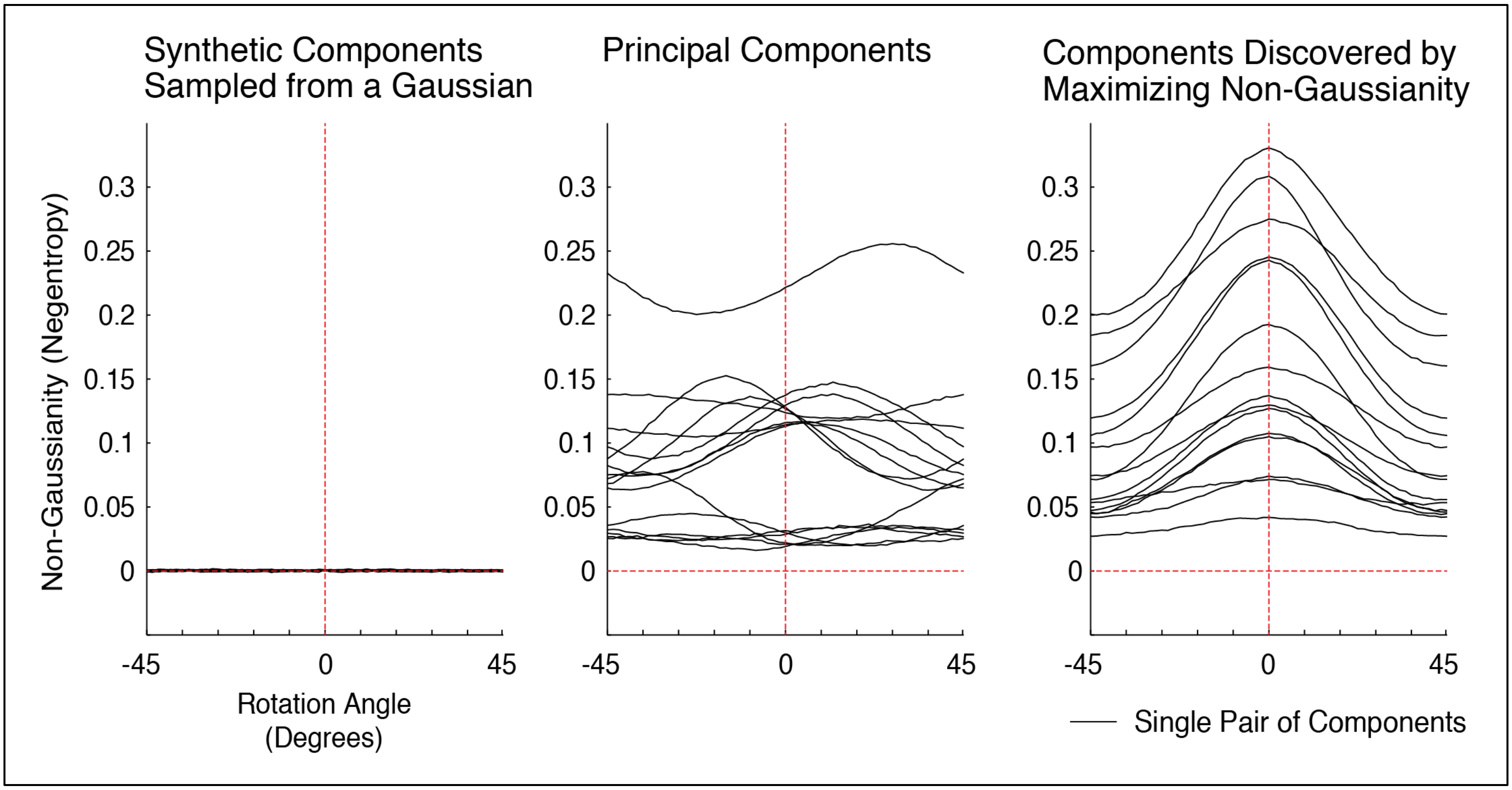


Figure S7. Testing Assumptions of Non-Gaussianity

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 ($\mathbf{v}^{\text{scan1}}$ and $\mathbf{v}^{\text{scan2}}$), 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):

$$\mathbf{R}_{new}, \boldsymbol{\beta}_{new} = \\ \arg\max_{\mathbf{R},\boldsymbol{\beta}} \mathbf{E}\big[\log p(\{\mathbf{v}_{i,j}\}\{\mathbf{w}_i\}|\mathbf{R},\boldsymbol{\beta}) \;\big|\; p(\{\mathbf{w}_i\}|\{\mathbf{v}_{i,j}\},\mathbf{R}_{fixed},\boldsymbol{\beta}_{fixed})\big] \tag{26}$$

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$_{c,i}$]):

$$\begin{aligned}
&\mathbf{E}\big[\ \log p(\{\mathbf{v}_{i,j}\}\{\mathbf{w}_i\}|\mathbf{R},\boldsymbol{\beta})\ \mid\ p(\{\mathbf{w}_i\}|\{\mathbf{v}_{i,j}\},\mathbf{R}_{fixed},\boldsymbol{\beta}_{fixed})\ \big] = \\
&\sum_{i=1}^{V}\sum_{c=1}^{N}\beta_c\log\beta_c - \log\Gamma(\beta_c) + (\beta_c-1)\mathbf{E}[\log w_{c,i}] - \beta_c\mathbf{E}[w_{c,i}] \\
&-\sum_{i=1}^{V}\sum_{j=1}^{M}\frac{1}{2\sigma_i^2}\Big(\mathbf{v}_{i,j}^T\mathbf{v}_{i,j} - 2\mathbf{v}_{i,j}^T\mathbf{R}\mathbf{E}[\mathbf{w}_i] + \sum_{l,m=1}^{C}\mathbf{E}[w_{l,i}w_{m,i}]\sum_{k=1}^{S}R_{k,l}R_{k,m}\Big) \\
&-\sum_{i=1}^{V}\sum_{j=1}^{M}\Big(\frac{S}{2}\log 2\pi + \frac{S}{2}\log\sigma_i^2\Big)
\end{aligned} \tag{27}$$

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 (**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 ($\mathbf{z}_i^{(n)} \sim \mathrm{G}_i$) 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 (**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:

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

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:

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

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

where $\mathbf{r_1}$ and $\mathbf{r_2}$ 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 $\mathbf{D}_3$):

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

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

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

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.