Intracranial recordings from human auditory cortex reveal a neural population selective for musical song

Sam V Norman-Haignere[1,2,3]*, Jenelle Feather[4], Peter Brunner[5,6], Anthony Ritaccio[7], Josh H McDermott[4,8,9,10], Gerwin Schalk[5,6,11], Nancy Kanwisher[4,9,10]

[1]Zuckerman Institute, Columbia University
[2]HHMI Fellow of the Life Sciences Research Foundation
[3]Laboratoire des Sytèmes Perceptifs, Département d'Études Cognitives, ENS, PSL University, CNRS, Paris France
[4]Department of Brain and Cognitive Sciences, MIT
[5]Department of Neurology, Albany Medical College, Albany, NY
[6]National Center for Adaptive Neurotechnologies, Wadsworth Center, Albany, NY
[7]Department of Neurology, Mayo Clinic, Jacksonville, Florida, United States of America
[8]Program in Speech and Hearing Biosciences and Technology, Harvard University, Cambridge, Massachusetts, United States of America
[9]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America
[10]Center for Brains, Minds and Machines, Cambridge, Massachusetts, United States of America
[11]Department of Biomed. Sciences, State University of New York

*Corresponding author: sn2776@columbia.edu

Abstract

What is the neural basis of the human capacity for music? Neuroimaging has suggested some segregation between responses to music and other sounds, like speech. But it remains unclear whether finer-grained neural organization exists within the domain of music. Here, using intracranial recordings from the surface of the human brain, we demonstrate a selective response to music with vocals, distinct from responses to speech and to music more generally. Song selectivity was evident using both data-driven component modeling and single-electrode analyses, and could not be explained by standard acoustic features. These results suggest that music is represented by multiple neural populations selective for different aspects of music, at least one of which is specialized for the analysis of song.

Music is a quintessentially human capacity that is present in some form in nearly every society (Savage et al., 2015; Lomax, 2017; Mehr et al., 2018), and that differs substantially from its closest analogues in non-human animals (Patel, 2019). Researchers have long debated whether the human brain has neural mechanisms dedicated to music, and if so, what computations those mechanisms perform (Patel, 2012; Peretz et al., 2015). These questions have important implications for understanding the organization of auditory cortex (Leaver and Rauschecker, 2010; Norman-Haignere et al., 2015), the neural basis of sensory deficits such as amusia (Peterson and Pennington, 2015; Norman-Haignere et al., 2016; Peretz, 2016), the consequences of auditory expertise (Herholz and Zatorre, 2012), and the computational underpinnings of auditory behavior (Casey, 2017; Kell et al., 2018).

Neuroimaging studies have suggested that representations of music diverge from those of other sound categories in non-primary human auditory cortex: some non-primary voxels show partial selectivity for music compared with other categories (Leaver and Rauschecker, 2010; Fedorenko et al., 2012; Angulo-Perkins et al., 2014), and a recent study from our lab, which modeled voxels as weighted sums of multiple response profiles, inferred a component of the fMRI response with clear selectivity for music (Norman-Haignere et al., 2015). However, there are few reports of finer-grained organization within the domain of music (Casey, 2017), potentially due to the coarse resolution of fMRI. As a consequence, we know little about the neural code for music.

Here, we tested for finer-grained selectivity for music using intracranial recordings from the human brain (electrocorticography or ECoG) (**Fig 1A**). We measured ECoG responses to a diverse set of 165 natural sounds, and submitted these responses to a novel decomposition method that is well-suited to the statistical structure of ECoG to reveal dominant response components of auditory cortex. This data-driven method revealed multiple music- and speech-selective response components. Our key finding is that one of these components responded nearly exclusively to music with vocals, suggesting the existence of neural populations that are selective for singing. We then used model-based sound synthesis (Norman-Haignere and McDermott, 2018) to show that these components could not be explained by generic acoustic representations often used to model cortical responses. Finally, we demonstrate direct evidence for music, speech, and song selectivity in individual electrodes without component modeling or statistical assumptions.

Results

Electrode decomposition. We recorded ECoG responses from thirteen patients undergoing surgery for intractable epilepsy. We identified a set of 271 electrodes across all subjects with reliable broadband gamma (70-140 Hz) power responses to the sound set (split-half correlation > 0.2) (**Fig 1B** plots the split-half correlation for all electrodes). We focused on broadband gamma, because it is thought to reflect aggregate spiking in a local region (Steinschneider et al., 2008; Whittingstall and Logothetis, 2009; Ray and Maunsell, 2011). Sound-responsive electrodes were nearly always located near the superior temporal gyrus (STG). Based on prior work, we expected speech selectivity to be prominent in the STG (**Fig 1C**) (Mesgarani et al., 2014; Norman-Haignere et al., 2015; Overath et al., 2015). By contrast, music selectivity is strongest in the lateral sulcus (**Fig 1C**) (Leaver and Rauschecker, 2010; Angulo-Perkins et al., 2014; Norman-Haignere et al., 2015), whose activity cannot be detected with surface electrodes. Thus, we expected music-selective electrodes, if present at all, to be relatively rare.

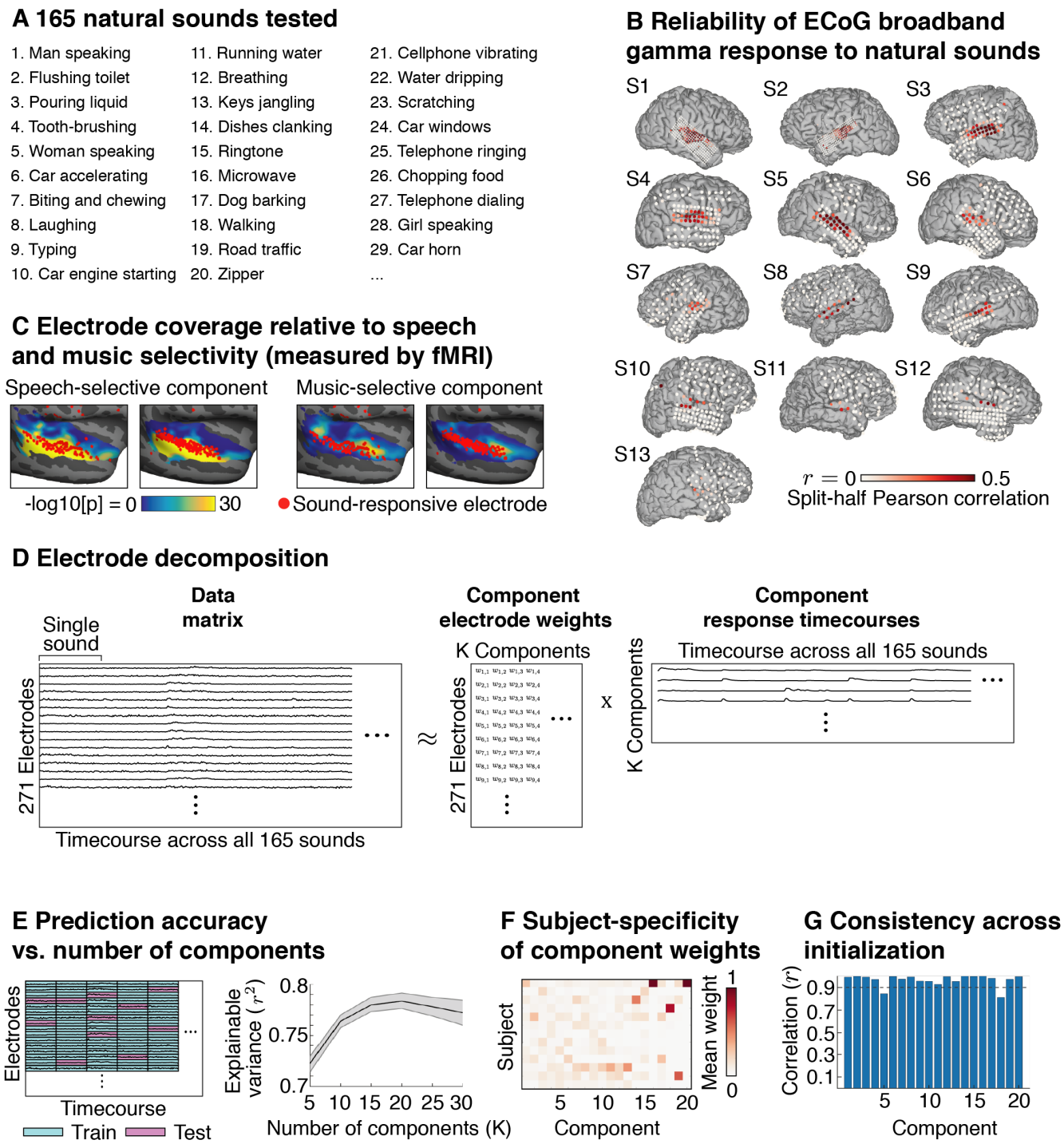

Figure 1. Overview of experiment and electrode decomposition analysis. A, The sound set was composed of 165 commonly heard sounds, each 2-seconds in duration (Norman-Haignere et al., 2015). **B**, Electrode map showing the reliability of broadband gamma responses (70-140 Hz) to natural sounds (split-half Pearson correlation). For each patient, we plot electrodes from the hemisphere in which most or all electrodes were implanted. **C,** Group maps of speech and music selectivity from a prior fMRI study (Norman-Haignere et al., 2015) with the locations of all sound-responsive electrodes overlaid. Maps show the average weight of the speech and music selective components from Norman-Haignere et al., transformed to a measure of significance. Electrodes were projected onto the cortical surface in Freesurfer and aligned to a common template brain. **D**, Schematic of electrode decomposition. The data was represented as a matrix, where each row contains the full response timecourse of each electrode across all 165 sounds tested (the data matrix included responses from 271 sound-responsive electrodes, defined as having a test-retest correlation greater than 0.2). For each sound, we measured responses from a three-second window time-locked to the onset of each sound. The data matrix was approximated as the product of two component matrices: a electrode weight matrix expressing the

contribution of each component to each electrode, and a response matrix containing the response timecourse of each component to the sound set. **E,** Cross-validation was used to compare models (**Fig S2C**) and determine the number of components. The data matrix was divided into cells, with one cell containing the response timecourse of a single electrode to a single sound. The model was trained on a randomly chosen subset of 80% of cells, and responses were then predicted for the remaining 20% of cells. This plot shows the squared test correlation between the measured and predicted response (averaged across all electrodes) as a function of the number of components. The correlation has been noise-corrected using the test-retest reliability of the electrode responses so that it provides a measure of explainable variance. Error bars plot the median and central 68% of the sampling distribution (equivalent to 1 standard error for a Gaussian), computed via bootstrapping across subjects. **F**, The average weight of each component in each subject, normalized so that the weights across subjects sum to 1. Large values indicate that a component primarily explained responses from a single subject. We focused our analyses on components that were not subject-specific, operationalized as having a maximum value across subjects below 0.5 (components 14, 16, 18, 19, and 20 had maximum values greater than 0.5). **G**, The component decomposition algorithm was run 1000 times with different random initializations. This figure plots the median correlation of the inferred response timecourses between the best solution (lowest cost) and the next 99 best solutions. We focused on components with median correlation >0.9.

We sought to identify a small number of response timecourses across the sound set (components) that when weighted together could explain much of the response variance across all 271 electrodes. Each component timecourse could potentially reflect the response of a different neuronal subpopulation in auditory cortex, with the weights providing an estimate for the contribution of each subpopulation to each electrode. To identify components, we represented the response of all 271 electrodes in a matrix, in which each row represented the response timecourse of a single electrode across all 165 sounds (**Fig 1D**). We then tried to approximate this matrix as the product of a component response timecourse matrix and a component electrode weight matrix.

In general, the problem of matrix factorization – finding a set of response components whose weighted sum best explains the data – is ill-posed and needs to be constrained by additional statistical criteria. We identified three statistical properties of auditory broadband gamma activity that are relevant to component modeling (**Fig S1**): (1) broadband gamma responses to sounds are nearly always larger than those to silence (smaller relative responses to sound accounted for <1% of the response power); (2) responses are sparse across both time/stimuli and space/electrodes; (3) responses are temporally smooth, and the extent of this smoothness varies across electrodes. We designed a model that captured all of these statistical properties by convolving a set of sparse/non-negative components with a learned smoothing kernel (**Fig S2**; see Methods for details). We focus on the results of this model because it yielded better prediction accuracy in held-out data than competing models (**Fig S2C**). But we note that our key results were evident using a model that only imposed non-negativity on the responses and weights (**Fig S3**), and were also evident in individual electrodes without using any component modeling (see *Single-electrode analyses* below).

Using a simple cross-validation procedure, in which we trained and tested on separate sounds/electrodes, we found that we could estimate ~15-20 components before overfitting (**Fig 1E**). We show results from a model with 20 components, though all of the speech, music, and song-selective components were evident in a 15-component model (**Fig S4**). Collectively, the 20 components inferred by the model accounted for approximately 78% of the explainable response variation (i.e. the variation that was reliable across repeated presentations). Of these 20 components, fourteen explained responses across multiple subjects (rather than primarily weighting on just a single subject; **Fig 1F**) and were stable across random initializations of the algorithm (**Fig 1G**). We focused on these fourteen components since they are more likely to reflect consistent features of auditory cortical responses.

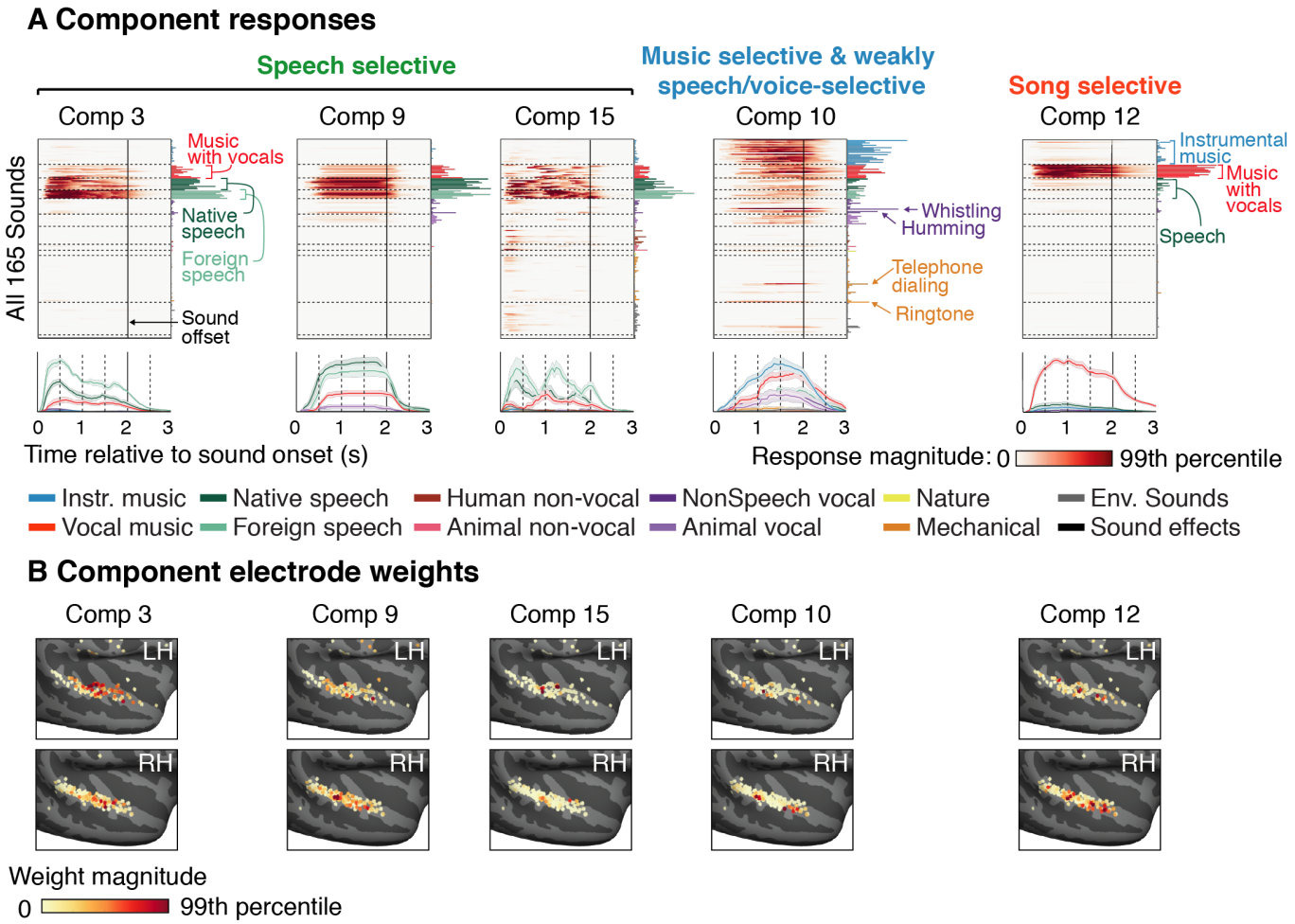

Figure 2. Components responses and electrode weights for five components that responded selectively to speech, music, and or song (Fig S5 plots all reliable components). **A,** The response timecourse of each component to all 165 sounds is plotted as a raster. The time-averaged response to each sound is plotted to the right of the raster. The sounds have been grouped and colored based on membership in one of 12 sound categories (determined primarily based on subject ratings; see *Sound Category Assignments* in Methods). Below each raster, we plot the average response to each category with greater than 5 exemplars. Error bars plot the median and central 68% of the sampling distribution (equivalent to 1 standard error for a Gaussian), computed via bootstrapping across sounds. **B,** Anatomical maps of the electrode weights for each component. To produce this map, each electrode was projected onto the cortical surface, as computed by Freesurfer, and their brain was aligned to a common anatomical template (FsAverage brain).

Component Responses and Weights. For each component, we plot the response timecourse to each of the 165 sounds as a stack of raster plots (**Fig 2A** shows five components that responded selectively to speech, music or song**; Fig S5** shows all fourteen components). The sounds have been grouped based on their membership in one of 12 categories (see *Sound Category Assignments* in Methods). Below each raster, we plot the average response timecourse for each category, and to the right, the time-averaged response to each sound, colored based on category membership. For each component, a map is plotted showing the anatomical distribution of electrode weights (**Fig 2B**; electrode anatomy played no role in the component analysis). Components were numbered based on the overall magnitude of their responses and weights.

Five components responded nearly exclusively to speech or music (**Fig 2**). Three of these components responded selectively to speech (components 1, 9, & 15; average[English speech, foreign speech] > average[all non-speech categories]: $p < 0.001$ via bootstrapping, Bonferroni-corrected for multiple components, see Methods for details). Music with vocals produced an

intermediate response, presumably due to the presence of speech structure (e.g. phonemes, words). The response to English and foreign speech was similar in these components, suggesting a response to speech acoustics rather than linguistic meaning, consistent with prior studies of speech selectivity in the STG (Mesgarani et al., 2014; Norman-Haignere et al., 2015; Overath et al., 2015) (all of the subjects were native English speakers; the response to foreign speech was higher in Components 3 & 15, plausibly because the foreign speech was spoken at faster rate and thus had more speech content). Speech selectivity in Components 3 & 15 developed within a few hundred milliseconds, while Component 9 showed a slower response. The speech-selective components clustered in the middle STG, as expected (Scott et al., 2000; Mesgarani et al., 2014; Overath et al., 2015). The weights for Components 3 & 15 were stronger on average in the left hemisphere, but this effect did not reach significance ($p > 0.07$, uncorrected for multiple components), consistent with prior fMRI studies showing bilateral speech selectivity (Norman-Haignere et al., 2015; Overath et al., 2015) (**Fig S6** plots the mean difference in electrodes weights between right and left hemisphere for all components).

Two components exhibited selectivity for music (Component 10 & 12). Component 10 responded strongly to both instrumental and vocal music (average[Instrumental music, vocal music] > average[all non-music categories]: $p < 0.001$ via bootstrapping, Bonferroni-corrected), and produced an intermediate response to speech, suggesting that music and speech were not perfectly disentangled by our component analysis (perhaps due to limited coverage of the lateral sulcus where music selectivity is prominent; **Fig 1C**). All other non-music and non-speech sounds produced weak responses in these components. Moreover, the response of Component 10 was considerably slower than many of the other components, with music selectivity taking nearly a second to build up, suggesting selectivity for longer-term temporal structure.

Component 12 responded nearly exclusively to music with vocals: every single vocal music stimulus produced a high response and all other sounds, including both speech and instrumental music, produced a weak response. As a consequence, the response to vocal music was significantly higher than the summed response to speech and instrumental music, suggesting nonlinear selectivity for song (vocal music > max[English speech, foreign speech] + instrumental music: $p < 0.001$ via bootstrapping, Bonferroni-corrected). This finding of nonlinear selectivity for vocal music is strengthened by the fact that our decomposition method explicitly models each electrode as a weighted sum of multiple components, and thus if song selectivity simply reflected a sum of speech and music selectivity, the model should not have needed a separate component selective for just vocal music.

Unlike most other components, Components 10 (music selective) and 12 (song selective) showed high weights for electrodes in anterior auditory cortex, similar to what would be expected based on prior work (Leaver and Rauschecker, 2010; Angulo-Perkins et al., 2014; Norman-Haignere et al., 2015). There were also electrodes in middle/posterior STG with substantial weight for these components, which has also been observed with fMRI (Norman-Haignere et al., 2015), though less prominently than the anterior region of music selectivity.

Many components did not exhibit clear selectivity for categories (**Fig S5**). Some components showed strong responses at the onset (Components 1, 2, 4, 6, 7, 8) or offset (Component 17) of sound, although the strength of this onset response varied across stimuli for several components. Most of these onset/offset selective components had weights that were clustered in the middle or posterior STG, but rarely in the anterior STG, consistent with a recent study (Hamilton et al., 2018). Several components were weakly selective for music or speech (Component 7, 8, 13), producing higher

average responses to these categories, but also strong responses for sounds other than speech or music.

Selectivity for spectrotemporal modulation statistics. Can speech, music and song selectivity be explained by generic acoustic representations, such as spectrotemporal modulations that appear to drive much of the functional organization of human primary auditory cortex (Schönwiesner and Zatorre, 2009; Barton et al., 2012; Santoro et al., 2014)? This question is relevant since speech and music are known to have distinctive modulation rates (Singh and Theunissen, 2003; Ding et al., 2017). We recently introduced an algorithm for synthesizing sounds that are matched to natural sounds in their spectrotemporal modulation statistics, despite being acoustically distinct (**Fig 3A**) (Norman-Haignere and McDermott, 2018). We found previously that primary auditory regions produced very similar responses to natural and modulation-matched synthetic sounds, but that non-primary regions produced weak responses to the synthetic sounds, presumably because they lack higher-order structure necessary to drive neurons in non-primary regions.

We measured responses to 36 natural and 36 corresponding modulation-matched synthetic sounds in a subset of ten patients. We used different natural sounds from the 165 sounds tested in the main experiment because we needed longer stimuli for the synthesis procedure (4 seconds vs. 2 seconds; see Methods for details). Of these 36 sounds, there were 8 speech stimuli and 10 music stimuli, two of which contained vocals (these stimuli were designed prior to the discovery of a song-selective component and so were not explicitly designed to examine song selectivity). Using the electrode weights from the 165 natural sounds experiment, we inferred the response of the same 20 components to the new sound set, thus providing an independent validation of their selectivity. We plot the response timecourse of each component to natural and modulation-matched sounds separately for speech, vocal music, instrumental music, and all other non-speech and non-music sounds (**Fig 3B & S7**), as well as the time-averaged response for each pair of natural and modulation-matched sounds (**Fig 3C,D**).

For all category-selective components, we observed a clear difference between the natural and modulation-matched synthetic sounds. The speech-selective components (3, 9, & 15) replicated their selectivity for natural speech with the new stimulus set (with an intermediate response to vocal music) and produced weak responses to the modulation-matched speech ($p < 0.01$ via a sign test across sounds comparing natural and modulation-matched speech). The music-selective component (10) replicated its selectivity for natural music and responded weakly to modulation-matched music ($p < 0.01$ via a sign test comparing natural and modulation-matched music). Critically, the song-selective component (12) responded nearly exclusively to the natural vocal music, producing weak responses to natural speech, natural instrumental music, and the modulation-matched vocal music ($p < 0.01$ via a sign test comparing natural and modulation-matched vocal music; because there were only 2 vocal music sounds, the response to those two stimuli was subdivided into 500 ms segments to increase the number of samples). In contrast, most non-category selective components responded similarly to natural and modulation-matched sounds (**Fig 3D**; Comp 7 showed modest selectivity for natural instrumental music, consistent with its response intermediate selectivity for instrumental music in the 165 natural sounds; see **Fig S5**). This finding demonstrates that speech, music, and song selectivity cannot be accounted for by spectrotemporal modulation statistics that appear to robustly drive responses throughout much of the rest of auditory cortex.

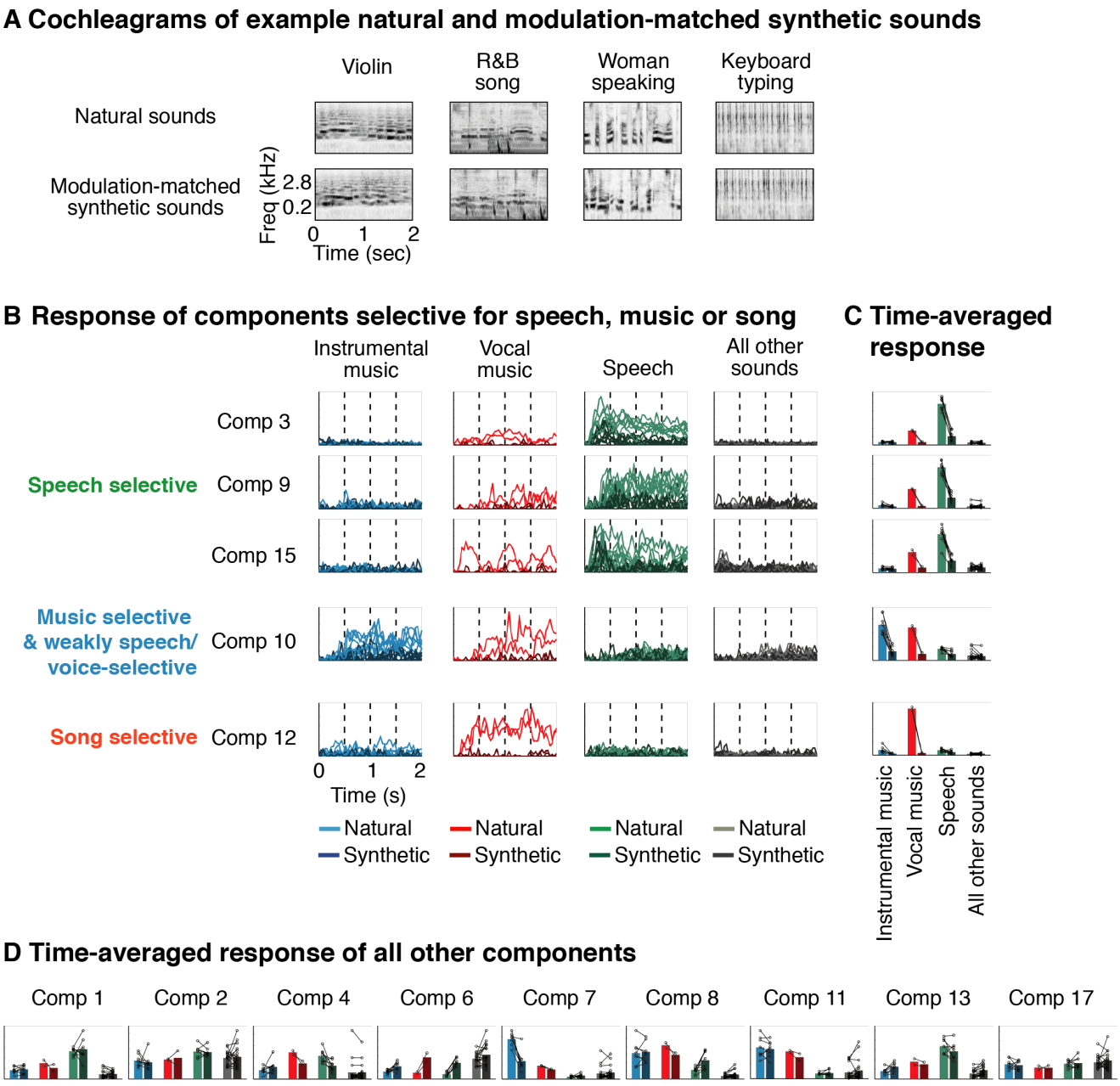

Figure 3. Component responses to natural and modulation-matched synthetic sounds. **A,** Cochleagrams of example natural and corresponding synthetic sounds with matched spectrotemporal modulation statistics (Norman-Haignere and McDermott, 2018). Cochleagrams plot energy as a function of time and frequency, similar to a spectrogram, but measured from filters designed to mimic cochlear frequency tuning. Each sound was 4 seconds in duration (cochleagrams show just the first two seconds of each sound). **B,** The response of the speech, music, and song-selective components, identified in the 165-natural sound experiment, to the natural and modulation-matched sounds of the control experiment. We plot the response timecourse (first 2-seconds) of each component to each natural (lighter colors) and modulation-matched synthetic sound (darker colors). The sounds are grouped into four categories: instrumental music (blue), music with vocals (red), speech (green, both English and foreign), and all other sounds (black/gray). **C**, The time-averaged component response to each pair of natural and modulation-matched sounds (connected circles indicate pairs), along with the mean component response across the natural (lighter bars) and modulation-matched (darker bars) sounds from each category. **D,** Same as panel **C**, but showing all other reliable components, most of which showed a similar response magnitude for natural and modulation-matched sounds.

Single-electrode analyses. We next sought to test whether we could observe evidence for speech, music, and song selectivity in individual electrodes without the need for statistical assumptions or modeling. Using a subset of data, we identified electrodes selective for speech, music or song, and

then measured their response in independent data. The electrode selection stage involved three steps (all performed on the same data and distinct from that used to measure the response). First, we measured the average response across time and stimuli to all sound categories with more than five exemplars. Second, we identified a pool of electrodes with a highly selective (selectivity > 0.6) and significant ($p < 0.001$ via bootstrapping) response to either speech, music or song. Selectivity was measured by contrasting the maximum response across all speech and music categories (English speech, foreign, speech, vocal music, instrumental music) with the maximum response across all other non-music and non-speech categories. Third, from this pool of music- or speech-selective electrodes, we formed three groups: those that responded significantly more ($p < 0.01$ via bootstrapping) to speech than all else (max[English speech, foreign speech] > max[non-speech categories except vocal music]), music than all else (instrumental music > max[non-music categories]), or that exhibited super-additive selectivity for vocal music (vocal music > max[English speech, foreign speech] + instrumental music).

We plot the response of the top electrodes most significantly responsive to each contrast (**Fig 4A**) as well as the average response across all electrodes identified using this procedure (**Fig 4B**). We measured responses to the same natural sounds used to identify the electrodes (in independent data), as well as the natural and synthetic sounds from our control experiment (**Fig 4C**). As expected, given the coverage of ECoG grids relative to speech and music-selectivity (**Fig 1C**), we observed many more speech-selective electrodes than music or song-selective electrodes (169 speech-selective electrodes across all 13 subjects, 12 music-selective electrodes across 4 subjects, and 7 song-selective electrodes across 3 subjects). But each of the music and song-selective electrodes identified replicated their selectivity for music or speech in independent data ($p < 0.05$ for every electrode; $p < 0.001$ for responses averaged across all music and song-selective electrodes; via bootstrapping the same contrast used to select electrodes but in independent data); and modulation-matched synthetic sounds produced a much weaker responses than natural sounds from the preferred category ($p < 0.01$ via a sign test between responses to natural and model-matched sounds applied to the average response of speech, music, and song-selective electrodes). Some of the music-selective electrodes were strikingly selective. For example, S1-E147 (from a patient with small, high-density electrodes with 1 mm exposed diameters) responded in a near binary fashion, producing a high response for nearly all of the music sounds and a near-zero response for all other sounds.

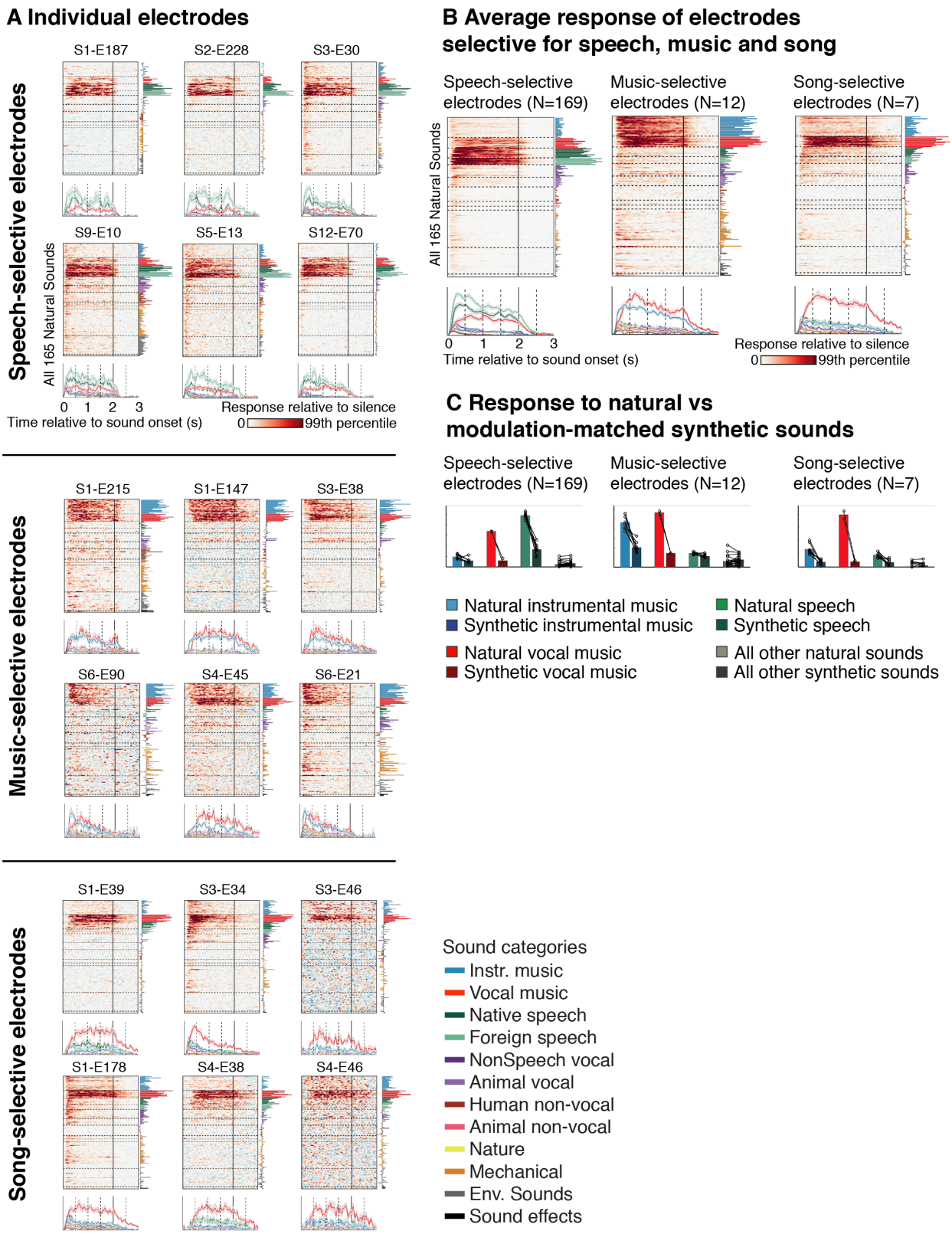

Figure 4. **The response of individual electrodes selective for speech, music or song.** We selected speech (top), music (middle), and song-selective (bottom) electrodes, and then measured their response in independent data. **A,** The top six electrodes that showed the most significant response preference for each category in the subset of data used to select electrodes. For speech-selective electrodes, the top 6 electrodes came from 2 subjects (2 from S1 and 4 from S2), and so we instead plot the top electrode from 6 different subjects to show greater diversity. Conventions as in **Fig 2A**. **B,** The average response of all electrodes identified as speech, music, or song-selective to the 165 natural sounds (same conventions as panel A). **C**, The average response of speech, music, and song-selective electrodes to natural and modulation-matched sounds from the control experiment (conventions as in **Fig 3C**).

The fact that we observed clear selectivity for vocal music in individual electrodes confirms that our component analysis did not infer a form of selectivity not present in the data. At the same time, the song-selective electrodes identified in this analysis were less selective than the component inferred by our decomposition analysis ($p < 0.001$ via bootstrapping the super-additive song selectivity metric), which suggests that our component analysis disentangled overlapping selectivity for music, speech and song within individual electrodes. Moreover, the song-selective component explained responses in a much wider range of electrodes than the 7 electrodes identified in our single-electrode analysis; indeed, the top 7 electrodes with the greatest weight for the song-selective component (Component 12) accounted for less than 20% of the total electrode weights. Thus, by de-mixing selectivity within individual electrodes, our component analysis isolated selectivity for song more cleanly and enabled us to better characterize the spatial distribution of song selectivity across the cortex.

Discussion

Using intracranial recordings from the human brain, our study reveals two distinct forms of music selectivity: one selective for a wide range of music, and one selective for music with vocals, suggesting selectivity for song. Both types of selectivity emerged from data-driven component modeling and were also evident in analyses of individual electrodes. Neither form of selectivity could be explained by a generic acoustic model based on spectrotemporal modulation. Our results suggest that music is represented by multiple distinct neural populations, selective for different aspects of music, at least one of which responds specifically to musical song.

Song selectivity. Although vocal music has frequently been used to explore the neural basis of music and speech perception (Merrill et al., 2012; Tierney et al., 2013), our findings provide the first evidence for a neural population specifically involved in the perception of song. Because our component method explicitly models electrodes as weighted sums of multiple response patterns, the method would not have inferred a component selective for vocal music if each electrode reflected a weighted sum of speech and music selectivity. Thus, the fact that our component analysis inferred a component that responded nearly exclusively to vocal music provides evidence for a super-additive response to singing, a hypothesis that we directly confirmed by analyzing the response of song-selective electrodes.

Why might the human brain have neural populations selectively responsive to song? Vocals are pervasive in music, and typically carry the main melodic line. Thus, the brain may develop neural mechanisms specialized for representing song, simply because it is one of the first and/or most prominent components of the music that people hear. Alternatively, neural specializations for song may be partly innate, reflecting the biological importance of singing (Mehr and Krasnow, 2017).

Why has song selectivity not been clearly observed before, including in our prior fMRI study using the same sound set (Norman-Haignere et al., 2015)? One possibility is that ECoG signals have greater spatial and temporal precision because they directly sample electrophysiological activity rather than using changes in blood flow to track neural activity. Consistent with this hypothesis, in our prior fMRI study, we were only able to infer six reliable response patterns across all of auditory cortex before overfitting to noise in the data. Here, we were able to infer a much larger number of components despite having access to only a fraction of auditory cortex (since surface electrodes do not provide coverage of the lateral sulcus). Most of the inferred components had distinct responses to the sound set even when averaging responses across time, suggesting that the increase in dimensionality is not solely due to improved temporal resolution.

It will be important in future work to identify the features of singing that drive song selectivity. For example, one could explore sensitivity to the types of pitch variation that characterize singing (Tierney et al., 2013), or test for an interaction between speech-like vocal tract information and musical pitch variation (Merrill et al., 2012).

Music selectivity. Researchers have long debated the extent to which music perception relies on specialized vs. general-purpose neural mechanisms (Patel, 2012; Peretz et al., 2015)**.** Our study provides the first direct electrophysiological evidence that the human brain has neural populations that are highly selective for music.

Our results also help validate the voxel decomposition method used in our prior work (Norman-Haignere et al., 2015). Using voxel decomposition, we inferred a component that was substantially more selective for music than were individual voxels, which we hypothesized was due to the overlap of distinct neural populations within a voxel. Our findings support this hypothesis by showing clear music selectivity using a more direct measure of neural activity. Moreover, many of the electrodes that showed the strongest selectivity for music (e.g. S1-E147, S1-E215) were sampled by a high-density grid with particularly small electrodes (1 mm exposed diameter), suggesting that high spatial resolution is useful for detecting clear music selectivity. Thus, our study both demonstrates the existence of music-selective neural populations, and helps explain why this type of selectivity has not been clearly observed with fMRI in standard voxel-wise analyses.

Speech and voice selectivity. Many prior studies have reported selectivity for speech (Mesgarani et al., 2014; Norman-Haignere et al., 2015; Overath et al., 2015) and non-speech vocalizations (e.g. crying, laughing) (Belin et al., 2000) in the superior temporal gyrus. Distinguishing responses to speech and voice has been difficult, because speech-selective responses typically show at least some response to non-speech vocalizations and vice-versa. Here, we found multiple components (3, 15) and electrodes (e.g. S2-E54, S2-E222) that produced essentially no response to non-speech vocalizations, demonstrating that pure speech selectivity exists in the human brain. Thus, as with the music selectivity, the fact that fMRI voxels reflect a mixture of speech and voice selectivity may in part reflect the blurring together of nearby neural populations.

Onset/offset selectivity. Many of the components we observed responded substantially more strongly at the onset or offset of sound, consistent with a recent study showing the onset selectivity is a prominent feature of human STG responses (Hamilton et al., 2018). Our study highlights the diversity of these responses across a wide variety of natural sounds: some components responded at the onset (Component 2) or offset (Component 17) of any sound, some were strongest for speech or vocalization stimuli (Components 1 & 4), and some were strongest for non-speech sounds (Component 6). Why so much of the STG is onset-selective is unclear. Some of these responses might reflect a generic/low-level adaptation mechanism in response to a sudden increment or decrement in sound energy. Others might reflect adaptation to higher-level stimulus statistics (Kvale and Schreiner, 2004), perhaps in the service of creating a more noise-robust (Mesgarani et al., 2014) or efficient (Barlow, 1961; Fairhall et al., 2001) representation of sound by suppressing responses to features that are predictable (Heilbron and Chait, 2017).

Component modeling: strengths, limitations and relationship to prior methods. Component modeling provides a way to: (1) infer prominent response patterns; (2) suggest novel hypotheses that might not be obvious a-priori; and (3) disentangle spatially overlapping responses. Our results illustrate each of these benefits. We were able to infer a set of 20 response components that explained much of the response variation across hundreds of electrodes. We found evidence for a novel form of music selectivity (song selectivity) that we did not hypothesize a priori. And the

selectivity that we observed in the song selective component was often clearer than that evident in individual electrodes, some of which appeared to reflect a mixture of music, speech and song selectivity.

The key challenge of component modeling is that matrix approximation is inherently ill-posed, and hence, the solution depends on statistical assumptions. Most component methods rely on just one of the following three assumptions: (1) non-negativity (Lee and Seung, 1999); (2) sparsity across time or space (Olshausen and Field, 1997; Hyvarinen, 1999); or (3) temporal smoothness (Wiskott and Sejnowski, 2002; Byron et al., 2009). We showed that all of these properties are evident in auditory ECoG responses. We developed a simple model to embody these assumptions and showed that the model better predicted ECoG responses compared with baseline models. We also showed that all of our category-selective components were evident using a model that imposed only non-negativity on the responses, suggesting that our key results were robust to the particular statistical assumptions imposed. Nonetheless, the assumptions of a component model are never perfect; and thus, it is useful to validate the results of a model with more direct analyses. Here, we found that speech, music and song selectivity were evident in individual electrodes, which demonstrates that our key findings were not dependent on statistical assumptions.

Our prior fMRI voxel decomposition method used statistical constraints on the high-dimensional voxel weights to infer components (Norman-Haignere et al., 2015). By contrast, ECoG grids have many fewer electrodes than voxels, but each electrode has a richly structured timecourse. We thus chose to constrain the solution with statistics of the high-dimensional response timecourses. Our method is also distinct from a number of other component models that have been applied to high-dimensional neural data. Unlike many sparse convolutional models (Bouchard et al., 2017), each component of our model was defined by a single timecourse and a single pattern of electrode weights rather than by a time-varying spatial pattern, and thus can be more easily interpreted as the response of an underlying neuronal population. Unlike clustering methods (or convex NMF (Hamilton et al., 2018)), our method can disentangle responses that overlap within individual electrodes. And unlike most tensor decomposition methods (Williams et al., 2018), our method does not require the shape of a component's response timecourse to be identical across different stimuli, which is critical for modeling responses to sensory features that are not necessarily aligned to stimulus onset.

Conclusions and future directions

By revealing a neural population selective for song, our study begins to unravel the neural code for music in the human brain, raising many questions for future research: What features of music underlie selective responses to music and song? Do these responses reflect note-level structure (e.g. pitch and timbre) (Casey et al., 2012) or the way notes are patterned to create music (e.g. melodies, harmonies and rhythms) (Schindler et al., 2013)? How can we describe the tuning of music and song-selective neural populations in computational terms, given that standard acoustic features appear insufficient (Kell et al., 2018)? And how is music and song selectivity constructed over the development of each individual, or over the history of our species (Wallin et al., 2001)? The findings and methods presented here provide a path towards answering these longstanding questions.

Methods

Participants. Thirteen epilepsy patients participated in our study (mean age: 37 years, age standard deviation: 14 years; 8 right-handed; 8 female). These subjects underwent temporary implantation of subdural electrode arrays at Albany Medical College to localize the epileptogenic zones and to delineate these zones from eloquent cortical areas before brain resection. All of the subjects gave

informed written consent to participate in the study, which was approved by the Institutional Review Board of Albany Medical College.

Electrode grids. Most subjects had electrodes implanted in a single hemisphere, and STG coverage was much better in one of the two hemispheres in all subjects (8 right hemisphere patients and 5 left hemisphere patients; **Fig 1B** shows the coverage of the primary hemisphere for all subjects). In most subjects, electrodes had a 2.3 mm exposed diameter with a 6 mm center-to-center spacing for temporal lobe grids (10 mm spacing for grids in frontal, parietal and occipital lobe, but electrodes from these grids typically did not show reliable sound-driven responses; electrodes were embedded in silicone; PMT Corp., Chanhassen, MN). Two subjects were implanted with a higher-density grid (1 mm exposed diameter, 3 mm center-to-center spacing).

Natural sounds. The sound set was the same as in our prior study (Norman-Haignere et al., 2015). To generate the 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 (the 160-sound set included only two foreign speech stimuli: "Spanish" and "Chinese"). In total, there were 10 English speech stimuli, 7 foreign speech stimuli, 21 instrumental music stimuli, and 11 vocal music stimuli (see *Sound category assignments*). Sounds were RMS-normalized and presented at a comfortable volume using sound isolating over-the-ear headphones (Panasonic RP-HTX7, 10 dB isolation). The sound set is freely available:

http://mcdermottlab.mit.edu/svnh/Natural-Sound/Stimuli.html

Subjects completed between three and seven runs of the experiment (S11: 3 runs, S6, S12: 4 runs, S13: 5 runs, S1: 7 runs; all other subjects: 6 runs). In each run, each natural sound was presented at least once. Between 14 and 17 of the sounds were repeated exactly back-to-back, and subjects were asked to press a button when they detected this repetition. This second instance of the sound was excluded from the analysis, because the presence of a target could otherwise bias responses in favor of the repeated stimuli. Each run used a different random ordering of stimuli. There was a 1.4-2 second gap (randomly chosen) between consecutive stimuli.

Modulation-matched synthetic sounds. In ten of the subjects, we also measured responses to a distinct set of 36 natural sounds and 36 corresponding synthetic sounds that were individually matched to each natural sound in their spectrotemporal modulations statistics using the approach described in Norman-Haignere & McDermott (2018). The synthesis algorithm starts with an unstructured noise stimulus, and iteratively modifies the noise stimulus to match the modulation statistics of a natural sound. Modulations are measured using a standard model of auditory cortical responses in which a cochleagram is passed through a set of linear filters tuned to modulations at a particular audio frequency, temporal rate, and spectral scale (i.e. how coarse vs fine the modulations are in frequency) (Chi et al., 2005). The spectrotemporal filters were created by crossing 9 temporal rates (0.5, 1, 2, 4, 8, 16, 32, 128 Hz) with 7 spectral scales (0.125, 0.25, 0.5, 1, 2, 4, 8 cycles per octave), and replicating each filter at each audio frequency. The synthesis procedure alters the noise stimulus to match the histogram of response magnitudes across time for each filter in the model, which has the effect of matching all time-averaged statistics (such as mean and variance) of the filter

responses. The stimuli and synthesis procedures were very similar to those used in Norman-Haignere & McDermott with a few minor exceptions that are noted next.

All stimuli were 4 seconds long. We used shorter stimuli than the 10-second stimuli used in Norman-Haignere & McDermott (2018) due to limitations in the recording time. Because the stimuli were shorter, we did not include the very low-rate filters (0.125 and 0.25 Hz), which were necessary for longer stimuli to prevent energy from clumping unnaturally at particular moments in the synthetic recording. We also did not include "DC filters" as in Norman-Haignere & McDermott, but instead simply matched the mean value of the cochleagram across time and frequency at each iteration of the algorithm. Norman-Haignere & McDermott describe two versions of the algorithm: one in which the histogram-matching procedure was applied to the raw filter outputs and one where the matching procedure was applied to the envelopes of the filter responses. We found that the resulting stimuli were very similar, both perceptually and in terms of the cortical response. The stimuli tested here were created by applying the histogram matching procedure to the envelopes.

The stimuli were presented in a random order with a 1.4-1.8 second gap between stimuli (for the first subject tested, a 2-2.2 second gap was used). The natural sounds were repeated to make it possible to assess the reliability of stimulus-driven responses. For all analyses, we simply averaged responses across the two repetitions. The sound set was presented across 4 runs. In one subject (S1), the entire experiment was repeated (we averaged responses across the two repeats).

Sound category assignments. In an online experiment, Mechanical Turk participants chose the category that best described each of the 165 sounds tested, and we assigned each sound to its most frequently chosen category (30–33 participants per sound) (Norman-Haignere et al., 2015). Category assignments were highly reliable (split-half kappa = 0.93). We chose to re-assign three sounds ("cymbal crash", "horror film sound effects", and "drum roll") from the "instrumental music" category to a new "sound effects" category. There were two motivations for this re-assignment: (1) these three sounds were the only sounds assigned to the music category that produced weak fMRI responses in the music-selective component we inferred in our prior study, presumably because they lack canonical types of musical structure (i.e. clear notes, melody, rhythm, harmony, key, etc.); and (2) excluding these sounds makes our song selectivity contrast (vocal music – (instrumental music + speech)) more conservative as it is not biased upwards by the presence of instrumental music sounds that lack rich musical structure.

Preprocessing. Preprocessing was implemented in MATLAB. The scripts used to implement the preprocessing steps are available here (we reference specific scripts within these directories in describing our analyses):

https://github.com/snormanhaignere/ecog-analysis-code
https://github.com/snormanhaignere/general-analysis-code

The responses from all electrodes were common-average referenced to the grand mean across all electrodes (separately for each subject). We excluded noisy electrodes from the common-average reference by detecting anomalies in the 60 Hz power (see channel_selection_from_60Hz_noise.m; a IIR resonance filter with a 3dB down bandwidth of 0.6 Hz was used to measure the RMS power). Specifically, we excluded electrodes whose 60 Hz power exceeded 5 standard deviations of the median across electrodes. Because the standard deviation is itself sensitive to outliers, we estimated the standard deviation using the central 20% of samples, which are unlikely to be influenced by outliers (by dividing the range of the central 20% of samples by that which would be expected from a Gaussian of unit variance; see zscore_using_central_samples.m). After common-average

referencing, we used a notch filter to remove 60 Hz noise and its harmonics (60, 120, 180, and 240 Hz; see notch_filt.m; an IIR notch filter with a 3dB down bandwidth of 1 Hz was used to remove individual frequency components; the filter was applied forward and backward using filtfilt.m).

We computed broadband gamma power by measuring the envelope of the preprocessed signal filtered between 70 and 140 Hz (see bandpass_envelopes.m; bandpass filtering was implemented using a 6th order Butterworth filter with 3dB down cutoffs of 70 and 140 Hz; the filter was applied forward and backward using filtfilt.m). The envelope was measured as the absolute value of the analytic signal after bandpassing. For the single-electrode analyses (**Fig 4**), we downsampled the envelopes to 100 Hz (from the 1200 Hz recording rate), and smoothed the timecourses with a 50 ms FWHM kernel to reduce noise and make it easier to distinguish the timecourses for different categories in the plots. For component analyses, we downsampled the envelopes to 25 Hz, because this enabled us to fit the data in the limited memory available on the GPU used to perform the optimization (no smoothing was applied since the model inferred an appropriate smoothing kernel for each component).

Occasionally, we observed visually obvious artifacts in the broadband gamma power for a small number of timepoints. These artifacts were typically localized to a small fraction of electrodes; thus, we detected artifacts separately for each electrode. For each electrode, we computed the 90th percentile of its response magnitudes across all timepoints, which is by definition near the upper-end of that electrode's response distribution, but which should also be unaffected by outliers assuming the number of outliers accounts for less than 10% of time points (which we generally found to be the case). We classified a timepoint as an outlier if it exceeded 5 times the 90th percentile value for each electrode. We found this value to be relatively conservative in that only a small number of timepoints were excluded (<1% for all sound-responsive electrodes), and the vast majority of the excluded timepoints were artifactual based on visual inspection of the broadband gamma timecourses. Because there were only a small number of outlier timepoints, we replaced the outliers values with interpolated values from nearby non-outlier timepoints. We also manually excluded some or all of the runs from 11 electrodes where there were a large number of artifacts.

For each 2-second stimulus, we measured the response of each electrode during a three-second window locked to sound onset (for the 4-second natural and modulation-matched stimuli we used a 5-second window). We detected the onset of sound in each stimulus by computing the waveform power in 10 ms bins (with a 2 ms hop), and selecting the first bin in which the audio power exceeded 50 dB of the maximum power across all windows and stimuli. Following standard practice, the audio and ECoG data were synced by sending a copy of the audio signal to the same system used to record ECoG signals. This setup allowed us to measure the time delay between when the system initiated a trial and the onset of sound (which we measured using pure tones).

Responses were converted to units of percent signal change relative to silence by subtracting and then dividing the response of each electrode by the average response during the 300 ms before each stimulus.

Session effects. For five of the thirteen subjects, runs were collected across two sessions with a gap in between (typically a day; the 7th run of S1 was collected in a third session). For the vast majority of electrodes, we found that their response properties were stable across sessions. Occasionally, we observed electrodes that substantially changed their selectivity, potentially due to small changes in the positioning of the electrodes over the cortex. To identify such changes, from each run of data, we measured the time-averaged response of each electrode to each of the 165 sounds tested. We then detected electrodes for which the test-retest correlation from runs of the

same session was significantly greater than the test-retest correlation from runs of different sessions ($p < 10^{-5}$; we used time-averaged response profiles rather than the raw timecourses, because we found them to be more reliable, and thus a better target for detecting selectivity changes across sessions; for S1 we grouped the runs from the 2nd and 3rd session together since there was only a single run in the 3rd session). Significance was evaluated via a permutation test (Nichols and Holmes, 2002) in which we permuted the correspondence between stimuli across runs (10,000). We used this approach to build up a null distribution for our test statistic (the difference between the correlation within and across sessions), fit this null distribution with a Gaussian (so that we could estimate small p-values that would have been impossible to estimate via counting), and used the null to calculate a two-sided p-value (by measuring the tail probability that exceeded the test statistic and multiplying by 2). Seven electrodes passed our conservative significance threshold. For these electrodes, we simply treated the data from different sessions as coming from different electrodes, since they likely sampled distinct neural populations.

Electrode selection. We selected electrodes with a reliable response to the sound set. Specifically, we measured the test-retest correlation of each electrode's broadband gamma response timecourse across all sounds, measured in two splits of data (odd and even runs). We kept all electrodes with a test-retest correlation greater than 0.2 (electrodes with a test-retest correlation less than 0.2 were quite noisy upon inspection). Results were similar using a more liberal threshold of 0.1.

Electrode localization. We localized electrodes in order to be able to visualize the electrode weights for each component. Electrode locations played no role in the identification of components or category-selective electrodes.

Following standard practice, we identified electrodes as bright spots on a post-operative computer tomography (CT) image. The CT was the aligned to a high-resolution, pre-operative magnetic resonance image (MRI) using a rigid-body transformation. We then projected each electrode onto the cortical surface, computed by Freesurfer from the MRI scan. This projection is error-prone because far-away points on the cortical surface can be spatially nearby due to cortical folding. As a consequence, it was not uncommon for electrodes very near STG, where sound-driven responses are common, to be projected to a spatially nearby point on middle temporal or supramarginal/inferior frontal gyrus, where sound-driven responses are much less common (**Fig S8**). To minimize gross errors, we preferentially localized sound-driven electrodes to regions where sound-driven responses are likely to occur. Specifically, using a recently collected fMRI dataset, where responses were measured to the same 165 sounds in a large cohort of 20 subjects with whole-brain coverage (our prior published study only had partial brain coverage (Norman-Haignere et al., 2015)), we calculated the fraction of subjects that showed a significant response to sound at each point in the brain ($p < 10^{-5}$, measured using a permutation test (Norman-Haignere et al., 2016)). We treated this map as a prior and multiplied it by a likelihood map, computed separately for each electrode based on the distance of that electrode to each point on the cortical surface (using a Gaussian error distribution; 10 mm FWHM). We then assigned each electrode to the point on the cortical surface where the product of the prior and likelihood was greatest (which can be thought of as the maximum posterior probability solution). We smoothed the prior probability map so that it would only effect the localization of electrodes at a coarse level, and not bias the location of electrodes locally, and we set the minimum prior probability to be 0.05 to ensure every point had non-zero prior probability. We plot the prior map and the effect it has on localization in **Fig S8**.

Responses statistics relevant to component modeling. Our component model approximated the response of each electrodes as the weighted sum of a set of canonical response timecourses ("components"). The component timecourses are shared across all electrodes, but the weights are

unique. We modeled each electrode as the weighted sum of multiple components because each electrode reflects the pooled activity of many neurons. When the electrode response timecourses are compiled into a matrix, our analysis corresponds to matrix factorization: approximating the data matrix as a product of a component response matrix and a component weight matrix.

Matrix factorization is inherently ill-posed (that is, there are many equally good approximations). Thus, we constrained our factorization by additional statistical criteria. Most component methods rely on one of three statistical assumptions: (1) non-negativity (Lee and Seung, 1999); (2) a non-Gaussian distribution of response magnitudes across time or space (Olshausen and Field, 1997; Hyvarinen, 1999); or (3) temporal smoothness of the component responses (Wiskott and Sejnowski, 2002; Byron et al., 2009). We investigated each of these statistical properties in broadband gamma responses to sound (**Fig S1**).

To evaluate non-negativity, we measured the percent of the total RMS power accounted for by positive vs. negative responses during the presentation of sound (measured relative to 300 ms of silence preceding the onset of each sound):

$$100 * \sqrt{\frac{\sum p^2}{\sum p^2 + \sum n^2}} \qquad 1$$

where p and n are shorthand for positive and negative values. We applied the above equation to the response of individual electrodes (summing over all timepoints for all sounds; **Fig S1A,B**), as well as to the pooled response of all sound-responsive electrodes (summing over all timepoints, sounds, and electrodes; **Fig S1D**). To minimize the effect of measurement noise, which will create negative values even if none are present (since measurement noise will not depend on the stimulus and thus noise fluctuations will be symmetric around the silent baseline), we measured the response of all electrodes in two splits of data (average across odd and even runs). We then: (1) sorted the response magnitudes of all timepoints by their magnitude in the first split; (2) measured their response in the second split; and (3) applied a median filter to the sorted response magnitudes from the second splits, thus suppressing unreliable response variation (filter size = 100 when applied to individual electrodes, filter size = 10,000 when pooling responses across all electrodes) (**Fig S1B&D** show the results of applying this procedure to individual electrodes and the pooled response of all electrodes). When equation 1 was applied to the de-noised response distributions (i.e. median filtered responses from the second split), we found that positive responses accounted for 99.97% of the RMS power across all sound-responsive electrodes. Note that sound-responsive electrodes were selected based on the reliability of their responses, not based on a greater response to sounds compared with silence, and thus our analysis is not biased by our selection criterion.

To investigate whether and how the distribution of responses might differ from a Gaussian, we measured the skewness (normalized 3rd moment) and sparsity (excess kurtosis relative to a Gaussian) of the responses:

$$\text{skewness} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}\right)^3} \qquad 2$$

$$\text{sparsity} = \log\left[\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^2} - 3\right] \qquad 3$$

We applied the above equations to the response distribution of each electrode across all timepoints and sounds (i.e. concatenating the timecourses from all sounds into a single vector), denoised using the procedure described in the preceding paragraph. **Fig S1F** plots a histogram of these skewness and sparsity values across all electrodes. We found that all electrodes were skewed and sparse relative to a Gaussian, and relative to what would be expected given just noise in the data ($p < 0.001$ via sign test; see *Statistics* for details). This observation implies that the response distribution of each electrode across time/stimuli has a heavy rightward tail, with a relatively small fraction of timepoints yielding large responses for any given electrode.

We also tested the skewness and sparsity of responses across electrodes by applying equations 2 and 3 to the distribution of responses across electrodes. Specifically, we measured the averaged response of each electrode to each sound, and then for each sound, we applied equations 2 and 3 to the distribution of responses across the 271 sound-responsive electrodes. **Fig S1G** plots the histogram of these skewness and sparsity measures for all 165 sounds. We did not apply our de-noising procedure since we only had 271 electrodes which made the sorting/median-filtering procedure infeasible (in contrast, for each electrode we had 12,375 timepoints across all sounds); instead we time-averaged the response of each electrode to each sound to reduce noise. We again found that this distribution was significantly skewed and sparse relative to a Gaussian and relative to what would be expected given just noise in the data ($p < 0.001$ via sign test).

Finally, to investigate the temporal smoothness of auditory ECoG responses, we measured the normalized autocorrelation of each electrode's response (**Fig S1C,E**). To prevent noise from influencing the result, we correlated responses measured in independent runs (odd and even runs). This analysis revealed substantial long-term dependencies over more than a second, and the strength of these dependencies varied substantially across electrodes. This substantial variation across electrodes demonstrates that these long-term dependencies are not a by-product of measuring broadband gamma power (in simulations, we have found that our measurement procedure can resolve power fluctuations up to ~30 Hz, assuming a 70-140 Hz carrier).

Together, the results from our analysis reveal three key properties of auditory ECoG: (1) nearly all responses are positive/excitatory relative to sound onset; (2) responses are skewed/sparse across time/stimuli and electrodes; and (3) responses are temporally smooth and the extent of this smoothness varies across electrodes. We sought to design a simple component model that captures these three essential properties. We refer to this model as the "Sparse and Smooth Component" (SSC) model.

Component model. Each electrode is represented by its response timecourse across all sounds ($\boldsymbol{e}_i(t)$) (**Fig S2A**). We approximate this response timecourse as the weighted sum of K component response timecourses ($\boldsymbol{r}_k(t)$):

$$\mathbf{e}_i(t) \approx \sum_{k=1}^{K} \mathbf{r}_k(t) w_{i,k} \qquad \text{s4}$$

The component timecourses are shared across all electrodes, but the weights are separate for each electrode, allowing the model to approximate different response patterns. We constrain all of the component responses and weights to be positive, since we found that nearly all of the sound-driven responses were positive. To encourage the components to be both sparse and smooth, we model the response timecourse of each component as the convolution of a set of sparse activations ($\boldsymbol{a}_k(t)$) with a smoothing kernel ($\boldsymbol{h}_k(t)$):

$$\mathbf{r}_k(t) = \mathbf{a}_k(t) * \mathbf{h}_k(t) \quad (5)$$

The activations effectively determine when responses occur and the smoothing kernel determines their smoothness. The activations, smoothing kernel, and electrode weights are all learned from the data. The learning algorithm proceeds by minimizing the cost function below, which has two parts: (1) a reconstruction penalty that encourages the model to be close to the data; and (2) an L1 penalty that encourages the component activations and weights to be sparse.

$$\min_{\{w_{k,i}\},\{\mathbf{a}_k\},\{\mathbf{h}_k\}} \sum_i \Big(\mathbf{e}_i(t) - \sum_{k=1}^{K} \mathbf{r}_k(t) w_{i,k}\Big)^2 + \lambda\Big(\sum_i \sum_{k=1}^{K} w_{i,k} + \sum_{k=1}^{K} \sum_t \mathbf{a}_k(t)\Big) \quad (6)$$

We allowed the smoothing kernel to vary across components to capture the fact that different electrodes have variable levels of smoothness. We forced the kernel to be smooth by constraining it to be unimodal (see *Constraining the Smoothing Kernel* below). The learned smoothing kernels for each component are shown in **Fig S9**. The kernels vary substantially in their extent/duration, thus capturing varying levels of smoothness across components. The model has two hyper-parameters: the number of components (K) and the strength of the sparsity penalty (λ), which we chose using cross-validation (see next section).

We implemented and optimized the model in TensorFlow, which provides efficient, general-purpose routines for optimizing models composed of common mathematical operations. We used the built-in ADAM optimizer to minimize the loss. We ran the optimizer for 10,000 iterations, decreasing the step size after each 2,000 iterations (in logarithmically spaced intervals; from 0.01 to 0.0001). Positivity of the activations and electrode weights was enforced by representing each element as the absolute value of a real-valued latent variable.

As with any sparse component model, our cost function is not convex, and the optimization algorithm could potentially arrive at local optima, leading to unstable results across different random initializations of the algorithm. To address this issue, we ran the analysis many times (1,000 times), using different random initializations (activations and electrode weights were initialized with random samples from a truncated normal distribution; see **Fig S10** for the structure and initialization of the smoothing kernels). Components that are stable should be consistently present for all solutions with low cost, which we quantified by correlating the component response profiles for the solution with the lowest cost with those for the 99 next-best solutions (using the “Hungarian algorithm” to determine the correspondence between components from different solutions (Kuhn, 1955)). As a measure of stability, we computed the median correlation value for each component across the 99 next-best solutions (**Fig 1G**). The responses and weights shown are from the model with the lowest cost.

We ordered components based on their total contribution to explaining the data matrix, measured by summing the response timecourse and electrode weights for each component, and then multiplying them together:

$$(\sum_t \mathbf{r}_k(t))(\sum_i w_{i,k}) \quad (7)$$

Cross-validation analyses. We used cross-validated prediction accuracy to determine the number of components and the sparsity parameter (**Fig 1E & S2B**), as well as to compare the SSC model with several baseline models (**Fig S2C**). For the purposes of cross-validation, we separated the timecourses for different sounds into cells, thus creating an electrode x sound matrix (**Fig 1E**). We then trained the model on a random subset of 80% of cells and measured the model's prediction accuracy (squared Pearson correlation) in the left-out 20% of cells. We trained models starting from 10 different random initializations, and selected the model with the lowest error in the training data. We repeated our analyses using 5 different random splits of train and test data, averaging the test correlations across splits. For each split, we ensured an even and broad sampling of train and test stimuli using the following procedure: (1) we created a random ordering of stimuli and electrodes (2) we assigned the first 20% of sounds to be test sounds for the first electrode, the next 20% of sounds to be test sounds for electrodes 2, and so on. After using up all 165 sounds (every 8-9 electrodes), we refreshed the pool of available test sounds using a new random ordering of stimuli.

To prevent correlated noise across electrodes from influencing the results, we used non-overlapping sets of runs (odd and even runs) to compute the training and test data (i.e. training on odd runs and testing on even runs, and vice-versa; again averaging test correlations across the two splits). For a given set of hyper-parameters, we then averaged the test correlations across all electrodes to arrive at a summary measure of that model's performance (**Fig 1E & S2B**). We noise-corrected this measure by dividing it by the average test-retest correlation of the electrode responses (using the unsquared Pearson correlation), which gives an upper bound on the model's prediction accuracy (Norman-Haignere et al., 2015; Schoppe et al., 2016).

We considered several baseline models that did not use the convolutional decomposition of the SSC model (specifically, we constrained the smoothing kernel to be a delta function such that the component activations, $\boldsymbol{a}_k(t)$, equaled the component responses, $\boldsymbol{r}_k(t)$). We tested four baseline models: (1) we removed the sparseness and smoothness constraints entirely but maintained the non-negativity constraint (i.e. non-negative matrix factorization / NMF); (2) we imposed sparsity but not smoothness via an L1 penalty the component responses and weights (3) we imposed smoothness but not sparsity via an L2 smoothness penalty on the derivative of the component responses (the first-order difference of adjacent time-points); and (4) we applied both the L1 sparsity and L2 smoothness constraint. To prevent the number of hyper-parameters from biasing the results, for each electrode, we selected the hyper-parameters that led to the best performance across electrodes from other subjects (**Fig S2C**). We used grid-search over the following range of hyper-parameters: K (number of components) $= [5,10,15,20,25,30]$, λ (sparsity) $= [0, 0.033, 0.1, 0.33, 1, 3.3]$, ω (smoothness) $= [0, 0.033, 0.1, 0.33]$ (we verified that the best-performing models were not on the boundary of these values, except in cases where the best-performing model had a parameter value of 0). We found that all of the baseline models performed worse than the SSC model ($p < 0.001$ via bootstrapping across subjects, see *Statistics*; including the model with both an L1 sparsity and L2 smoothness penalty, which had more hyper-parameters). This result shows that our convolutional decomposition is an effective way of capturing both the smoothness and sparsity of auditory broadband gamma responses, and is more effective than simply imposing sparsity and smoothing penalties directly on the component responses.

Constraining the smoothing kernel. We investigated three potential methods for forcing the smoothing kernel to be smooth: (1) using a parametric kernel (e.g. Gamma distribution); (2) placing a smoothness penalty on the derivative of the kernel; and (3) constraining the kernel to be unimodal. We found that the optimizer had difficulty minimizing the loss when using parametric kernels (likely because the low-dimensional parameters of the kernel interacted in complex ways with the other high-dimensional parameters). We found that penalizing the derivative and constraining the kernel

to be unimodal were both effective (yielding similar cross-validated prediction accuracy), but penalizing the derivative introduces a third hyper-parameter that must be chosen with cross-validation, so we chose the unimodal constraint.

We constrained the kernel to be unimodal by placing two constraints on its derivative: (1) the first N points of the derivative must be positive and the remaining points must be negative (which forces the kernel to go up and then down, but not oscillate); and (2) the sum of the derivative must equal 0 (ensuring that the kernel starts and ends at zero). The set of operations used to implement these constraints in TensorFlow is described in **Fig S10**. Many of the learned smoothing kernels were asymmetric, with a rapid rise and a slower falloff (**Fig S9**). There is nothing in the constraints that encourages asymmetry, and so this property must reflect an asymmetry in the cortical responses themselves.

Specificity of components for individual subjects. The sparse and clinically-driven coverage of ECoG grids virtually guarantees that some response types will only be present in a subset of subjects. Thus, one might expect to find components that are subject-specific. To evaluate this possibility, we measured the average weight of each component in each subject, and then normalized these mean weights to sum to one across subjects (**Fig 1F**). Most components had substantial weights for multiple subjects, but for five of the 20 components, one subject accounted for more than half of the normalized subject weights (Components 14, 16, 18, 19, 20). We thus chose to focus on the components that were more general.

For the 15-component model (**Fig S4**), three components had normalized subject weights greater than 0.5 (one other component was omitted because it was not stable across random re-initializations of the algorithm). For component model constrained only by non-negativity (**Fig S3**), two components had normalized subject weights greater than 0.5, and three other components weighted strongly on a single electrode (with one electrode accounting for more than 25% of the total weights across all electrodes), and were thus excluded from the plots shown.

Component responses to modulation-matched sounds. The components were inferred using responses to just the 165 natural sounds from the main experiment. But since a subset of ten subjects were tested in both experiments, we could estimate the response of these same components to the natural and synthetic sounds from our control experiment. Specifically, we fixed the component electrode weights to the values inferred from the responses in our main experiment, and learned a new set of component response timecourses that best approximated the measured responses in the modulation-matching experiment. Since the electrode weights are known, this analysis is no longer ill-posed, and we thus removed all of the additional sparsity and smoothness constraints and simply estimated a set of non-negative response profiles that minimized the squared reconstruction error (we left the non-negativity constraint because we found that nearly all of the measured responses were non-negative).

Single electrode analyses. To identify electrodes selective for music, speech and song, we defined a number of contrasts based on the average response to different categories (the contrasts are described in the Results). We then divided each contrast by the maximum response across all categories to compute a measure of selectivity, or we bootstrapped the contrast to determine if it was significantly greater than zero (see *Statistics* below). In all cases, we used independent data to identify electrodes and measure their response. Specifically, we used two runs (first and last) to select electrodes and the remaining runs to evaluate their response.

Statistics. The significance of all category contrasts was evaluated using bootstrapping (Efron, 1982). Specifically, we sampled sounds from each category with replacement (100,000 times), averaged responses across the sampled sounds for each category, and then recomputed the contrast of interest (all of the contrasts tested are specified in the Results). We then counted the fraction of samples that fell below zero and multiplied by 2 to compute a two-sided p-value. For p-values smaller than 0.001, counting becomes unreliable, and so we instead fit the distribution of bootstrapped samples with a Gaussian and measured the tail probability that fell below zero (and multiplied by 2 to compute a two-sided p-value). For the component analyses, we corrected for multiple comparisons by multiplying these p-values by the number of components (corresponding to Bonferroni correction).

We compared the song-selective component (Component 12) with the average response of all song-selective electrodes by counting the fraction of bootstrapped samples where the component showed greater super-additive selectivity for vocal music (vocal music > max(English speech, foreign speech) + instrumental music). We found that across all 100,000 bootstrapped samples, the component always showed greater selectivity.

We also used bootstrapping to compute error bars for the category timecourses (**Fig 2A, Figs S3-5**). In these figures we plot the central 68% of the sampling distribution (equivalent to one standard error for a Gaussian distributed variable). We only plot categories for which there were more than 5 exemplars.

To test for laterality effects, we computed the mean difference in the component electrode weights between the right and left hemispheres (**Fig S6**). We then bootstrapped this difference score by sampling subjects with replacement, and recomputing the mean difference using only electrodes from the sampled subjects. We repeated this procedure 100,000 times, and computed a p-value by counting the fraction of samples falling below or above zero (whichever was smaller) and multiplying by 2. We again Bonferroni-corrected by simply multiplying the p-value by the number of components. Only one component (Component 17, which was offset-selective) was significant after correction (p = 0.032 after correction).

We also used bootstrapping across subjects to place error bars on model prediction scores. Specifically, we (1) sampled subjects with replacement (10,000 times); (2) averaged the test correlation values (squared Pearson correlation) across the electrodes from the sampled subjects; and (3) divided by the average test-retest correlation (unsquared Pearson correlation) of the sampled electrodes to noise-correct our measure. We tested whether the SSC model outperformed our baseline models by counting the fraction of bootstrapped samples where the average test predictions were lower than each baseline model and multiplying by 2 to arrive at a two-sided p-value. When plotting the test predictions for different models (**Fig S2C**), we used “within-subject” error bars (Loftus and Masson, 1994), computed by subtracting off the mean of each bootstrapped sample across all models before measuring the central 68% of the sampling distribution. We multiplied the central 68% interval by the correction factor shown below to account for a downward bias in the standard error induced by mean-subtraction (Loftus and Masson, 1994):

$$\sqrt{\frac{N}{N-1}} \qquad 8$$

We used a sign test to evaluate whether the response to natural sounds was consistently greater than responses to corresponding modulation-matched sounds. A sign test is natural choice, because

the natural and modulation matched sounds are organized as pairs (**Fig 3A**). For components selective for speech / music (song selective components described in the next paragraph), we compared the time-averaged response to natural speech / music with the corresponding modulation-matched controls (there were eight speech stimuli, eight instrumental music stimuli and two vocal music stimuli). We performed the same analysis on the average response of speech and music-selective electrodes (**Fig 4C**). For both components and electrodes, the response to natural sounds of the preferred category was always greater than the response to modulation-matched sound, and thus significant with a sign test ($p < 0.01$).

Although there were only two vocal music stimuli in the modulation-matching experiment, the stimuli were relatively long (4 seconds). We thus subdivided the response to each stimulus into seven 500 ms segments (discarding the first 500 ms to account for the build-up in the response), and measured the average response to each segment. For both the song-selective component and the average response of song-selective electrodes, we found that for all fourteen 500-ms segments (7 segments across 2 stimuli), the response to natural vocal music was higher than the response to the modulation-matched controls, and thus is significant with a sign test ($p < 0.001$).

To determine whether the electrode responses were significantly more skewed and sparse than would be expected given noise (i.e. to evaluate the significance of the skewness/sparsity measures described in *Response statistics relevant to component modeling*), we computed skewness/sparsity using two data quantities: (1) the residual error after subtracting the response to even and odd runs; and (2) the summed response across even and odd runs. The properties of the noise should be the same for these two quantities, but the second quantity will also contain the reliable stimulus-driven component of the response. Thus, if the second quantity is more skewed/sparse than the first quantity, then the stimulus-driven response must be more skewed/sparse than the noise. To assess skewness/sparsity across time/stimuli, we measured the skewness and sparsity (equations 2 and 3) separately for each electrode using the residual error and summed response (pooling responses across all timepoints and stimuli). In every subject, we found that the average skewness/sparsity of the summed responses was greater than the skewness/sparsity of the residual error, and thus significant with a sign test ($p < 0.001$). We used the same approach to evaluate the skewness/sparsity of responses across electrodes, measured separately for each sound. Using a sign test across sounds, we found both the skewness and sparsity of the summed response to be significantly greater than that for the residual error ($p < 0.001$).

Acknowledgements

This work was supported by the National Institutes of Health (EY13455 to N.G.K., P41-EB018783 to G.S., P50-MH109429 to G.S., R01-EB026439 to G.S., U24-NS109103 to G.S., U01-NS108916 to G.S., and R25-HD088157 to G.S.), the U.S. Army Research Office (W911NF-15-1-0440 to G.S.), the National Science Foundation (Grant BCS-1634050 to J.H.M.), the NSF Science and Technology Center for Brains, Minds, and Machines (CCF-1231216), Fondazione Neurone (Grant to G.S.), and the Howard Hughes Medical Institute (LSRF Postdoctoral Fellowship to S.N.H.).

Competing interests

Authors declare no competing financial and/or non-financial interests in relation to the work described in this paper.

References

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

Barlow HB (1961) Possible principles underlying the transformation of sensory messages. Sens Commun 1:217–234.

Barton B, Venezia JH, Saberi K, Hickok G, Brewer AA (2012) Orthogonal acoustic dimensions define auditory field maps in human cortex. Proc Natl Acad Sci 109:20738–20743.

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

Bouchard KE, Bujan AF, Chang EF, Sommer FT (2017) Sparse coding of ECoG signals identifies interpretable components for speech control in human sensorimotor cortex. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp 3636–3639.

Byron MY, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, Sahani M (2009) Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In: Advances in neural information processing systems, pp 1881–1888.

Casey M, Thompson J, Kang O, Raizada R, Wheatley T (2012) Population codes representing musical timbre for high-level fMRI categorization of music genres. In: Machine Learning and Interpretation in Neuroimaging, pp 34–41. Springer.

Casey MA (2017) Music of the 7Ts: Predicting and decoding multivoxel fMRI responses with acoustic, schematic, and categorical Music Features. Front Psychol 8:1179.

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

Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D (2017) Temporal modulations in speech and music. Neurosci Biobehav Rev.

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

Fairhall AL, Lewen GD, Bialek W, de Ruyter van Steveninck RR (2001) Efficiency and ambiguity in an adaptive neural code. Nature 412:787–792.

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

Hamilton LS, Edwards E, Chang EF (2018) A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. Curr Biol 28:1860-1871.e4.

Heilbron M, Chait M (2017) Great expectations: is there evidence for predictive coding in auditory cortex? Neuroscience.

Herholz SC, Zatorre RJ (2012) Musical training as a framework for brain plasticity: behavior, function, and structure. Neuron 76:486–502.

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

Kell AJ, Yamins DL, Shook EN, Norman-Haignere SV, McDermott JH (2018) A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron.

Kuhn HW (1955) The Hungarian method for the assignment problem. Nav Res Logist Q 2:83–97.

Kvale MN, Schreiner CE (2004) Short-term adaptation of auditory receptive fields to dynamic stimuli. J Neurophysiol 91:604–612.

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

Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791.

Loftus GR, Masson ME (1994) Using confidence intervals in within-subject designs. Psychon Bull Rev 1:476–490.

Lomax A (2017) Folk song style and culture. Routledge.

Mehr S, Singh M, Knox D, Ketter D, Pickens-Jones D, Atwood S, Lucas C, Egner A, Jacoby N, Hopkins EJ (2018) A natural history of song.

Mehr SA, Krasnow MM (2017) Parent-offspring conflict and the evolution of infant-directed song. Evol Hum Behav 38:674–684.

Merrill J, Sammler D, Bangert M, Goldhahn D, Lohmann G, Turner R, Friederici AD (2012) Perception of words and pitch patterns in song and speech. Front Psychol 3.

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

Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1–25.

Norman-Haignere SV, Albouy P, Caclin A, McDermott JH, Kanwisher NG, Tillmann B (2016) Pitch-responsive cortical regions in congenital amusia. J Neurosci.

Norman-Haignere SV, Kanwisher NG, McDermott JH (2015) Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88:1281–1296.

Norman-Haignere SV, McDermott JH (2018) Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. PLoS Biol 16:e2005127.

Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Res 37:3311–3325.

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

Patel AD (2012) Language, music, and the brain: a resource-sharing framework. Lang Music Cogn Syst:204–223.

Patel AD (2019) Evolutionary music cognition: Cross-species studies. In: Foundations in Music Psychology: Theory and Research, pp 459–501.

Peretz I (2016) Neurobiology of congenital amusia. Trends Cogn Sci 20:857–867.

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

Peterson RL, Pennington BF (2015) Developmental dyslexia. Annu Rev Clin Psychol 11:283–307.

Ray S, Maunsell JHR (2011) Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLOS Biol 9:e1000610.

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

Savage PE, Brown S, Sakai E, Currie TE (2015) Statistical universals reveal the structures and functions of human music. Proc Natl Acad Sci 112:8987–8992.

Schindler A, Herdener M, Bartels A (2013) Coding of Melodic Gestalt in Human Auditory Cortex. Cereb Cortex 23:2987–2993.

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

Schoppe O, Harper NS, Willmore BD, King AJ, Schnupp JW (2016) Measuring the performance of neural models. Front Comput Neurosci 10:10.

Scott SK, Blank CC, Rosen S, Wise RJ (2000) Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123:2400–2406.

Singh NC, Theunissen FE (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. J Acoust Soc Am 114:3394–3411.

Steinschneider M, Fishman YI, Arezzo JC (2008) Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. Cereb Cortex 18:610–625.

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

Wallin NL, Merker B, Brown S (2001) The origins of music. MIT press.

Whittingstall K, Logothetis NK (2009) Frequency-band coupling in surface EEG reflects spiking activity in monkey visual cortex. Neuron 64:281–289.

Williams AH, Kim TH, Wang F, Vyas S, Ryu SI, Shenoy KV, Schnitzer M, Kolda TG, Ganguli S (2018) Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis. Neuron 98:1099–1115.

Wiskott L, Sejnowski TJ (2002) Slow feature analysis: Unsupervised learning of invariances. Neural Comput 14:715–770.

Supplemental Figures

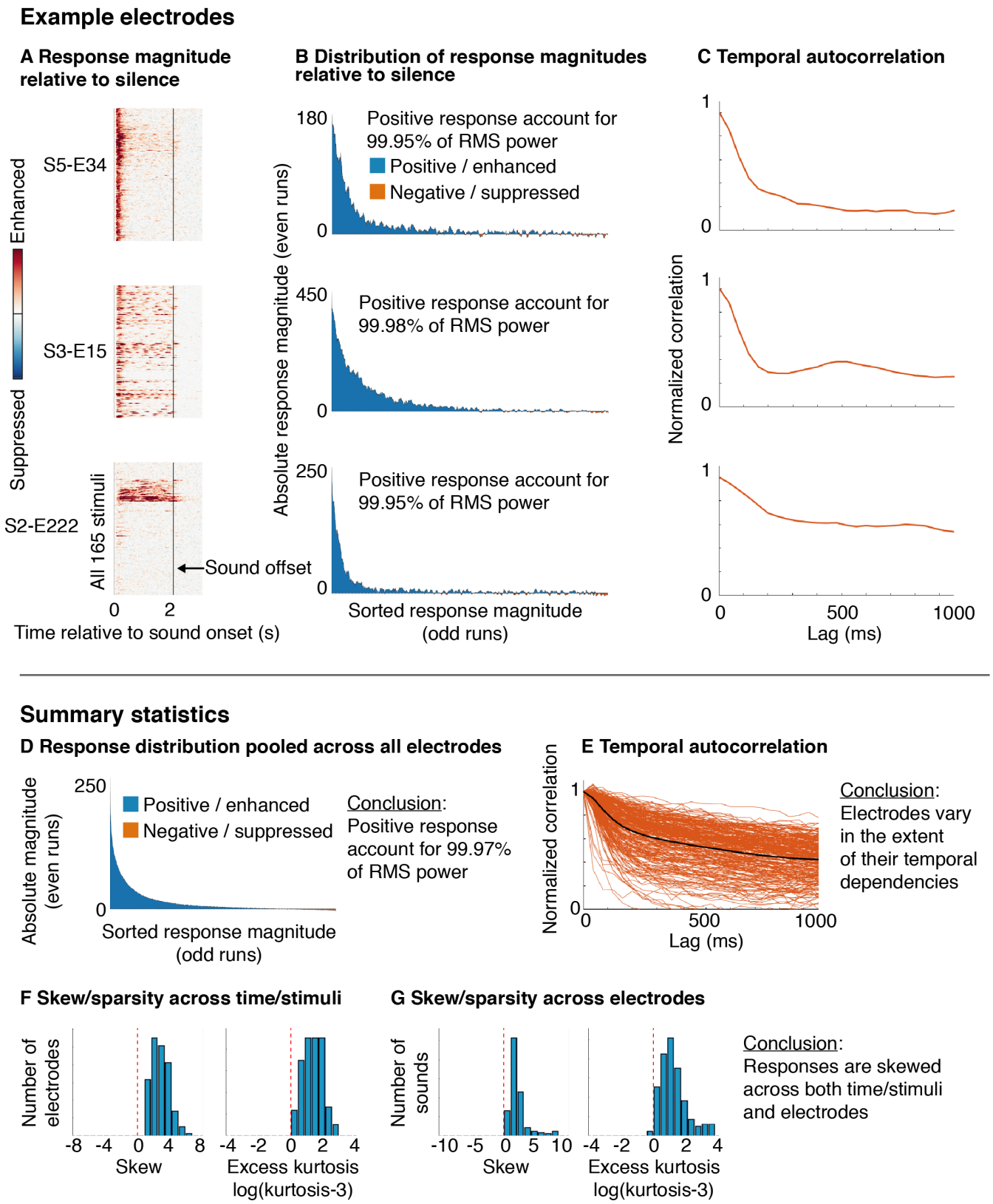

Figure S1. **Response statistics relevant to component modeling. A-C,** Response statistics from three example electrodes with distinct selectivities, but a shared set of statistical properties (positivity, sparsity/skew, and temporal smoothness). **A**, Broadband gamma power response of each electrode to all 165 sounds as a raster. Responses are measured relative to the response during silence (300 milliseconds preceding sound onset). Positive values (red)

indicate an enhanced response to sound, and negative responses indicate a suppressed response (blue). The color scales shows values from 0 to the 99th percentile of the response magnitude distribution for each electrode. **B,** Distribution of response magnitudes, measured in a cross-validated fashion to reduce effects of noise: using data from the odd runs, we sorted all of the bins of the raster on the left based on their magnitude (pooling across all timepoints and stimuli). The response of each bin was then measured using the even runs, and then smoothed using a median filter to suppress noise. Positive responses accounted for >99% of the RMS response power in all three electrodes. All three electrodes show a skewed and sparse distribution of response magnitudes (quantified in panel F, below) because negative responses were practically non-existent (yielding an asymmetric, rightward-skewed distribution) and strong positive responses were present for only a small fraction of bins (yielding a sparse distribution). **C,** The normalized autocorrelation (normalized by the correlation at zero lag) of each electrode's response measured in a cross-validated fashion by correlating the response in odd and even runs at different lags. **D-G,** Summary statistics across all sound-responsive electrodes. **D,** Distribution of response magnitude pooled across all electrodes, sounds and timepoints (measured in a cross-validated fashion, as described above). Positive responses accounted for >99% of the RMS power. **E,** Normalized autocorrelation of all sound-responsive electrodes. The extent of temporal dependencies varied substantially across electrodes. **F,** We measured the skew (3rd moment) and sparsity (excess kurtosis) of each electrode's response using its distribution of response magnitudes across all timepoints/stimuli (i.e. using the distributions shown in panel B). This figure plots a histogram of the skew and sparsity values across all electrodes. We subtracted the measured kurtosis from that which would be expected from a Gaussian (which has a kurtosis of 3). All electrodes were skewed and sparse relative to a Gaussian. **G,** For each sound, we measured the skew and sparsity of responses across electrodes, after averaging the response of each electrode to each sound. This figure plots a histogram of the skew and sparsity values across all sounds.

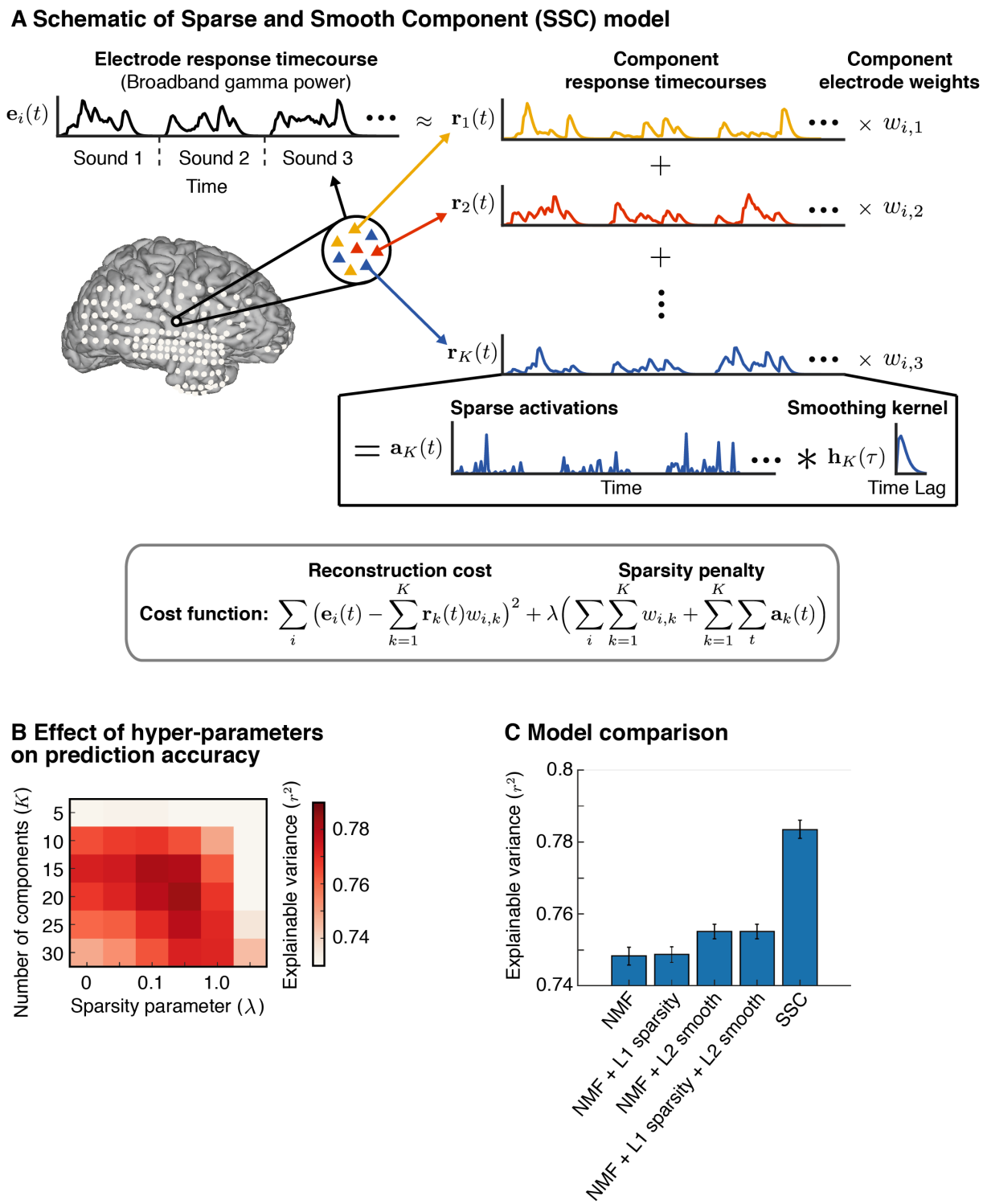

Figure S2. Component model and its evaluation via cross-validation. A, Schematic of the "sparse and smooth" component model, which was motivated by the statistical properties shown in **Fig S1**. Each electrode was represented by its response timecourse (broadband gamma) across all sounds (measured relative to silence). This timecourse was modeled as the weighted sum of multiple component timecourses to capture the fact that each electrode is influenced by many neurons and thus might reflect multiple underlying neuronal populations. The component response timecourses were the same across electrodes, but the weights varied to account for different response patterns. Both the component responses and weights were constrained to be positive. To encourage the component response patterns to be sparse and skewed, we modeled each component as the convolution of a set of sparse activations with a smoothing kernel. The activations, weights and smoothing kernel were all learned by minimizing a cost function with two terms: (1) a reconstruction penalty encouraging the components to closely approximate the data; and (2) a sparsity penalty encouraging the activations and weights to be sparse. The smoothing kernel was learned separately for each component to account for variable levels of smoothness in the responses across electrodes. **B,** Average squared correlation between

the measured and model-predicted response in test data as a function of the number of components and sparsity penalty (the correlation has been noise-corrected; **Fig 1E** shows results for the best sparsity parameter ($\lambda = 0.33$)). **C,** Comparison of the prediction accuracy (average correlation in test data) of the SSC model with several baseline models that did not rely on the convolutional decomposition used by the SSC model: (1) non-negative matrix factorization (NMF) where the components and weights were constrained only to be positive; (2) NMF with a sparsity penalty applied directly to the responses and weights; (3) NMF with a L2 smoothness penalty applied to the derivative (first-order difference) of the component responses; and (4) NMF with both an L1 sparsity and L2 smoothness penalty. Data from independent subjects was used to select the hyper-parameters for each model and evaluate prediction accuracy. Error bars show the median and central 68 percent of the sampling distribution measured via bootstrapping across subjects.

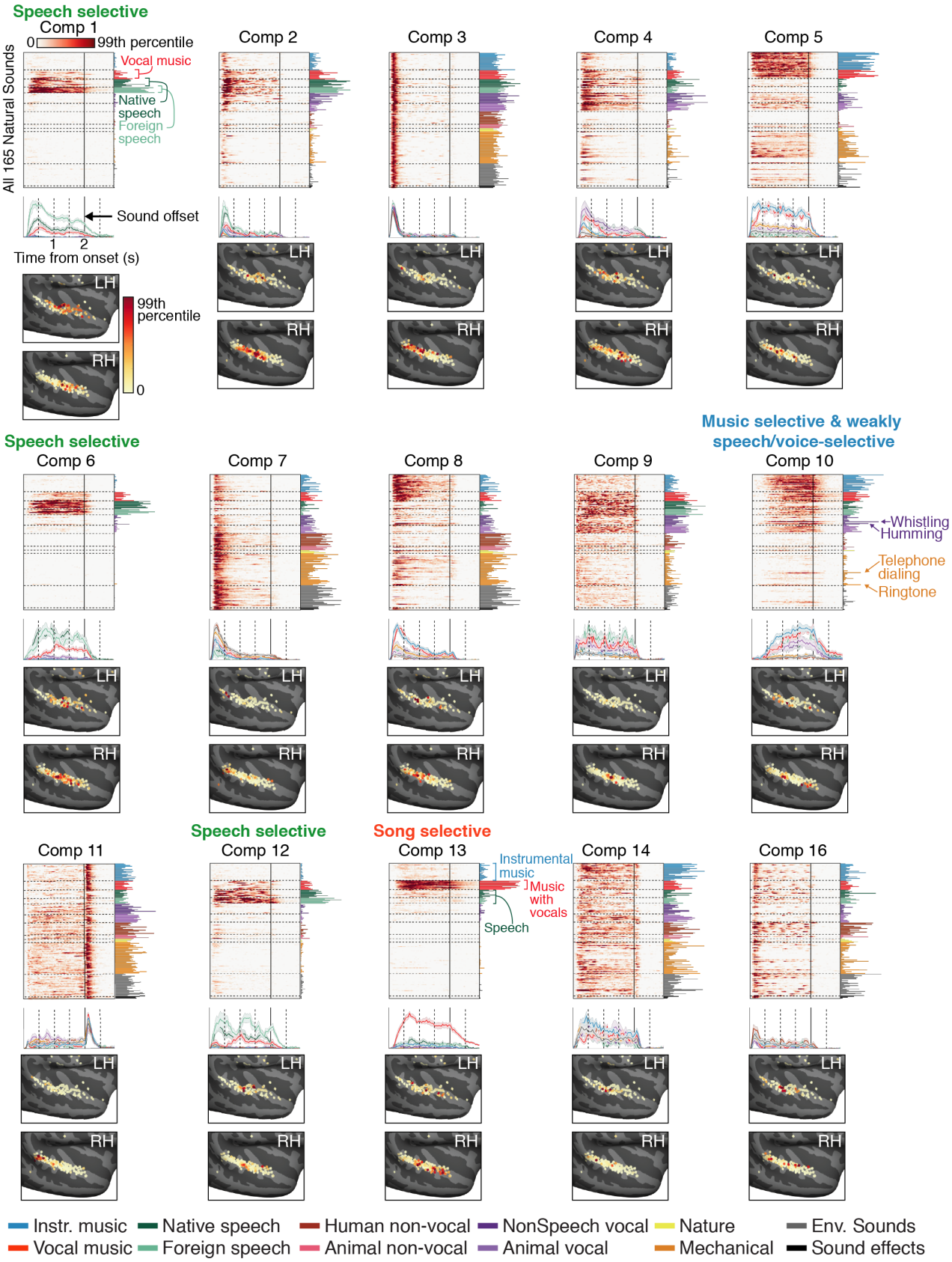

Figure S3. Components from non-negative matrix factorization (NMF) model. Component responses and weights from a model that only imposed non-negativity on the responses/weights. Conventions the same as **Fig 2&S5** which show components from the SSC model (which had the best prediction accuracy). As with the SSC model, we focus on components that were consistent across subjects and reliable across random re-initializations of the algorithm. All of the speech, music and song-selective components inferred from the SSC model have clear analogues to those inferred by NMF.

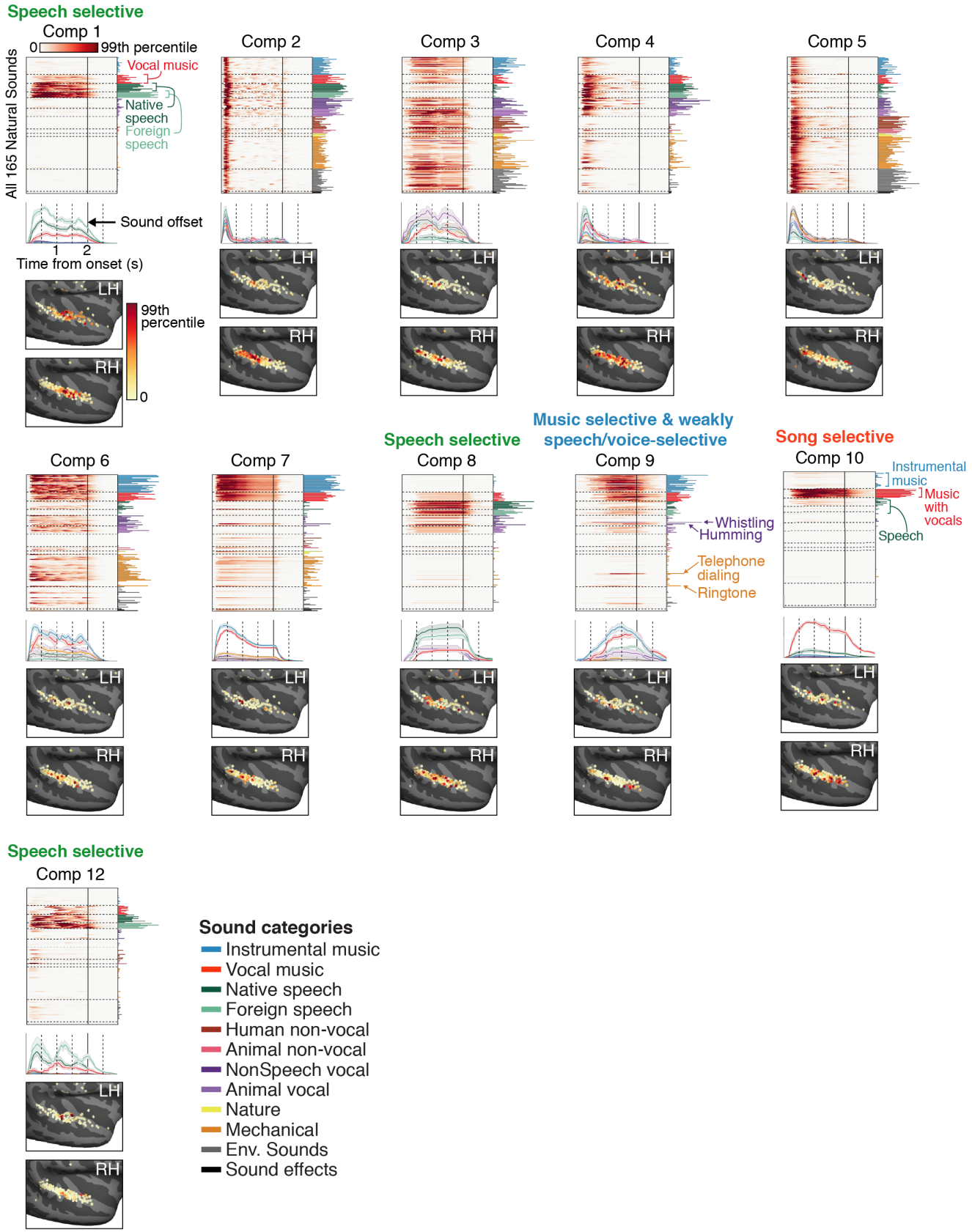

Figure S4. Results from 15-component model. Component responses and weights from a model with only 15 components. Conventions the same as **Fig 2&S5**, which show results from a 20-component model. We focus on components that were consistent across subjects and reliable across random re-initializations of the algorithm. All of the speech, music and song-selective components inferred from the 20-component model were evident in the 15-component model.

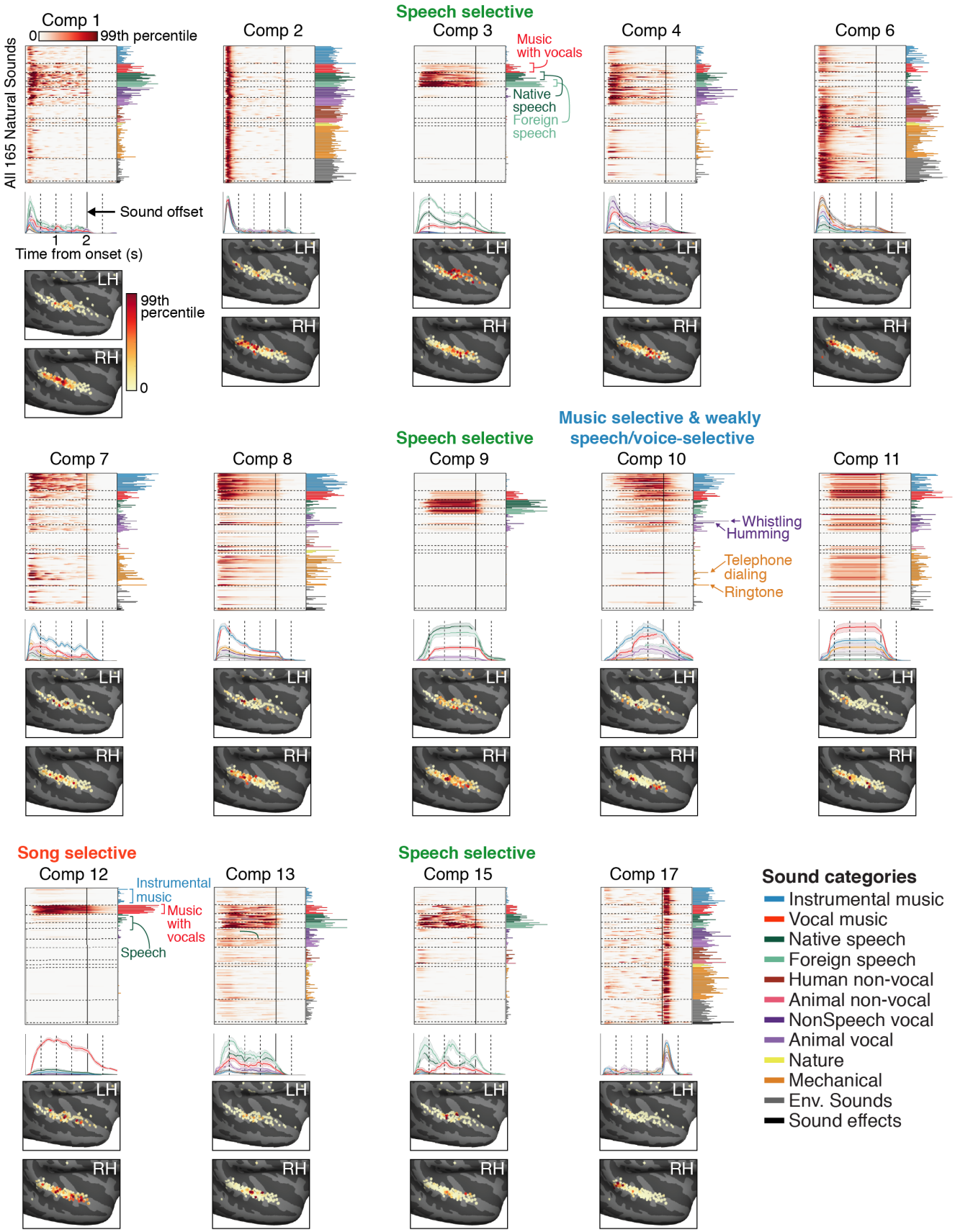

Figure S5. All reliable components from 20-component SSC model. This figure is the same as **Fig 2**, but shows component responses and weights from all of the reliable components rather than just the speech, music and song-selective components. Conventions the same as **Fig 2**.

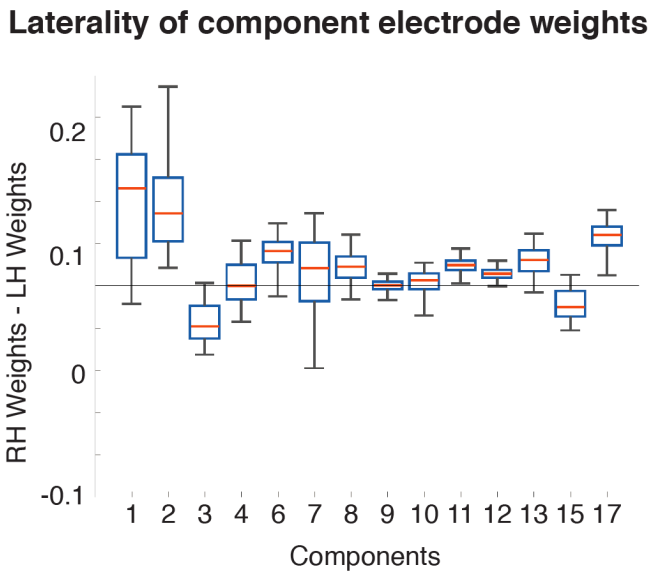

Figure S6. Laterality of component electrode weights. For each reliable component, we plot the average difference in the electrode weights between the right and left hemisphere. Bootstrapping across subjects was used to estimate the sampling distribution for each component. Boxes show the central 50% of the sampling distribution and whiskers show the central 95%.

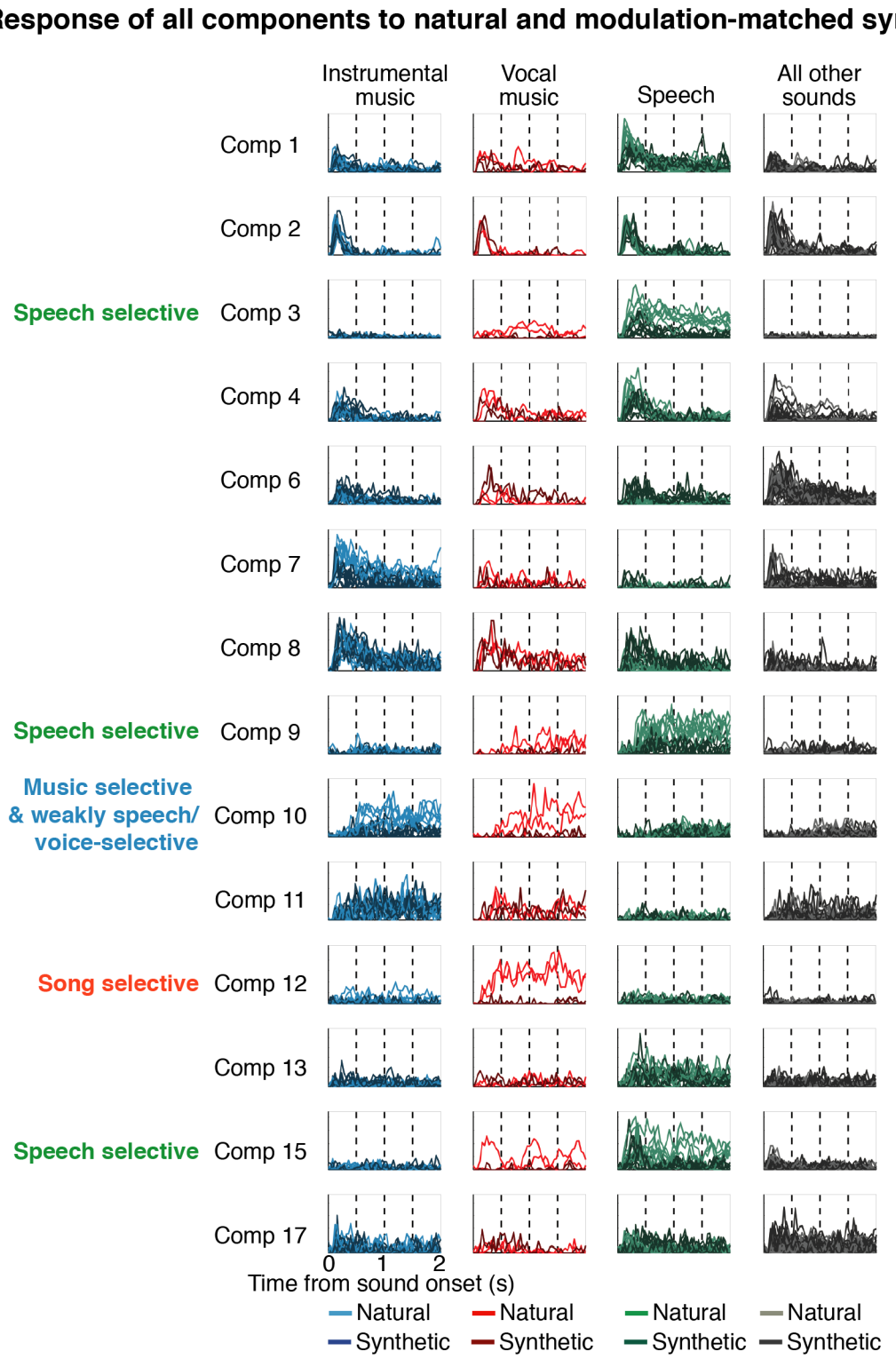

Figure S7. Response timecourse of all components to natural and modulation-matched synthetic sounds. Same as Fig 3B but showing responses from all components rather than just those selective for speech, music and song.

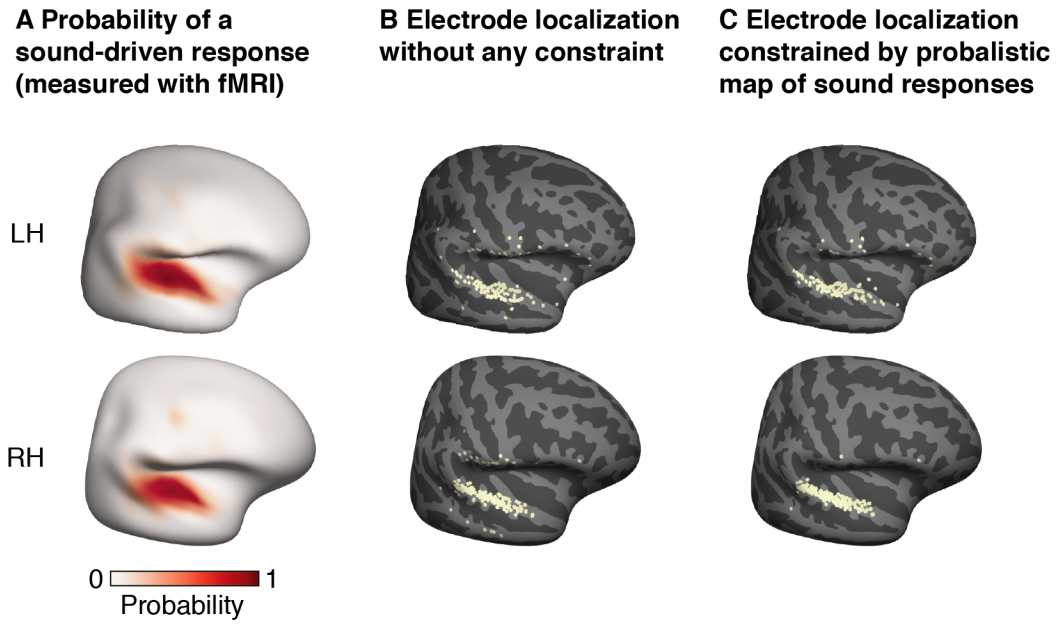

Figure S8. Constraining the anatomical localization of electrodes. A, Map showing the probability of observing a significant response to sound at each point in the brain. The map was computed using fMRI responses to the same sound set in a large cohort of 20 subjects. **B,** Electrode localization based purely on anatomical criteria. Small errors in localization likely explain why some electrodes have been localized to the middle temporal gyrus and supramarginal/inferior frontal gyrus, which abut the superior temporal gyrus where responses to sound are common. **C,** To minimize gross localization errors, we treated the probability map of sound-driven responses shown in panel A as a prior and used to it constrain the localization (see *Electrode localization* in the Methods). Our approach did not substantially affect the localization of electrodes at a fine scale, but encouraged electrodes to be mapped to the superior temporal gyrus rather than the middle temporal or supramarginal/inferior frontal gyrus.

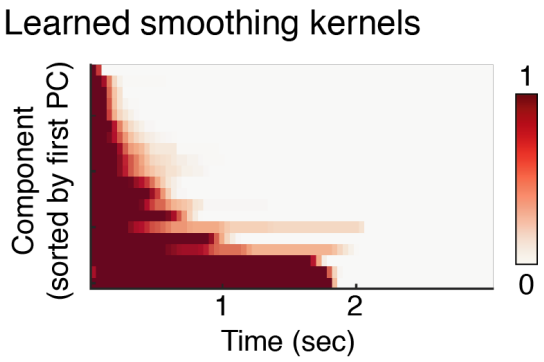

Figure S9. Learned smoothing kernels. This figure plots the learned smoothing kernels as a raster, with each row corresponding to a different kernel. The kernels have been sorted by the first principal component of the matrix. The kernels vary widely in their extent/duration. Many of the kernels are were also asymmetric with a fast/instantaneous rise and a slower falloff.

Constraining the smoothing kernel to be unimodal

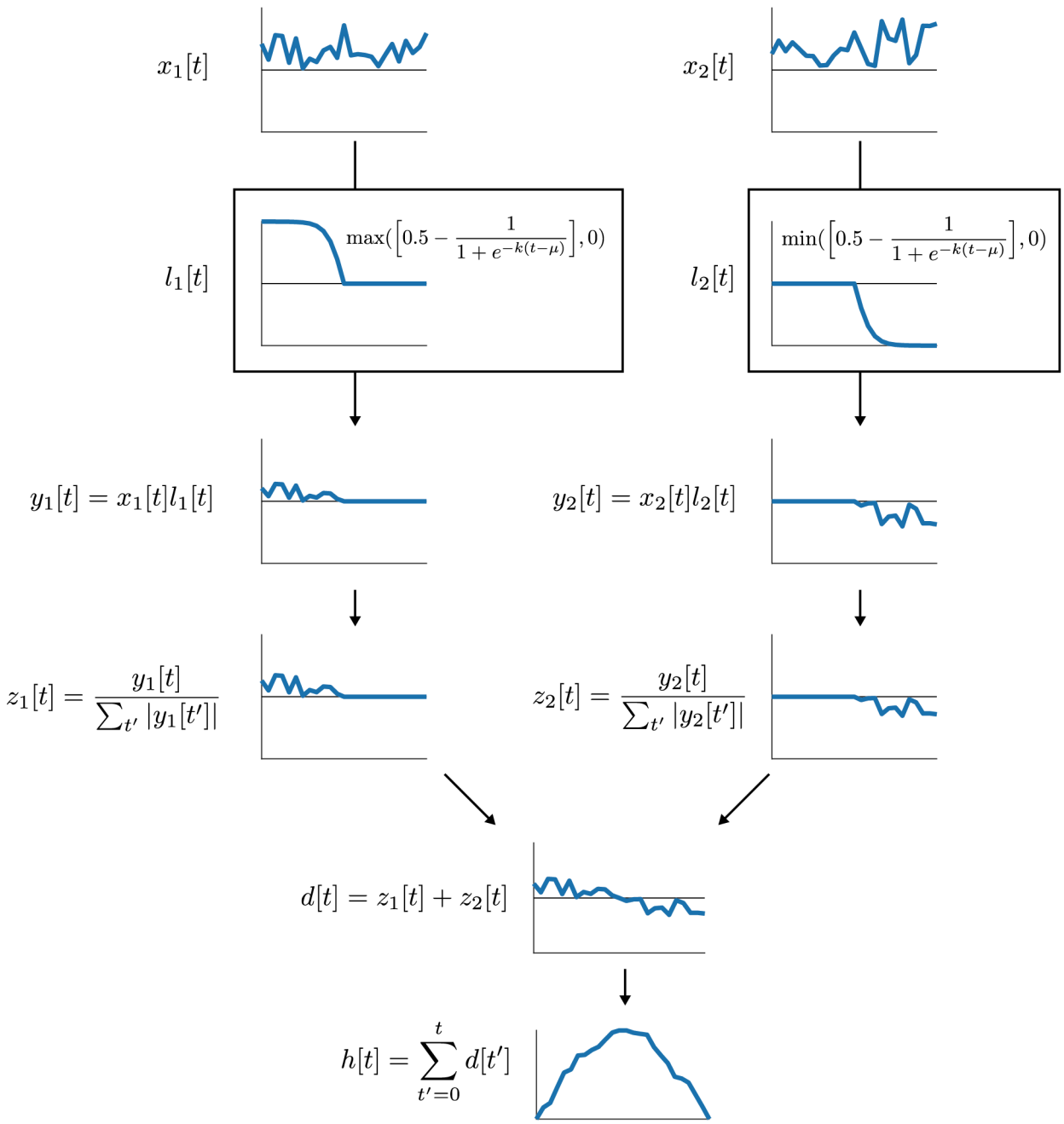

Figure S10. **Constraining the smoothing kernel to be unimodal.** This plot describes the set of operations (implemented in TensorFlow) that was used to constrain the smoothing kernel to be unimodal. Conceptually, the goal of these operations is to force the derivative to be exclusively positive for the first N time-points and then exclusively negative for the rest of the signal, thus preventing oscillations. We also must force the sum of the derivative to equal zero so that the kernel starts and ends at zero. Two positive vectors (themselves computed as the absolute value of real-valued vectors) were multiplied by a positively or negatively rectified logistic function with the same cross-over point. As a consequence, the first vector has positive values at the start of the signal, followed by zeros, and the second vector has negative values at the end of the signal, preceded by zeros. The two vectors were then normalized so that they sum to 1/-1. Finally, the two vectors were added and cumulatively summed, yielding a unimodal signal. The shape of the kernel is determined by the values of the two input vectors (x_1 and x_2) as well as the parameters of the logistic function (μ and k), all of which were learned. The input vectors were initialized with a vector of ones. μ was initialized to the value of the middle timepoint, and k was initialized to the value of 1 (and prevented from taking a value less than 0.001).