

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

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

6  
7 <sup>1</sup>Zuckerman Institute, Columbia University

8 <sup>2</sup>HHMI Fellow of the Life Sciences Research Foundation

9 <sup>3</sup>Laboratoire des Systèmes Perceptifs, Département d'Études Cognitives, ENS, PSL University,  
10 CNRS, Paris France

11 <sup>4</sup>Department of Brain and Cognitive Sciences, MIT

12 <sup>5</sup>Department of Neurology, Albany Medical College, Albany, NY

13 <sup>6</sup>National Center for Adaptive Neurotechnologies, Wadsworth Center, Albany, NY

14 <sup>7</sup>Department of Neurology, Mayo Clinic, Jacksonville, Florida, United States of America

15 <sup>8</sup>Program in Speech and Hearing Biosciences and Technology, Harvard University, Cambridge,  
16 Massachusetts, United States of America

17 <sup>9</sup>McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge,  
18 Massachusetts, United States of America

19 <sup>10</sup>Center for Brains, Minds and Machines, Cambridge, Massachusetts, United States of America

20 <sup>11</sup>Department of Biomed. Sciences, State University of New York

21

22 \*Corresponding author: [sn2776@columbia.edu](mailto:sn2776@columbia.edu)

23 **Abstract**

24

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

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

44  
45 Neuroimaging studies have suggested that representations of music diverge from those of other  
46 sound categories in non-primary human auditory cortex: some non-primary voxels show partial  
47 selectivity for music compared with other categories (Leaver and Rauschecker, 2010; Fedorenko et  
48 al., 2012; Angulo-Perkins et al., 2014), and a recent study from our lab, which modeled voxels as  
49 weighted sums of multiple response profiles, inferred a component of the fMRI response with clear  
50 selectivity for music (Norman-Haignere et al., 2015). However, there are few reports of finer-grained  
51 organization within the domain of music (Casey, 2017), potentially due to the coarse resolution of  
52 fMRI. As a consequence, we know little about the neural code for music.

53  
54 Here, we tested for finer-grained selectivity for music using intracranial recordings from the human  
55 brain (electrocorticography or ECoG) (**Fig 1A**). We measured ECoG responses to a diverse set of  
56 165 natural sounds, and submitted these responses to a novel decomposition method that is well-  
57 suited to the statistical structure of ECoG to reveal dominant response components of auditory  
58 cortex. This data-driven method revealed multiple music- and speech-selective response  
59 components. Our key finding is that one of these components responded nearly exclusively to music  
60 with vocals, suggesting the existence of neural populations that are selective for singing. We then  
61 used model-based sound synthesis (Norman-Haignere and McDermott, 2018) to show that these  
62 components could not be explained by generic acoustic representations often used to model cortical  
63 responses. Finally, we demonstrate direct evidence for music, speech, and song selectivity in  
64 individual electrodes without component modeling or statistical assumptions.

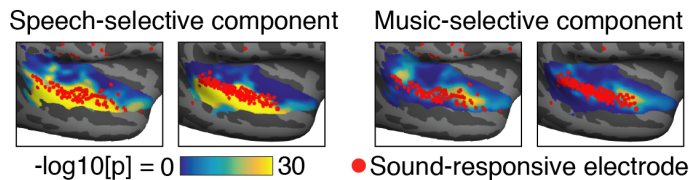
## 65 66 **Results**

67  
68 **Electrode decomposition.** We recorded ECoG responses from thirteen patients undergoing  
69 surgery for intractable epilepsy. We identified a set of 271 electrodes across all subjects with reliable  
70 broadband gamma (70-140 Hz) power responses to the sound set (split-half correlation > 0.2) (**Fig**  
71 **1B** plots the split-half correlation for all electrodes). We focused on broadband gamma, because it  
72 is thought to reflect aggregate spiking in a local region (Steinschneider et al., 2008; Whittingstall and  
73 Logothetis, 2009; Ray and Maunsell, 2011). Sound-responsive electrodes were nearly always  
74 located near the superior temporal gyrus (STG). Based on prior work, we expected speech selectivity  
75 to be prominent in the STG (**Fig 1C**) (Mesgarani et al., 2014; Norman-Haignere et al., 2015; Overath  
76 et al., 2015). By contrast, music selectivity is strongest in the lateral sulcus (**Fig 1C**) (Leaver and  
77 Rauschecker, 2010; Angulo-Perkins et al., 2014; Norman-Haignere et al., 2015), whose activity  
78 cannot be detected with surface electrodes. Thus, we expected music-selective electrodes, if  
79 present at all, to be relatively rare.

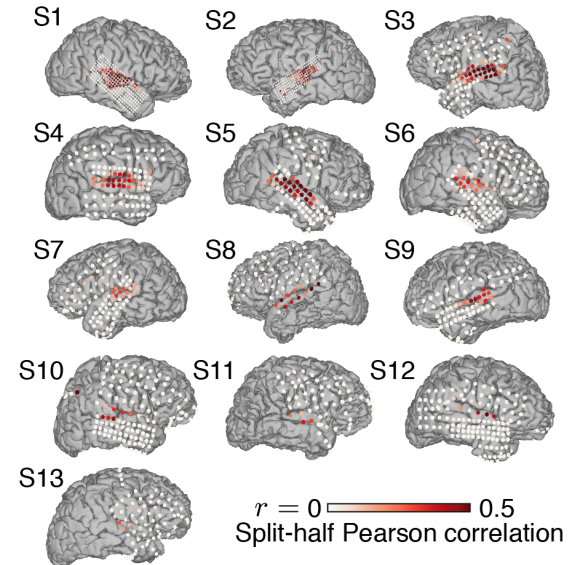
## A 165 natural sounds tested

- |                         |                     |                         |
|-------------------------|---------------------|-------------------------|
| 1. Man speaking         | 11. Running water   | 21. Cellphone vibrating |
| 2. Flushing toilet      | 12. Breathing       | 22. Water dripping      |
| 3. Pouring liquid       | 13. Keys jangling   | 23. Scratching          |
| 4. Tooth-brushing       | 14. Dishes clanking | 24. Car windows         |
| 5. Woman speaking       | 15. Ringtone        | 25. Telephone ringing   |
| 6. Car accelerating     | 16. Microwave       | 26. Chopping food       |
| 7. Biting and chewing   | 17. Dog barking     | 27. Telephone dialing   |
| 8. Laughing             | 18. Walking         | 28. Girl speaking       |
| 9. Typing               | 19. Road traffic    | 29. Car horn            |
| 10. Car engine starting | 20. Zipper          | ...                     |

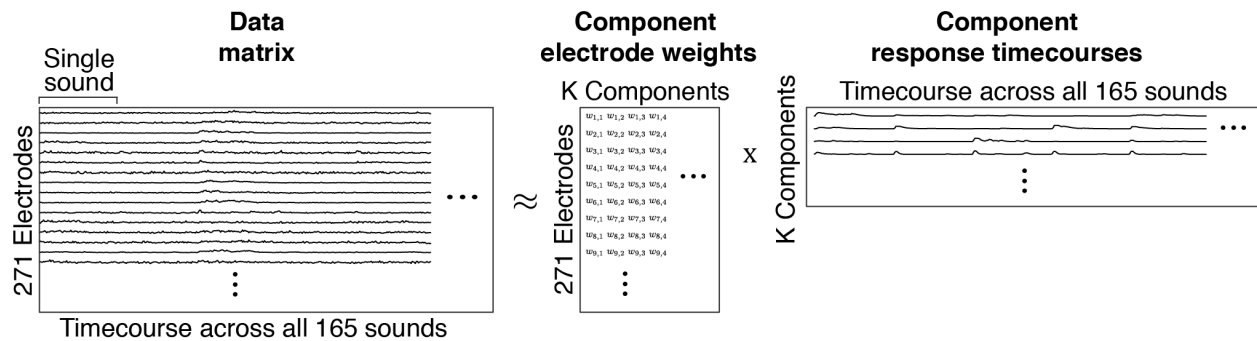
## C Electrode coverage relative to speech and music selectivity (measured by fMRI)



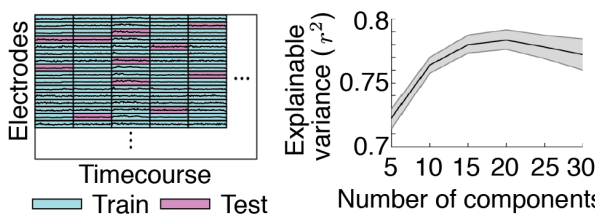
## B Reliability of ECoG broadband gamma response to natural sounds



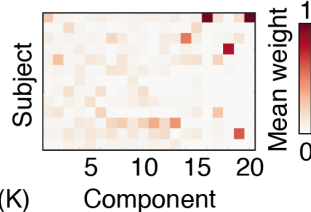
## D Electrode decomposition



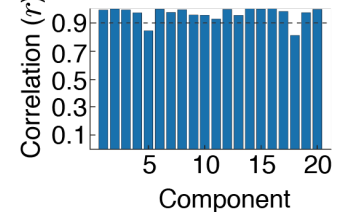
## E Prediction accuracy vs. number of components



## F Subject-specificity of component weights



## G Consistency across initialization



**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

80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93



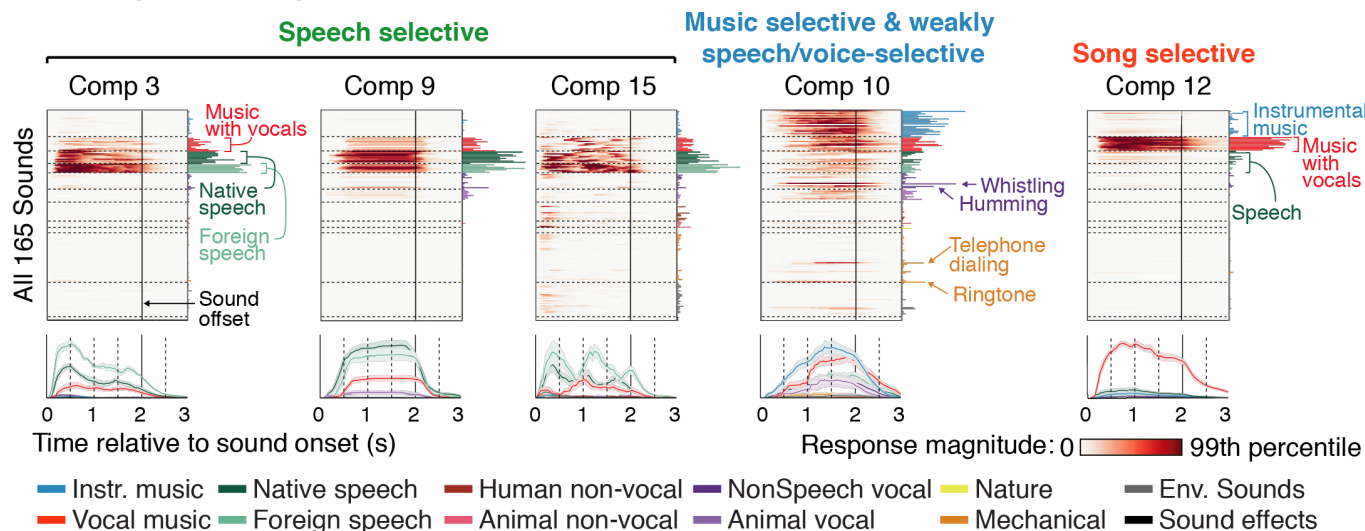
94 contribution of each component to each electrode, and a response matrix containing the response timecourse  
95 of each component to the sound set. **E**, Cross-validation was used to compare models (**Fig S2C**) and determine  
96 the number of components. The data matrix was divided into cells,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

110  
111 We sought to identify a small number of response timecourses across the sound set (components)  
112 that when weighted together could explain much of the response variance across all 271 electrodes.  
113 Each component timecourse could potentially reflect the response of a different neuronal  
114 subpopulation in auditory cortex, with the weights providing an estimate for the contribution of each  
115 subpopulation to each electrode. To identify components, we represented the response of all 271  
116 electrodes in a matrix, in which each row represented the response timecourse of a single electrode  
117 across all 165 sounds (**Fig 1D**). We then tried to approximate this matrix as the product of a  
118 component response timecourse matrix and a component electrode weight matrix.

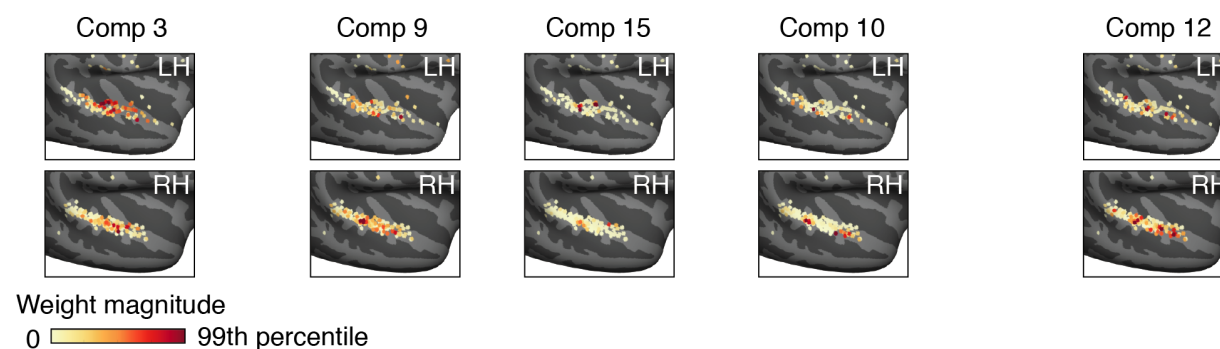
119  
120 In general, the problem of matrix factorization – finding a set of response components whose  
121 weighted sum best explains the data – is ill-posed and needs to be constrained by additional  
122 statistical criteria. We identified three statistical properties of auditory broadband gamma activity that  
123 are relevant to component modeling (**Fig S1**): (1) broadband gamma responses to sounds are nearly  
124 always larger than those to silence (smaller relative responses to sound accounted for <1% of the  
125 response power); (2) responses are sparse across both time/stimuli and space/electrodes; (3)  
126 responses are temporally smooth, and the extent of this smoothness varies across electrodes. We  
127 designed a model that captured all of these statistical properties by convolving a set of sparse/non-  
128 negative components with a learned smoothing kernel (**Fig S2**; see Methods for details). We focus  
129 on the results of this model because it yielded better prediction accuracy in held-out data than  
130 competing models (**Fig S2C**). But we note that our key results were evident using a model that only  
131 imposed non-negativity on the responses and weights (**Fig S3**), and were also evident in individual  
132 electrodes without using any component modeling (see *Single-electrode analyses* below).

133  
134 Using a simple cross-validation procedure, in which we trained and tested on separate  
135 sounds/electrodes, we found that we could estimate ~15-20 components before overfitting (**Fig 1E**).  
136 We show results from a model with 20 components, though all of the speech, music, and song-  
137 selective components were evident in a 15-component model (**Fig S4**). Collectively, the 20  
138 components inferred by the model accounted for approximately 78% of the explainable response  
139 variation (i.e. the variation that was reliable across repeated presentations). Of these 20  
140 components, fourteen explained responses across multiple subjects (rather than primarily weighting  
141 on just a single subject; **Fig 1F**) and were stable across random initializations of the algorithm (**Fig**  
142 **1G**). We focused on these fourteen components since they are more likely to reflect consistent  
143 features of auditory cortical responses.

## A Component responses



## B Component electrode weights



144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169

**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 weights 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

170 intermediate response, presumably due to the presence of speech structure (e.g. phonemes, words).  
171 The response to English and foreign speech was similar in these components, suggesting a  
172 response to speech acoustics rather than linguistic meaning, consistent with prior studies of speech  
173 selectivity in the STG (Mesgarani et al., 2014; Norman-Haignere et al., 2015; Overath et al., 2015)  
174 (all of the subjects were native English speakers; the response to foreign speech was higher in  
175 Components 3 & 15, plausibly because the foreign speech was spoken at faster rate and thus had  
176 more speech content). Speech selectivity in Components 3 & 15 developed within a few hundred  
177 milliseconds, while Component 9 showed a slower response. The speech-selective components  
178 clustered in the middle STG, as expected (Scott et al., 2000; Mesgarani et al., 2014; Overath et al.,  
179 2015). The weights for Components 3 & 15 were stronger on average in the left hemisphere, but this  
180 effect did not reach significance ( $p > 0.07$ , uncorrected for multiple components), consistent with  
181 prior fMRI studies showing bilateral speech selectivity (Norman-Haignere et al., 2015; Overath et al.,  
182 2015) (**Fig S6** plots the mean difference in electrodes weights between right and left hemisphere for  
183 all components).

184  
185 Two components exhibited selectivity for music (Component 10 & 12). Component 10 responded  
186 strongly to both instrumental and vocal music (average[Instrumental music, vocal music] >  
187 average[all non-music categories]:  $p < 0.001$  via bootstrapping, Bonferroni-corrected), and produced  
188 an intermediate response to speech, suggesting that music and speech were not perfectly  
189 disentangled by our component analysis (perhaps due to limited coverage of the lateral sulcus where  
190 music selectivity is prominent; **Fig 1C**). All other non-music and non-speech sounds produced weak  
191 responses in these components. Moreover, the response of Component 10 was considerably slower  
192 than many of the other components, with music selectivity taking nearly a second to build up,  
193 suggesting selectivity for longer-term temporal structure.

194  
195 Component 12 responded nearly exclusively to music with vocals: every single vocal music stimulus  
196 produced a high response and all other sounds, including both speech and instrumental music,  
197 produced a weak response. As a consequence, the response to vocal music was significantly higher  
198 than the summed response to speech and instrumental music, suggesting nonlinear selectivity for  
199 song (vocal music > max[English speech, foreign speech] + instrumental music:  $p < 0.001$  via  
200 bootstrapping, Bonferroni-corrected). This finding of nonlinear selectivity for vocal music is  
201 strengthened by the fact that our decomposition method explicitly models each electrode as a  
202 weighted sum of multiple components, and thus if song selectivity simply reflected a sum of speech  
203 and music selectivity, the model should not have needed a separate component selective for just  
204 vocal music.

205  
206 Unlike most other components, Components 10 (music selective) and 12 (song selective) showed  
207 high weights for electrodes in anterior auditory cortex, similar to what would be expected based on  
208 prior work (Leaver and Rauschecker, 2010; Angulo-Perkins et al., 2014; Norman-Haignere et al.,  
209 2015). There were also electrodes in middle/posterior STG with substantial weight for these  
210 components, which has also been observed with fMRI (Norman-Haignere et al., 2015), though less  
211 prominently than the anterior region of music selectivity.

212  
213 Many components did not exhibit clear selectivity for categories (**Fig S5**). Some components showed  
214 strong responses at the onset (Components 1, 2, 4, 6, 7, 8) or offset (Component 17) of sound,  
215 although the strength of this onset response varied across stimuli for several components. Most of  
216 these onset/offset selective components had weights that were clustered in the middle or posterior  
217 STG, but rarely in the anterior STG, consistent with a recent study (Hamilton et al., 2018). Several  
218 components were weakly selective for music or speech (Component 7, 8, 13), producing higher

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

221

222 **Selectivity for spectrotemporal modulation statistics.** Can speech, music and song selectivity  
223 be explained by generic acoustic representations, such as spectrotemporal modulations that appear  
224 to drive much of the functional organization of human primary auditory cortex (Schönwiesner and  
225 Zatorre, 2009; Barton et al., 2012; Santoro et al., 2014)? This question is relevant since speech and  
226 music are known to have distinctive modulation rates (Singh and Theunissen, 2003; Ding et al.,  
227 2017). We recently introduced an algorithm for synthesizing sounds that are matched to natural  
228 sounds in their spectrotemporal modulation statistics, despite being acoustically distinct (**Fig 3A**)  
229 (Norman-Haignere and McDermott, 2018). We found previously that primary auditory regions  
230 produced very similar responses to natural and modulation-matched synthetic sounds, but that non-  
231 primary regions produced weak responses to the synthetic sounds, presumably because they lack  
232 higher-order structure necessary to drive neurons in non-primary regions.

233

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

246

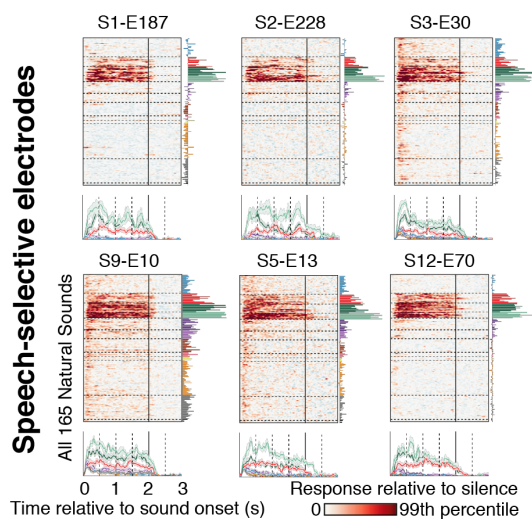
247 For all category-selective components, we observed a clear difference between the natural and  
248 modulation-matched synthetic sounds. The speech-selective components (3, 9, & 15) replicated  
249 their selectivity for natural speech with the new stimulus set (with an intermediate response to vocal  
250 music) and produced weak responses to the modulation-matched speech ( $p < 0.01$  via a sign test  
251 across sounds comparing natural and modulation-matched speech). The music-selective  
252 component (10) replicated its selectivity for natural music and responded weakly to modulation-  
253 matched music ( $p < 0.01$  via a sign test comparing natural and modulation-matched music). Critically,  
254 the song-selective component (12) responded nearly exclusively to the natural vocal music,  
255 producing weak responses to natural speech, natural instrumental music, and the modulation-  
256 matched vocal music ( $p < 0.01$  via a sign test comparing natural and modulation-matched vocal  
257 music; because there were only 2 vocal music sounds, the response to those two stimuli was  
258 subdivided into 500 ms segments to increase the number of samples). In contrast, most non-  
259 category selective components responded similarly to natural and modulation-matched sounds (**Fig**  
260 **3D**; Comp 7 showed modest selectivity for natural instrumental music, consistent with its response  
261 intermediate selectivity for instrumental music in the 165 natural sounds; see **Fig S5**). This finding  
262 demonstrates that speech, music, and song selectivity cannot be accounted for by spectrotemporal  
263 modulation statistics that appear to robustly drive responses throughout much of the rest of auditory  
264 cortex.



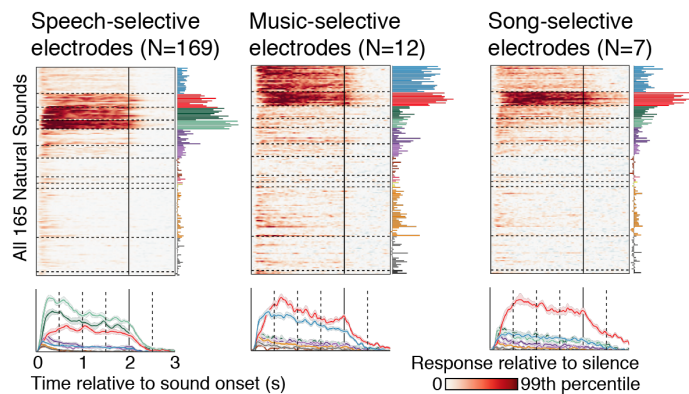
284 then measured their response in independent data. The electrode selection stage involved three  
285 steps (all performed on the same data and distinct from that used to measure the response). First,  
286 we measured the average response across time and stimuli to all sound categories with more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

297  
298 We plot the response of the top electrodes most significantly responsive to each contrast (**Fig 4A**)  
299 as well as the average response across all electrodes identified using this procedure (**Fig 4B**). We  
300 measured responses to the same natural sounds used to identify the electrodes (in independent  
301 data), as well as the natural and synthetic sounds from our control experiment (**Fig 4C**). As expected,  
302 given the coverage of ECoG grids relative to speech and music-selectivity (**Fig 1C**), we observed  
303 many more speech-selective electrodes than music or song-selective electrodes (169 speech-  
304 selective electrodes across all 13 subjects, 12 music-selective electrodes across 4 subjects, and 7  
305 song-selective electrodes across 3 subjects). But each of the music and song-selective electrodes  
306 identified replicated their selectivity for music or speech in independent data ( $p < 0.05$  for every  
307 electrode;  $p < 0.001$  for responses averaged across all music and song-selective electrodes; via  
308 bootstrapping the same contrast used to select electrodes but in independent data); and modulation-  
309 matched synthetic sounds produced a much weaker responses than natural sounds from the  
310 preferred category ( $p < 0.01$  via a sign test between responses to natural and model-matched sounds  
311 applied to the average response of speech, music, and song-selective electrodes). Some of the  
312 music-selective electrodes were strikingly selective. For example, S1-E147 (from a patient with  
313 small, high-density electrodes with 1 mm exposed diameters) responded in a near binary fashion,  
314 producing a high response for nearly all of the music sounds and a near-zero response for all other  
315 sounds.

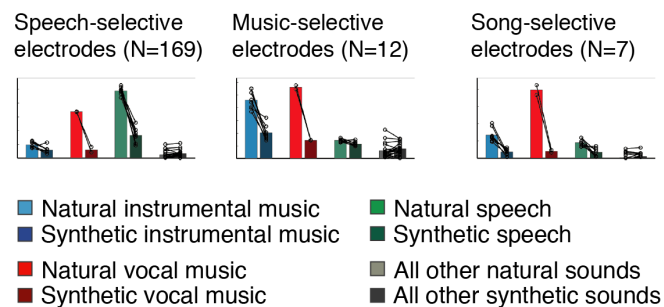
### A Individual electrodes



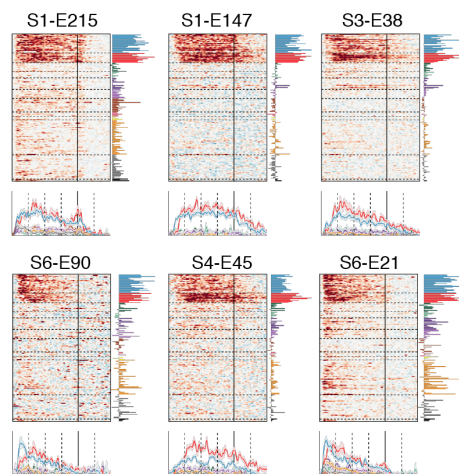
### B Average response of electrodes selective for speech, music and song



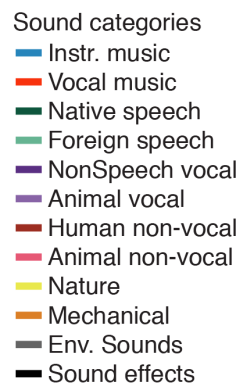
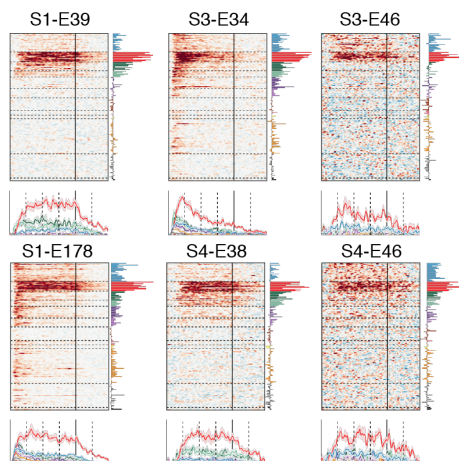
### C Response to natural vs modulation-matched synthetic sounds



### Music-selective electrodes



### Song-selective electrodes



316  
 317 **Figure 4. The response of individual electrodes selective for speech, music or song.** We selected speech  
 318 (top), music (middle), and song-selective (bottom) electrodes, and then measured their response in independent  
 319 data. **A**, The top six electrodes that showed the most significant response preference for each category in the  
 320 subset of data used to select electrodes. For speech-selective electrodes, the top 6 electrodes came from 2  
 321 subjects (2 from S1 and 4 from S2), and so we instead plot the top electrode from 6 different subjects to show  
 322 greater diversity. Conventions as in **Fig 2A**. **B**, The average response of all electrodes identified as speech,  
 323 music, or song-selective to the 165 natural sounds (same conventions as panel A). **C**, The average response of  
 324 speech, music, and song-selective electrodes to natural and modulation-matched sounds from the control  
 325 experiment (conventions as in **Fig 3C**).



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

338

## 339 Discussion

340

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

348

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

358

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

364

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

375

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

380  
381 **Music selectivity.** Researchers have long debated the extent to which music perception relies on  
382 specialized vs. general-purpose neural mechanisms (Patel, 2012; Peretz et al., 2015). Our study  
383 provides the first direct electrophysiological evidence that the human brain has neural populations  
384 that are highly selective for music.

385  
386 Our results also help validate the voxel decomposition method used in our prior work (Norman-  
387 Haignere et al., 2015). Using voxel decomposition, we inferred a component that was substantially  
388 more selective for music than were individual voxels, which we hypothesized was due to the overlap  
389 of distinct neural populations within a voxel. Our findings support this hypothesis by showing clear  
390 music selectivity using a more direct measure of neural activity. Moreover, many of the electrodes  
391 that showed the strongest selectivity for music (e.g. S1-E147, S1-E215) were sampled by a high-  
392 density grid with particularly small electrodes (1 mm exposed diameter), suggesting that high spatial  
393 resolution is useful for detecting clear music selectivity. Thus, our study both demonstrates the  
394 existence of music-selective neural populations, and helps explain why this type of selectivity has  
395 not been clearly observed with fMRI in standard voxel-wise analyses.

396  
397 **Speech and voice selectivity.** Many prior studies have reported selectivity for speech (Mesgarani  
398 et al., 2014; Norman-Haignere et al., 2015; Overath et al., 2015) and non-speech vocalizations (e.g.  
399 crying, laughing) (Belin et al., 2000) in the superior temporal gyrus. Distinguishing responses to  
400 speech and voice has been difficult, because speech-selective responses typically show at least  
401 some response to non-speech vocalizations and vice-versa. Here, we found multiple components  
402 (3, 15) and electrodes (e.g. S2-E54, S2-E222) that produced essentially no response to non-speech  
403 vocalizations, demonstrating that pure speech selectivity exists in the human brain. Thus, as with  
404 the music selectivity, the fact that fMRI voxels reflect a mixture of speech and voice selectivity may  
405 in part reflect the blurring together of nearby neural populations.

406  
407 **Onset/offset selectivity.** Many of the components we observed responded substantially more  
408 strongly at the onset or offset of sound, consistent with a recent study showing the onset selectivity  
409 is a prominent feature of human STG responses (Hamilton et al., 2018). Our study highlights the  
410 diversity of these responses across a wide variety of natural sounds: some components responded  
411 at the onset (Component 2) or offset (Component 17) of any sound, some were strongest for speech  
412 or vocalization stimuli (Components 1 & 4), and some were strongest for non-speech sounds  
413 (Component 6). Why so much of the STG is onset-selective is unclear. Some of these responses  
414 might reflect a generic/low-level adaptation mechanism in response to a sudden increment or  
415 decrement in sound energy. Others might reflect adaptation to higher-level stimulus statistics (Kvale  
416 and Schreiner, 2004), perhaps in the service of creating a more noise-robust (Mesgarani et al., 2014)  
417 or efficient (Barlow, 1961; Fairhall et al., 2001) representation of sound by suppressing responses  
418 to features that are predictable (Heilbron and Chait, 2017).

419  
420 **Component modeling: strengths, limitations and relationship to prior methods.** Component  
421 modeling provides a way to: (1) infer prominent response patterns; (2) suggest novel hypotheses  
422 that might not be obvious a-priori; and (3) disentangle spatially overlapping responses. Our results  
423 illustrate each of these benefits. We were able to infer a set of 20 response components that  
424 explained much of the response variation across hundreds of electrodes. We found evidence for a  
425 novel form of music selectivity (song selectivity) that we did not hypothesize a priori. And the

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

429  
430 The key challenge of component modeling is that matrix approximation is inherently ill-posed, and  
431 hence, the solution depends on statistical assumptions. Most component methods rely on just one  
432 of the following three assumptions: (1) non-negativity (Lee and Seung, 1999); (2) sparsity across  
433 time or space (Olshausen and Field, 1997; Hyvarinen, 1999); or (3) temporal smoothness (Wiskott  
434 and Sejnowski, 2002; Byron et al., 2009). We showed that all of these properties are evident in  
435 auditory ECoG responses. We developed a simple model to embody these assumptions and showed  
436 that the model better predicted ECoG responses compared with baseline models. We also showed  
437 that all of our category-selective components were evident using a model that imposed only non-  
438 negativity on the responses, suggesting that our key results were robust to the particular statistical  
439 assumptions imposed. Nonetheless, the assumptions of a component model are never perfect; and  
440 thus, it is useful to validate the results of a model with more direct analyses. Here, we found that  
441 speech, music and song selectivity were evident in individual electrodes, which demonstrates that  
442 our key findings were not dependent on statistical assumptions.

443  
444 Our prior fMRI voxel decomposition method used statistical constraints on the high-dimensional  
445 voxel weights to infer components (Norman-Haignere et al., 2015). By contrast, ECoG grids have  
446 many fewer electrodes than voxels, but each electrode has a richly structured timecourse. We thus  
447 chose to constrain the solution with statistics of the high-dimensional response timecourses. Our  
448 method is also distinct from a number of other component models that have been applied to high-  
449 dimensional neural data. Unlike many sparse convolutional models (Bouchard et al., 2017), each  
450 component of our model was defined by a single timecourse and a single pattern of electrode weights  
451 rather than by a time-varying spatial pattern, and thus can be more easily interpreted as the response  
452 of an underlying neuronal population. Unlike clustering methods (or convex NMF (Hamilton et al.,  
453 2018)), our method can disentangle responses that overlap within individual electrodes. And unlike  
454 most tensor decomposition methods (Williams et al., 2018), our method does not require the shape  
455 of a component's response timecourse to be identical across different stimuli, which is critical for  
456 modeling responses to sensory features that are not necessarily aligned to stimulus onset.

## 457 458 **Conclusions and future directions**

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

## 468 469 **Methods**

470  
471 **Participants.** Thirteen epilepsy patients participated in our study (mean age: 37 years, age standard  
472 deviation: 14 years; 8 right-handed; 8 female). These subjects underwent temporary implantation of  
473 subdural electrode arrays at Albany Medical College to localize the epileptogenic zones and to  
474 delineate these zones from eloquent cortical areas before brain resection. All of the subjects gave

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

477

478 **Electrode grids.** Most subjects had electrodes implanted in a single hemisphere, and STG coverage  
479 was much better in one of the two hemispheres in all subjects (8 right hemisphere patients and 5 left  
480 hemisphere patients; **Fig 1B** shows the coverage of the primary hemisphere for all subjects). In most  
481 subjects, electrodes had a 2.3 mm exposed diameter with a 6 mm center-to-center spacing for  
482 temporal lobe grids (10 mm spacing for grids in frontal, parietal and occipital lobe, but electrodes  
483 from these grids typically did not show reliable sound-driven responses; electrodes were embedded  
484 in silicone; PMT Corp., Chanhassen, MN). Two subjects were implanted with a higher-density grid  
485 (1 mm exposed diameter, 3 mm center-to-center spacing).

486

487 **Natural sounds.** The sound set was the same as in our prior study (Norman-Haignere et al., 2015).  
488 To generate the stimulus set, we began with a set of 280 everyday sounds for which we could find  
489 a recognizable, 2-second recording. Using an online experiment (via Amazon's Mechanical Turk),  
490 we excluded sounds that were difficult to recognize (below 80% accuracy on a ten-way multiple  
491 choice task; 55–60 participants for each sound), yielding 238 sounds. We then selected a subset of  
492 160 sounds that were rated as most frequently heard in everyday life (in a second Mechanical Turk  
493 study; 38–40 ratings per sound). Five additional “foreign speech” sounds were included (“German,”  
494 “French,” “Italian,” “Russian,” “Hindi”) to distinguish responses to acoustic speech structure from  
495 responses to linguistic structure (the 160-sound set included only two foreign speech stimuli:  
496 “Spanish” and “Chinese”). In total, there were 10 English speech stimuli, 7 foreign speech stimuli,  
497 21 instrumental music stimuli, and 11 vocal music stimuli (see *Sound category assignments*).  
498 Sounds were RMS-normalized and presented at a comfortable volume using sound isolating over-  
499 the-ear headphones (Panasonic RP-HTX7, 10 dB isolation). The sound set is freely available:

500

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

502

503 Subjects completed between three and seven runs of the experiment (S11: 3 runs, S6, S12: 4 runs,  
504 S13: 5 runs, S1: 7 runs; all other subjects: 6 runs). In each run, each natural sound was presented  
505 at least once. Between 14 and 17 of the sounds were repeated exactly back-to-back, and subjects  
506 were asked to press a button when they detected this repetition. This second instance of the sound  
507 was excluded from the analysis, because the presence of a target could otherwise bias responses  
508 in favor of the repeated stimuli. Each run used a different random ordering of stimuli. There was a  
509 1.4-2 second gap (randomly chosen) between consecutive stimuli.

510

511 **Modulation-matched synthetic sounds.** In ten of the subjects, we also measured responses to a  
512 distinct set of 36 natural sounds and 36 corresponding synthetic sounds that were individually  
513 matched to each natural sound in their spectrotemporal modulations statistics using the approach  
514 described in Norman-Haignere & McDermott (2018). The synthesis algorithm starts with an  
515 unstructured noise stimulus, and iteratively modifies the noise stimulus to match the modulation  
516 statistics of a natural sound. Modulations are measured using a standard model of auditory cortical  
517 responses in which a cochleagram is passed through a set of linear filters tuned to modulations at a  
518 particular audio frequency, temporal rate, and spectral scale (i.e. how coarse vs fine the modulations  
519 are in frequency) (Chi et al., 2005). The spectrotemporal filters were created by crossing 9 temporal  
520 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  
521 octave), and replicating each filter at each audio frequency. The synthesis procedure alters the noise  
522 stimulus to match the histogram of response magnitudes across time for each filter in the model,  
523 which has the effect of matching all time-averaged statistics (such as mean and variance) of the filter



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

526  
527 All stimuli were 4 seconds long. We used shorter stimuli than the 10-second stimuli used in Norman-  
528 Haignere & McDermott (2018) due to limitations in the recording time. Because the stimuli were  
529 shorter, we did not include the very low-rate filters (0.125 and 0.25 Hz), which were necessary for  
530 longer stimuli to prevent energy from clumping unnaturally at particular moments in the synthetic  
531 recording. We also did not include “DC filters” as in Norman-Haignere & McDermott, but instead  
532 simply matched the mean value of the cochleagram across time and frequency at each iteration of  
533 the algorithm. Norman-Haignere & McDermott describe two versions of the algorithm: one in which  
534 the histogram-matching procedure was applied to the raw filter outputs and one where the matching  
535 procedure was applied to the envelopes of the filter responses. We found that the resulting stimuli  
536 were very similar, both perceptually and in terms of the cortical response. The stimuli tested here  
537 were created by applying the histogram matching procedure to the envelopes.

538  
539 The stimuli were presented in a random order with a 1.4-1.8 second gap between stimuli (for the first  
540 subject tested, a 2-2.2 second gap was used). The natural sounds were repeated to make it possible  
541 to assess the reliability of stimulus-driven responses. For all analyses, we simply averaged  
542 responses across the two repetitions. The sound set was presented across 4 runs. In one subject  
543 (S1), the entire experiment was repeated (we averaged responses across the two repeats).

544  
545 **Sound category assignments.** In an online experiment, Mechanical Turk participants chose the  
546 category that best described each of the 165 sounds tested, and we assigned each sound to its most  
547 frequently chosen category (30–33 participants per sound) (Norman-Haignere et al., 2015).  
548 Category assignments were highly reliable (split-half kappa = 0.93). We chose to re-assign three  
549 sounds (“cymbal crash”, “horror film sound effects”, and “drum roll”) from the “instrumental music”  
550 category to a new “sound effects” category. There were two motivations for this re-assignment: (1)  
551 these three sounds were the only sounds assigned to the music category that produced weak fMRI  
552 responses in the music-selective component we inferred in our prior study, presumably because  
553 they lack canonical types of musical structure (i.e. clear notes, melody, rhythm, harmony, key, etc.);  
554 and (2) excluding these sounds makes our song selectivity contrast (vocal music – (instrumental  
555 music + speech)) more conservative as it is not biased upwards by the presence of instrumental  
556 music sounds that lack rich musical structure.

557  
558 **Preprocessing.** Preprocessing was implemented in MATLAB. The scripts used to implement the  
559 preprocessing steps are available here (we reference specific scripts within these directories in  
560 describing our analyses):

561  
562 <https://github.com/snormanhaignere/ecog-analysis-code>  
563 <https://github.com/snormanhaignere/general-analysis-code>

564  
565 The responses from all electrodes were common-average referenced to the grand mean across all  
566 electrodes (separately for each subject). We excluded noisy electrodes from the common-average  
567 reference by detecting anomalies in the 60 Hz power (see `channel_selection_from_60Hz_noise.m`;  
568 a IIR resonance filter with a 3dB down bandwidth of 0.6 Hz was used to measure the RMS power).  
569 Specifically, we excluded electrodes whose 60 Hz power exceeded 5 standard deviations of the  
570 median across electrodes. Because the standard deviation is itself sensitive to outliers, we estimated  
571 the standard deviation using the central 20% of samples, which are unlikely to be influenced by  
572 outliers (by dividing the range of the central 20% of samples by that which would be expected from  
573 a Gaussian of unit variance; see `zscore_using_central_samples.m`). After common-average

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

577  
578 We computed broadband gamma power by measuring the envelope of the preprocessed signal  
579 filtered between 70 and 140 Hz (see bandpass\_envelopes.m; bandpass filtering was implemented  
580 using a 6<sup>th</sup> order Butterworth filter with 3dB down cutoffs of 70 and 140 Hz; the filter was applied  
581 forward and backward using filtfilt.m). The envelope was measured as the absolute value of the  
582 analytic signal after bandpassing. For the single-electrode analyses (**Fig 4**), we downsampled the  
583 envelopes to 100 Hz (from the 1200 Hz recording rate), and smoothed the timecourses with a 50 ms  
584 FWHM kernel to reduce noise and make it easier to distinguish the timecourses for different  
585 categories in the plots. For component analyses, we downsampled the envelopes to 25 Hz, because  
586 this enabled us to fit the data in the limited memory available on the GPU used to perform the  
587 optimization (no smoothing was applied since the model inferred an appropriate smoothing kernel  
588 for each component).

589  
590 Occasionally, we observed visually obvious artifacts in the broadband gamma power for a small  
591 number of timepoints. These artifacts were typically localized to a small fraction of electrodes; thus,  
592 we detected artifacts separately for each electrode. For each electrode, we computed the 90<sup>th</sup>  
593 percentile of its response magnitudes across all timepoints, which is by definition near the upper-  
594 end of that electrode's response distribution, but which should also be unaffected by outliers  
595 assuming the number of outliers accounts for less than 10% of time points (which we generally found  
596 to be the case). We classified a timepoint as an outlier if it exceeded 5 times the 90<sup>th</sup> percentile value  
597 for each electrode. We found this value to be relatively conservative in that only a small number of  
598 timepoints were excluded (<1% for all sound-responsive electrodes), and the vast majority of the  
599 excluded timepoints were artifactual based on visual inspection of the broadband gamma  
600 timecourses. Because there were only a small number of outlier timepoints, we replaced the outliers  
601 values with interpolated values from nearby non-outlier timepoints. We also manually excluded some  
602 or all of the runs from 11 electrodes where there were a large number of artifacts.

603  
604 For each 2-second stimulus, we measured the response of each electrode during a three-second  
605 window locked to sound onset (for the 4-second natural and modulation-matched stimuli we used a  
606 5-second window). We detected the onset of sound in each stimulus by computing the waveform  
607 power in 10 ms bins (with a 2 ms hop), and selecting the first bin in which the audio power exceeded  
608 50 dB of the maximum power across all windows and stimuli. Following standard practice, the audio  
609 and ECoG data were synced by sending a copy of the audio signal to the same system used to  
610 record ECoG signals. This setup allowed us to measure the time delay between when the system  
611 initiated a trial and the onset of sound (which we measured using pure tones).

612  
613 Responses were converted to units of percent signal change relative to silence by subtracting and  
614 then dividing the response of each electrode by the average response during the 300 ms before  
615 each stimulus.

616  
617 **Session effects.** For five of the thirteen subjects, runs were collected across two sessions with a  
618 gap in between (typically a day; the 7th run of S1 was collected in a third session). For the vast  
619 majority of electrodes, we found that their response properties were stable across sessions.  
620 Occasionally, we observed electrodes that substantially changed their selectivity, potentially due to  
621 small changes in the positioning of the electrodes over the cortex. To identify such changes, from  
622 each run of data, we measured the time-averaged response of each electrode to each of the 165  
623 sounds tested. We then detected electrodes for which the test-retest correlation from runs of the



624 same session was significantly greater than the test-retest correlation from runs of different sessions  
625 ( $p < 10^{-5}$ ; we used time-averaged response profiles rather than the raw timecourses, because we  
626 found them to be more reliable, and thus a better target for detecting selectivity changes across  
627 sessions; for S1 we grouped the runs from the 2<sup>nd</sup> and 3<sup>rd</sup> session together since there was only a  
628 single run in the 3<sup>rd</sup> session). Significance was evaluated via a permutation test (Nichols and Holmes,  
629 2002) in which we permuted the correspondence between stimuli across runs (10,000). We used  
630 this approach to build up a null distribution for our test statistic (the difference between the correlation  
631 within and across sessions), fit this null distribution with a Gaussian (so that we could estimate small  
632 p-values that would have been impossible to estimate via counting), and used the null to calculate a  
633 two-sided p-value (by measuring the tail probability that exceeded the test statistic and multiplying  
634 by 2). Seven electrodes passed our conservative significance threshold. For these electrodes, we  
635 simply treated the data from different sessions as coming from different electrodes, since they likely  
636 sampled distinct neural populations.

637  
638 **Electrode selection.** We selected electrodes with a reliable response to the sound set. Specifically,  
639 we measured the test-retest correlation of each electrode's broadband gamma response timecourse  
640 across all sounds, measured in two splits of data (odd and even runs). We kept all electrodes with a  
641 test-retest correlation greater than 0.2 (electrodes with a test-retest correlation less than 0.2 were  
642 quite noisy upon inspection). Results were similar using a more liberal threshold of 0.1.

643  
644 **Electrode localization.** We localized electrodes in order to be able to visualize the electrode weights  
645 for each component. Electrode locations played no role in the identification of components or  
646 category-selective electrodes.

647  
648 Following standard practice, we identified electrodes as bright spots on a post-operative computer  
649 tomography (CT) image. The CT was aligned to a high-resolution, pre-operative magnetic  
650 resonance image (MRI) using a rigid-body transformation. We then projected each electrode onto  
651 the cortical surface, computed by Freesurfer from the MRI scan. This projection is error-prone  
652 because far-away points on the cortical surface can be spatially nearby due to cortical folding. As a  
653 consequence, it was not uncommon for electrodes very near STG, where sound-driven responses  
654 are common, to be projected to a spatially nearby point on middle temporal or supramarginal/inferior  
655 frontal gyrus, where sound-driven responses are much less common (**Fig S8**). To minimize gross  
656 errors, we preferentially localized sound-driven electrodes to regions where sound-driven responses  
657 are likely to occur. Specifically, using a recently collected fMRI dataset, where responses were  
658 measured to the same 165 sounds in a large cohort of 20 subjects with whole-brain coverage (our  
659 prior published study only had partial brain coverage (Norman-Haignere et al., 2015)), we calculated  
660 the fraction of subjects that showed a significant response to sound at each point in the brain ( $p <$   
661  $10^{-5}$ , measured using a permutation test (Norman-Haignere et al., 2016)). We treated this map as a  
662 prior and multiplied it by a likelihood map, computed separately for each electrode based on the  
663 distance of that electrode to each point on the cortical surface (using a Gaussian error distribution;  
664 10 mm FWHM). We then assigned each electrode to the point on the cortical surface where the  
665 product of the prior and likelihood was greatest (which can be thought of as the maximum posterior  
666 probability solution). We smoothed the prior probability map so that it would only effect the  
667 localization of electrodes at a coarse level, and not bias the location of electrodes locally, and we set  
668 the minimum prior probability to be 0.05 to ensure every point had non-zero prior probability. We plot  
669 the prior map and the effect it has on localization in **Fig S8**.

670  
671 **Responses statistics relevant to component modeling.** Our component model approximated the  
672 response of each electrodes as the weighted sum of a set of canonical response timecourses  
673 ("components"). The component timecourses are shared across all electrodes, but the weights are

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

678  
679 Matrix factorization is inherently ill-posed (that is, there are many equally good approximations).  
680 Thus, we constrained our factorization by additional statistical criteria. Most component methods rely  
681 on one of three statistical assumptions: (1) non-negativity (Lee and Seung, 1999); (2) a non-  
682 Gaussian distribution of response magnitudes across time or space (Olshausen and Field, 1997;  
683 Hyvarinen, 1999); or (3) temporal smoothness of the component responses (Wiskott and Sejnowski,  
684 2002; Byron et al., 2009). We investigated each of these statistical properties in broadband gamma  
685 responses to sound (**Fig S1**).

686  
687 To evaluate non-negativity, we measured the percent of the total RMS power accounted for by  
688 positive vs. negative responses during the presentation of sound (measured relative to 300 ms of  
689 silence preceding the onset of each sound):  
690

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

691  
692 where  $p$  and  $n$  are shorthand for positive and negative values. We applied the above equation to the  
693 response of individual electrodes (summing over all timepoints for all sounds; **Fig S1A,B**), as well  
694 as to the pooled response of all sound-responsive electrodes (summing over all timepoints, sounds,  
695 and electrodes; **Fig S1D**). To minimize the effect of measurement noise, which will create negative  
696 values even if none are present (since measurement noise will not depend on the stimulus and thus  
697 noise fluctuations will be symmetric around the silent baseline), we measured the response of all  
698 electrodes in two splits of data (average across odd and even runs). We then: (1) sorted the response  
699 magnitudes of all timepoints by their magnitude in the first split; (2) measured their response in the  
700 second split; and (3) applied a median filter to the sorted response magnitudes from the second  
701 splits, thus suppressing unreliable response variation (filter size = 100 when applied to individual  
702 electrodes, filter size = 10,000 when pooling responses across all electrodes) (**Fig S1B&D** show the  
703 results of applying this procedure to individual electrodes and the pooled response of all electrodes).  
704 When equation 1 was applied to the de-noised response distributions (i.e. median filtered responses  
705 from the second split), we found that positive responses accounted for 99.97% of the RMS power  
706 across all sound-responsive electrodes. Note that sound-responsive electrodes were selected based  
707 on the reliability of their responses, not based on a greater response to sounds compared with  
708 silence, and thus our analysis is not biased by our selection criterion.

709  
710 To investigate whether and how the distribution of responses might differ from a Gaussian, we  
711 measured the skewness (normalized 3<sup>rd</sup> moment) and sparsity (excess kurtosis relative to a  
712 Gaussian) of the responses:  
713

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

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

714

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

723  
724 We also tested the skewness and sparsity of responses across electrodes by applying equations 2  
725 and 3 to the distribution of responses across electrodes. Specifically, we measured the averaged  
726 response of each electrode to each sound, and then for each sound, we applied equations 2 and 3  
727 to the distribution of responses across the 271 sound-responsive electrodes. **Fig S1G** plots the  
728 histogram of these skewness and sparsity measures for all 165 sounds. We did not apply our de-  
729 noising procedure since we only had 271 electrodes which made the sorting/median-filtering  
730 procedure infeasible (in contrast, for each electrode we had 12,375 timepoints across all sounds);  
731 instead we time-averaged the response of each electrode to each sound to reduce noise. We again  
732 found that this distribution was significantly skewed and sparse relative to a Gaussian and relative  
733 to what would be expected given just noise in the data ( $p < 0.001$  via sign test).

734  
735 Finally, to investigate the temporal smoothness of auditory ECoG responses, we measured the  
736 normalized autocorrelation of each electrode's response (**Fig S1C,E**). To prevent noise from  
737 influencing the result, we correlated responses measured in independent runs (odd and even runs).  
738 This analysis revealed substantial long-term dependencies over more than a second, and the  
739 strength of these dependencies varied substantially across electrodes. This substantial variation  
740 across electrodes demonstrates that these long-term dependencies are not a by-product of  
741 measuring broadband gamma power (in simulations, we have found that our measurement  
742 procedure can resolve power fluctuations up to  $\sim 30$  Hz, assuming a 70-140 Hz carrier).

743  
744 Together, the results from our analysis reveal three key properties of auditory ECoG: (1) nearly all  
745 responses are positive/excitatory relative to sound onset; (2) responses are skewed/sparse across  
746 time/stimuli and electrodes; and (3) responses are temporally smooth and the extent of this  
747 smoothness varies across electrodes. We sought to design a simple component model that captures  
748 these three essential properties. We refer to this model as the "Sparse and Smooth Component"  
749 (SSC) model.

750  
751 **Component model.** Each electrode is represented by its response timecourse across all sounds  
752 ( $e_i(t)$ ) (**Fig S2A**). We approximate this response timecourse as the weighted sum of  $K$  component  
753 response timecourses ( $r_k(t)$ ):

754

$$e_i(t) \approx \sum_{k=1}^K r_k(t) w_{i,k} \quad \text{s4}$$

755  
756 The component timecourses are shared across all electrodes, but the weights are separate for each  
757 electrode, allowing the model to approximate different response patterns. We constrain all of the  
758 component responses and weights to be positive, since we found that nearly all of the sound-driven  
759 responses were positive. To encourage the components to be both sparse and smooth, we model  
760 the response timecourse of each component as the convolution of a set of sparse activations ( $a_k(t)$ )  
761 with a smoothing kernel ( $h_k(t)$ ):

762

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

5

763

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

769

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

770

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

778

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

785

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

798

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

802

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

7

803



804 **Cross-validation analyses.** We used cross-validated prediction accuracy to determine the number  
805 of components and the sparsity parameter (**Fig 1E & S2B**), as well as to compare the SSC model  
806 with several baseline models (**Fig S2C**). For the purposes of cross-validation, we separated the  
807 timecourses for different sounds into cells, thus creating an electrode x sound matrix (**Fig 1E**). We  
808 then trained the model on a random subset of 80% of cells and measured the model's prediction  
809 accuracy (squared Pearson correlation) in the left-out 20% of cells. We trained models starting from  
810 10 different random initializations, and selected the model with the lowest error in the training data.  
811 We repeated our analyses using 5 different random splits of train and test data, averaging the test  
812 correlations across splits. For each split, we ensured an even and broad sampling of train and test  
813 stimuli using the following procedure: (1) we created a random ordering of stimuli and electrodes (2)  
814 we assigned the first 20% of sounds to be test sounds for the first electrode, the next 20% of sounds  
815 to be test sounds for electrodes 2, and so on. After using up all 165 sounds (every 8-9 electrodes),  
816 we refreshed the pool of available test sounds using a new random ordering of stimuli.

817  
818 To prevent correlated noise across electrodes from influencing the results, we used non-overlapping  
819 sets of runs (odd and even runs) to compute the training and test data (i.e. training on odd runs and  
820 testing on even runs, and vice-versa; again averaging test correlations across the two splits). For a  
821 given set of hyper-parameters, we then averaged the test correlations across all electrodes to arrive  
822 at a summary measure of that model's performance (**Fig 1E & S2B**). We noise-corrected this  
823 measure by dividing it by the average test-retest correlation of the electrode responses (using the  
824 unsquared Pearson correlation), which gives an upper bound on the model's prediction accuracy  
825 (Norman-Haignere et al., 2015; Schoppe et al., 2016).

826  
827 We considered several baseline models that did not use the convolutional decomposition of the SSC  
828 model (specifically, we constrained the smoothing kernel to be a delta function such that the  
829 component activations,  $\alpha_k(t)$ , equaled the component responses,  $r_k(t)$ ). We tested four baseline  
830 models: (1) we removed the sparseness and smoothness constraints entirely but maintained the  
831 non-negativity constraint (i.e. non-negative matrix factorization / NMF); (2) we imposed sparsity but  
832 not smoothness via an L1 penalty the component responses and weights (3) we imposed  
833 smoothness but not sparsity via an L2 smoothness penalty on the derivative of the component  
834 responses (the first-order difference of adjacent time-points); and (4) we applied both the L1 sparsity  
835 and L2 smoothness constraint. To prevent the number of hyper-parameters from biasing the results,  
836 for each electrode, we selected the hyper-parameters that led to the best performance across  
837 electrodes from other subjects (**Fig S2C**). We used grid-search over the following range of hyper-  
838 parameters:  $K$  (number of components) = [5,10,15,20,25,30],  $\lambda$  (sparsity) =  
839 [0,0.033,0.1,0.33,1,3.3],  $\omega$  (smoothness) = [0,0.033,0.1,0.33] (we verified that the best-performing  
840 models were not on the boundary of these values, except in cases where the best-performing model  
841 had a parameter value of 0). We found that all of the baseline models performed worse than the  
842 SSC model ( $p < 0.001$  via bootstrapping across subjects, see *Statistics*; including the model with  
843 both an L1 sparsity and L2 smoothness penalty, which had more hyper-parameters). This result  
844 shows that our convolutional decomposition is an effective way of capturing both the smoothness  
845 and sparsity of auditory broadband gamma responses, and is more effective than simply imposing  
846 sparsity and smoothing penalties directly on the component responses.

847  
848 **Constraining the smoothing kernel.** We investigated three potential methods for forcing the  
849 smoothing kernel to be smooth: (1) using a parametric kernel (e.g. Gamma distribution); (2) placing  
850 a smoothness penalty on the derivative of the kernel; and (3) constraining the kernel to be unimodal.  
851 We found that the optimizer had difficulty minimizing the loss when using parametric kernels (likely  
852 because the low-dimensional parameters of the kernel interacted in complex ways with the other  
853 high-dimensional parameters). We found that penalizing the derivative and constraining the kernel

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

857  
858 We constrained the kernel to be unimodal by placing two constraints on its derivative: (1) the first N  
859 points of the derivative must be positive and the remaining points must be negative (which forces  
860 the kernel to go up and then down, but not oscillate); and (2) the sum of the derivative must equal 0  
861 (ensuring that the kernel starts and ends at zero). The set of operations used to implement these  
862 constraints in TensorFlow is described in **Fig S10**. Many of the learned smoothing kernels were  
863 asymmetric, with a rapid rise and a slower falloff (**Fig S9**). There is nothing in the constraints that  
864 encourages asymmetry, and so this property must reflect an asymmetry in the cortical responses  
865 themselves.

866  
867 **Specificity of components for individual subjects.** The sparse and clinically-driven coverage of  
868 ECoG grids virtually guarantees that some response types will only be present in a subset of  
869 subjects. Thus, one might expect to find components that are subject-specific. To evaluate this  
870 possibility, we measured the average weight of each component in each subject, and then  
871 normalized these mean weights to sum to one across subjects (**Fig 1F**). Most components had  
872 substantial weights for multiple subjects, but for five of the 20 components, one subject accounted  
873 for more than half of the normalized subject weights (Components 14, 16, 18, 19, 20). We thus chose  
874 to focus on the components that were more general.

875  
876 For the 15-component model (**Fig S4**), three components had normalized subject weights greater  
877 than 0.5 (one other component was omitted because it was not stable across random re-  
878 initializations of the algorithm). For component model constrained only by non-negativity (**Fig S3**),  
879 two components had normalized subject weights greater than 0.5, and three other components  
880 weighted strongly on a single electrode (with one electrode accounting for more than 25% of the  
881 total weights across all electrodes), and were thus excluded from the plots shown.

882  
883 **Component responses to modulation-matched sounds.** The components were inferred using  
884 responses to just the 165 natural sounds from the main experiment. But since a subset of ten  
885 subjects were tested in both experiments, we could estimate the response of these same  
886 components to the natural and synthetic sounds from our control experiment. Specifically, we fixed  
887 the component electrode weights to the values inferred from the responses in our main experiment,  
888 and learned a new set of component response timecourses that best approximated the measured  
889 responses in the modulation-matching experiment. Since the electrode weights are known, this  
890 analysis is no longer ill-posed, and we thus removed all of the additional sparsity and smoothness  
891 constraints and simply estimated a set of non-negative response profiles that minimized the squared  
892 reconstruction error (we left the non-negativity constraint because we found that nearly all of the  
893 measured responses were non-negative).

894  
895 **Single electrode analyses.** To identify electrodes selective for music, speech and song, we defined  
896 a number of contrasts based on the average response to different categories (the contrasts are  
897 described in the Results). We then divided each contrast by the maximum response across all  
898 categories to compute a measure of selectivity, or we bootstrapped the contrast to determine if it  
899 was significantly greater than zero (see *Statistics* below). In all cases, we used independent data to  
900 identify electrodes and measure their response. Specifically, we used two runs (first and last) to  
901 select electrodes and the remaining runs to evaluate their response.

902



903 **Statistics.** The significance of all category contrasts was evaluated using bootstrapping (Efron,  
904 1982). Specifically, we sampled sounds from each category with replacement (100,000 times),  
905 averaged responses across the sampled sounds for each category, and then recomputed the  
906 contrast of interest (all of the contrasts tested are specified in the Results). We then counted the  
907 fraction of samples that fell below zero and multiplied by 2 to compute a two-sided p-value. For p-  
908 values smaller than 0.001, counting becomes unreliable, and so we instead fit the distribution of  
909 bootstrapped samples with a Gaussian and measured the tail probability that fell below zero (and  
910 multiplied by 2 to compute a two-sided p-value). For the component analyses, we corrected for  
911 multiple comparisons by multiplying these p-values by the number of components (corresponding to  
912 Bonferroni correction).

913  
914 We compared the song-selective component (Component 12) with the average response of all song-  
915 selective electrodes by counting the fraction of bootstrapped samples where the component showed  
916 greater super-additive selectivity for vocal music (vocal music > max(English speech, foreign  
917 speech) + instrumental music). We found that across all 100,000 bootstrapped samples, the  
918 component always showed greater selectivity.

919  
920 We also used bootstrapping to compute error bars for the category timecourses (**Fig 2A, Figs S3-**  
921 **5**). In these figures we plot the central 68% of the sampling distribution (equivalent to one standard  
922 error for a Gaussian distributed variable). We only plot categories for which there were more than 5  
923 exemplars.

924  
925 To test for laterality effects, we computed the mean difference in the component electrode weights  
926 between the right and left hemispheres (**Fig S6**). We then bootstrapped this difference score by  
927 sampling subjects with replacement, and recomputing the mean difference using only electrodes  
928 from the sampled subjects. We repeated this procedure 100,000 times, and computed a p-value by  
929 counting the fraction of samples falling below or above zero (whichever was smaller) and multiplying  
930 by 2. We again Bonferroni-corrected by simply multiplying the p-value by the number of components.  
931 Only one component (Component 17, which was offset-selective) was significant after correction (p  
932 = 0.032 after correction).

933  
934 We also used bootstrapping across subjects to place error bars on model prediction scores.  
935 Specifically, we (1) sampled subjects with replacement (10,000 times); (2) averaged the test  
936 correlation values (squared Pearson correlation) across the electrodes from the sampled subjects;  
937 and (3) divided by the average test-retest correlation (unsquared Pearson correlation) of the sampled  
938 electrodes to noise-correct our measure. We tested whether the SSC model outperformed our  
939 baseline models by counting the fraction of bootstrapped samples where the average test predictions  
940 were lower than each baseline model and multiplying by 2 to arrive at a two-sided p-value. When  
941 plotting the test predictions for different models (**Fig S2C**), we used “within-subject” error bars (Loftus  
942 and Masson, 1994), computed by subtracting off the mean of each bootstrapped sample across all  
943 models before measuring the central 68% of the sampling distribution. We multiplied the central 68%  
944 interval by the correction factor shown below to account for a downward bias in the standard error  
945 induced by mean-subtraction (Loftus and Masson, 1994):

946

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

8

947

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

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

958  
959 Although there were only two vocal music stimuli in the modulation-matching experiment, the stimuli  
960 were relatively long (4 seconds). We thus subdivided the response to each stimulus into seven 500  
961 ms segments (discarding the first 500 ms to account for the build-up in the response), and measured  
962 the average response to each segment. For both the song-selective component and the average  
963 response of song-selective electrodes, we found that for all fourteen 500-ms segments (7 segments  
964 across 2 stimuli), the response to natural vocal music was higher than the response to the  
965 modulation-matched controls, and thus is significant with a sign test ( $p < 0.001$ ).

966  
967 To determine whether the electrode responses were significantly more skewed and sparse than  
968 would be expected given noise (i.e. to evaluate the significance of the skewness/sparsity measures  
969 described in *Response statistics relevant to component modeling*), we computed skewness/sparsity  
970 using two data quantities: (1) the residual error after subtracting the response to even and odd runs;  
971 and (2) the summed response across even and odd runs. The properties of the noise should be the  
972 same for these two quantities, but the second quantity will also contain the reliable stimulus-driven  
973 component of the response. Thus, if the second quantity is more skewed/sparse than the first  
974 quantity, then the stimulus-driven response must be more skewed/sparse than the noise. To assess  
975 skewness/sparsity across time/stimuli, we measured the skewness and sparsity (equations 2 and 3)  
976 separately for each electrode using the residual error and summed response (pooling responses  
977 across all timepoints and stimuli). In every subject, we found that the average skewness/sparsity of  
978 the summed responses was greater than the skewness/sparsity of the residual error, and thus  
979 significant with a sign test ( $p < 0.001$ ). We used the same approach to evaluate the  
980 skewness/sparsity of responses across electrodes, measured separately for each sound. Using a  
981 sign test across sounds, we found both the skewness and sparsity of the summed response to be  
982 significantly greater than that for the residual error ( $p < 0.001$ ).

## 983 **Acknowledgements**

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

## 990 **Competing interests**

991  
992 Authors declare no competing financial and/or non-financial interests in relation to the work  
993 described in this paper.

## 994 **References**

995  
996  
997

- 998 Angulo-Perkins A, Aubé W, Peretz I, Barrios FA, Armony JL, Concha L (2014) Music listening  
999 engages specific cortical regions within the temporal lobes: Differences between musicians  
1000 and non-musicians. *Cortex* 59:126–137.
- 1001 Barlow HB (1961) Possible principles underlying the transformation of sensory messages. *Sens*  
1002 *Commun* 1:217–234.
- 1003 Barton B, Venezia JH, Saberi K, Hickok G, Brewer AA (2012) Orthogonal acoustic dimensions define  
1004 auditory field maps in human cortex. *Proc Natl Acad Sci* 109:20738–20743.
- 1005 Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex.  
1006 *Nature* 403:309–312.
- 1007 Bouchard KE, Bujan AF, Chang EF, Sommer FT (2017) Sparse coding of ECoG signals identifies  
1008 interpretable components for speech control in human sensorimotor cortex. In: 2017 39th  
1009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society  
1010 (EMBC), pp 3636–3639.
- 1011 Byron MY, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, Sahani M (2009) Gaussian-process  
1012 factor analysis for low-dimensional single-trial analysis of neural population activity. In:  
1013 *Advances in neural information processing systems*, pp 1881–1888.
- 1014 Casey M, Thompson J, Kang O, Raizada R, Wheatley T (2012) Population codes representing  
1015 musical timbre for high-level fMRI categorization of music genres. In: *Machine Learning and*  
1016 *Interpretation in Neuroimaging*, pp 34–41. Springer.
- 1017 Casey MA (2017) Music of the 7Ts: Predicting and decoding multivoxel fMRI responses with  
1018 acoustic, schematic, and categorical Music Features. *Front Psychol* 8:1179.
- 1019 Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. *J*  
1020 *Acoust Soc Am* 118:887–906.
- 1021 Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D (2017) Temporal modulations in speech and  
1022 music. *Neurosci Biobehav Rev*.
- 1023 Efron B (1982) The jackknife, the bootstrap, and other resampling plans. *Siam*.
- 1024 Fairhall AL, Lewen GD, Bialek W, de Ruyter van Steveninck RR (2001) Efficiency and ambiguity in  
1025 an adaptive neural code. *Nature* 412:787–792.
- 1026 Fedorenko E, McDermott JH, Norman-Haignere S, Kanwisher N (2012) Sensitivity to musical  
1027 structure in the human brain. *J Neurophysiol* 108:3289–3300.
- 1028 Hamilton LS, Edwards E, Chang EF (2018) A spatial map of onset and sustained responses to  
1029 speech in the human superior temporal gyrus. *Curr Biol* 28:1860-1871.e4.
- 1030 Heilbron M, Chait M (2017) Great expectations: is there evidence for predictive coding in auditory  
1031 cortex? *Neuroscience*.
- 1032 Herholz SC, Zatorre RJ (2012) Musical training as a framework for brain plasticity: behavior, function,  
1033 and structure. *Neuron* 76:486–502.

- 1034 Hyvarinen A (1999) Fast and robust fixed-point algorithms for independent component analysis.  
1035 Neural Netw IEEE Trans On 10:626–634.
- 1036 Kell AJ, Yamins DL, Shook EN, Norman-Haignere SV, McDermott JH (2018) A task-optimized neural  
1037 network replicates human auditory behavior, predicts brain responses, and reveals a cortical  
1038 processing hierarchy. Neuron.
- 1039 Kuhn HW (1955) The Hungarian method for the assignment problem. Nav Res Logist Q 2:83–97.
- 1040 Kvale MN, Schreiner CE (2004) Short-term adaptation of auditory receptive fields to dynamic stimuli.  
1041 J Neurophysiol 91:604–612.
- 1042 Leaver AM, Rauschecker JP (2010) Cortical representation of natural complex sounds: effects of  
1043 acoustic features and auditory object category. J Neurosci 30:7604–7612.
- 1044 Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature  
1045 401:788–791.
- 1046 Loftus GR, Masson ME (1994) Using confidence intervals in within-subject designs. Psychon Bull  
1047 Rev 1:476–490.
- 1048 Lomax A (2017) Folk song style and culture. Routledge.
- 1049 Mehr S, Singh M, Knox D, Ketter D, Pickens-Jones D, Atwood S, Lucas C, Egnér A, Jacoby N,  
1050 Hopkins EJ (2018) A natural history of song.
- 1051 Mehr SA, Krasnow MM (2017) Parent-offspring conflict and the evolution of infant-directed song.  
1052 Evol Hum Behav 38:674–684.
- 1053 Merrill J, Sammler D, Bangert M, Goldhahn D, Lohmann G, Turner R, Friederici AD (2012)  
1054 Perception of words and pitch patterns in song and speech. Front Psychol 3.
- 1055 Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior  
1056 temporal gyrus. Science 343:1006–1010.
- 1057 Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a  
1058 primer with examples. Hum Brain Mapp 15:1–25.
- 1059 Norman-Haignere SV, Albouy P, Caclin A, McDermott JH, Kanwisher NG, Tillmann B (2016) Pitch-  
1060 responsive cortical regions in congenital amusia. J Neurosci.
- 1061 Norman-Haignere SV, Kanwisher NG, McDermott JH (2015) Distinct cortical pathways for music and  
1062 speech revealed by hypothesis-free voxel decomposition. Neuron 88:1281–1296.
- 1063 Norman-Haignere SV, McDermott JH (2018) Neural responses to natural and model-matched stimuli  
1064 reveal distinct computations in primary and nonprimary auditory cortex. PLoS Biol  
1065 16:e2005127.
- 1066 Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: A strategy employed  
1067 by V1? Vision Res 37:3311–3325.
- 1068 Overath T, McDermott JH, Zarate JM, Poeppel D (2015) The cortical analysis of speech-specific  
1069 temporal structure revealed by responses to sound quilts. Nat Neurosci 18:903–911.

- 1070 Patel AD (2012) Language, music, and the brain: a resource-sharing framework. *Lang Music Cogn*  
1071 *Syst*:204–223.
- 1072 Patel AD (2019) Evolutionary music cognition: Cross-species studies. In: *Foundations in Music*  
1073 *Psychology: Theory and Research*, pp 459–501.
- 1074 Peretz I (2016) Neurobiology of congenital amusia. *Trends Cogn Sci* 20:857–867.
- 1075 Peretz I, Vuvan D, Lagrois M-É, Armony JL (2015) Neural overlap in processing music and speech.  
1076 *Philos Trans R Soc B Biol Sci* 370:20140090.
- 1077 Peterson RL, Pennington BF (2015) Developmental dyslexia. *Annu Rev Clin Psychol* 11:283–307.
- 1078 Ray S, Maunsell JHR (2011) Different origins of gamma rhythm and high-gamma activity in macaque  
1079 visual cortex. *PLOS Biol* 9:e1000610.
- 1080 Santoro R, Moerel M, De Martino F, Goebel R, Ugurbil K, Yacoub E, Formisano E (2014) Encoding  
1081 of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex.  
1082 *PLoS Comput Biol* 10.
- 1083 Savage PE, Brown S, Sakai E, Currie TE (2015) Statistical universals reveal the structures and  
1084 functions of human music. *Proc Natl Acad Sci* 112:8987–8992.
- 1085 Schindler A, Herdener M, Bartels A (2013) Coding of Melodic Gestalt in Human Auditory Cortex.  
1086 *Cereb Cortex* 23:2987–2993.
- 1087 Schönwiesner M, Zatorre RJ (2009) Spectro-temporal modulation transfer function of single voxels  
1088 in the human auditory cortex measured with high-resolution fMRI. *Proc Natl Acad Sci*  
1089 106:14611–14616.
- 1090 Schoppe O, Harper NS, Willmore BD, King AJ, Schnupp JW (2016) Measuring the performance of  
1091 neural models. *Front Comput Neurosci* 10:10.
- 1092 Scott SK, Blank CC, Rosen S, Wise RJ (2000) Identification of a pathway for intelligible speech in  
1093 the left temporal lobe. *Brain* 123:2400–2406.
- 1094 Singh NC, Theunissen FE (2003) Modulation spectra of natural sounds and ethological theories of  
1095 auditory processing. *J Acoust Soc Am* 114:3394–3411.
- 1096 Steinschneider M, Fishman YI, Arezzo JC (2008) Spectrotemporal analysis of evoked and induced  
1097 electroencephalographic responses in primary auditory cortex (A1) of the awake monkey.  
1098 *Cereb Cortex* 18:610–625.
- 1099 Tierney A, Dick F, Deutsch D, Sereno M (2013) Speech versus song: multiple pitch-sensitive areas  
1100 revealed by a naturally occurring musical illusion. *Cereb Cortex* 23:249–254.
- 1101 Wallin NL, Merker B, Brown S (2001) *The origins of music*. MIT press.
- 1102 Whittingstall K, Logothetis NK (2009) Frequency-band coupling in surface EEG reflects spiking  
1103 activity in monkey visual cortex. *Neuron* 64:281–289.

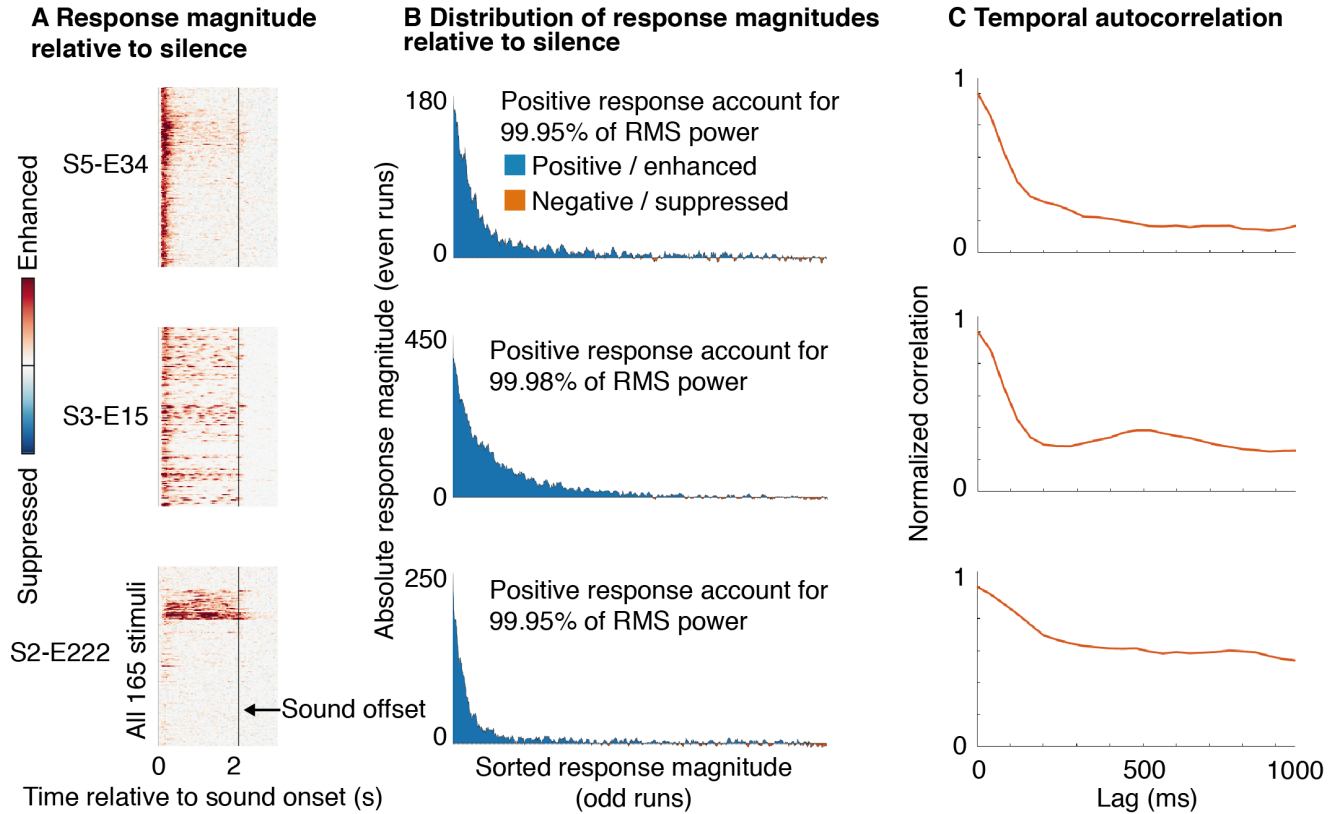
- 1104 Williams AH, Kim TH, Wang F, Vyas S, Ryu SI, Shenoy KV, Schnitzer M, Kolda TG, Ganguli S  
1105 (2018) Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple  
1106 timescales through tensor component analysis. *Neuron* 98:1099–1115.
- 1107 Wiskott L, Sejnowski TJ (2002) Slow feature analysis: Unsupervised learning of invariances. *Neural*  
1108 *Comput* 14:715–770.
- 1109



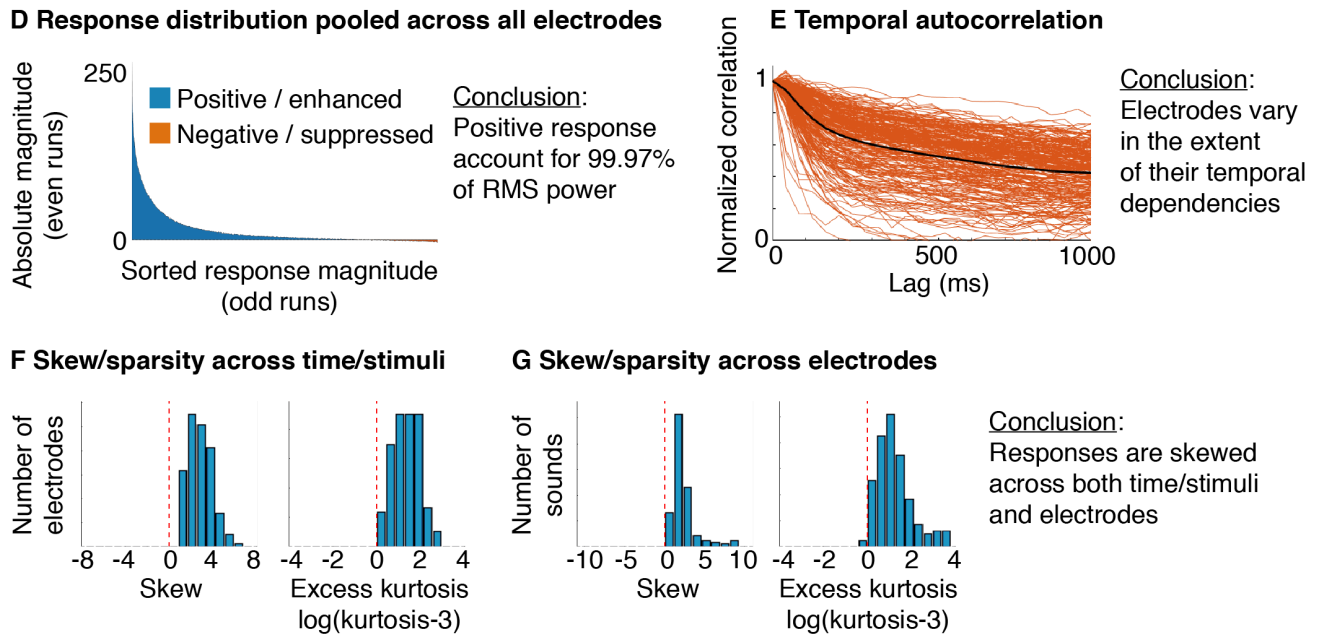
## 1110 Supplemental Figures

### Response statistics relevant to component modeling

#### Example electrodes



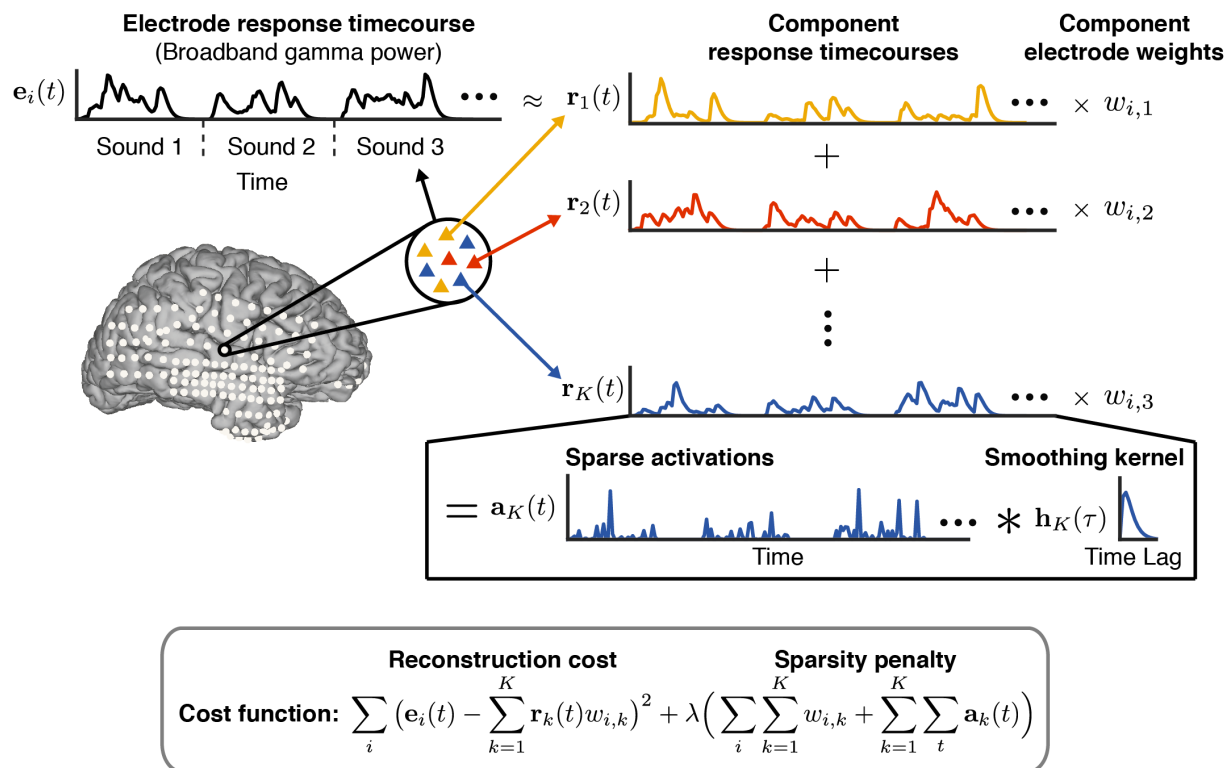
#### Summary statistics



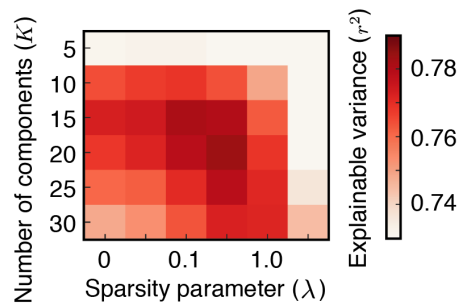
1111  
1112 **Figure S1. Response statistics relevant to component modeling.** A-C, Response statistics from three example  
1113 electrodes with distinct selectivities, but a shared set of statistical properties (positivity, sparsity/skew, and temporal  
1114 smoothness). A, Broadband gamma power response of each electrode to all 165 sounds as a raster. Responses are  
1115 measured relative to the response during silence (300 milliseconds preceding sound onset). Positive values (red)

1116 indicate an enhanced response to sound, and negative responses indicate a suppressed response (blue). The color  
1117 scales shows values from 0 to the 99<sup>th</sup> percentile of the response magnitude distribution for each electrode. **B**,  
1118 Distribution of response magnitudes, measured in a cross-validated fashion to reduce effects of noise: using data from  
1119 the odd runs, we sorted all of the bins of the raster on the left based on their magnitude (pooling across all timepoints  
1120 and stimuli). The response of each bin was then measured using the even runs, and then smoothed using a median  
1121 filter to suppress noise. Positive responses accounted for >99% of the RMS response power in all three electrodes. All  
1122 three electrodes show a skewed and sparse distribution of response magnitudes (quantified in panel F, below) because  
1123 negative responses were practically non-existent (yielding an asymmetric, rightward-skewed distribution) and strong  
1124 positive responses were present for only a small fraction of bins (yielding a sparse distribution). **C**, The normalized  
1125 autocorrelation (normalized by the correlation at zero lag) of each electrode's response measured in a cross-validated  
1126 fashion by correlating the response in odd and even runs at different lags. **D-G**, Summary statistics across all sound-  
1127 responsive electrodes. **D**, Distribution of response magnitude pooled across all electrodes, sounds and timepoints  
1128 (measured in a cross-validated fashion, as described above). Positive responses accounted for >99% of the RMS power.  
1129 **E**, Normalized autocorrelation of all sound-responsive electrodes. The extent of temporal dependencies varied  
1130 substantially across electrodes. **F**, We measured the skew (3<sup>rd</sup> moment) and sparsity (excess kurtosis) of each  
1131 electrode's response using its distribution of response magnitudes across all timepoints/stimuli (i.e. using the  
1132 distributions shown in panel B). This figure plots a histogram of the skew and sparsity values across all electrodes. We  
1133 subtracted the measured kurtosis from that which would be expected from a Gaussian (which has a kurtosis of 3). All  
1134 electrodes were skewed and sparse relative to a Gaussian. **G**, For each sound, we measured the skew and sparsity of  
1135 responses across electrodes, after averaging the response of each electrode to each sound. This figure plots a histogram  
1136 of the skew and sparsity values across all sounds.

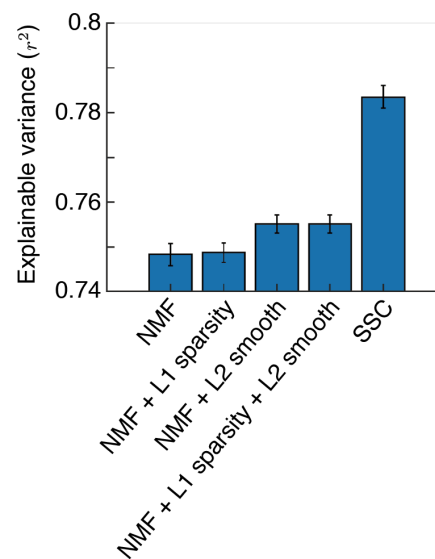
## A Schematic of Sparse and Smooth Component (SSC) model



## B Effect of hyper-parameters on prediction accuracy

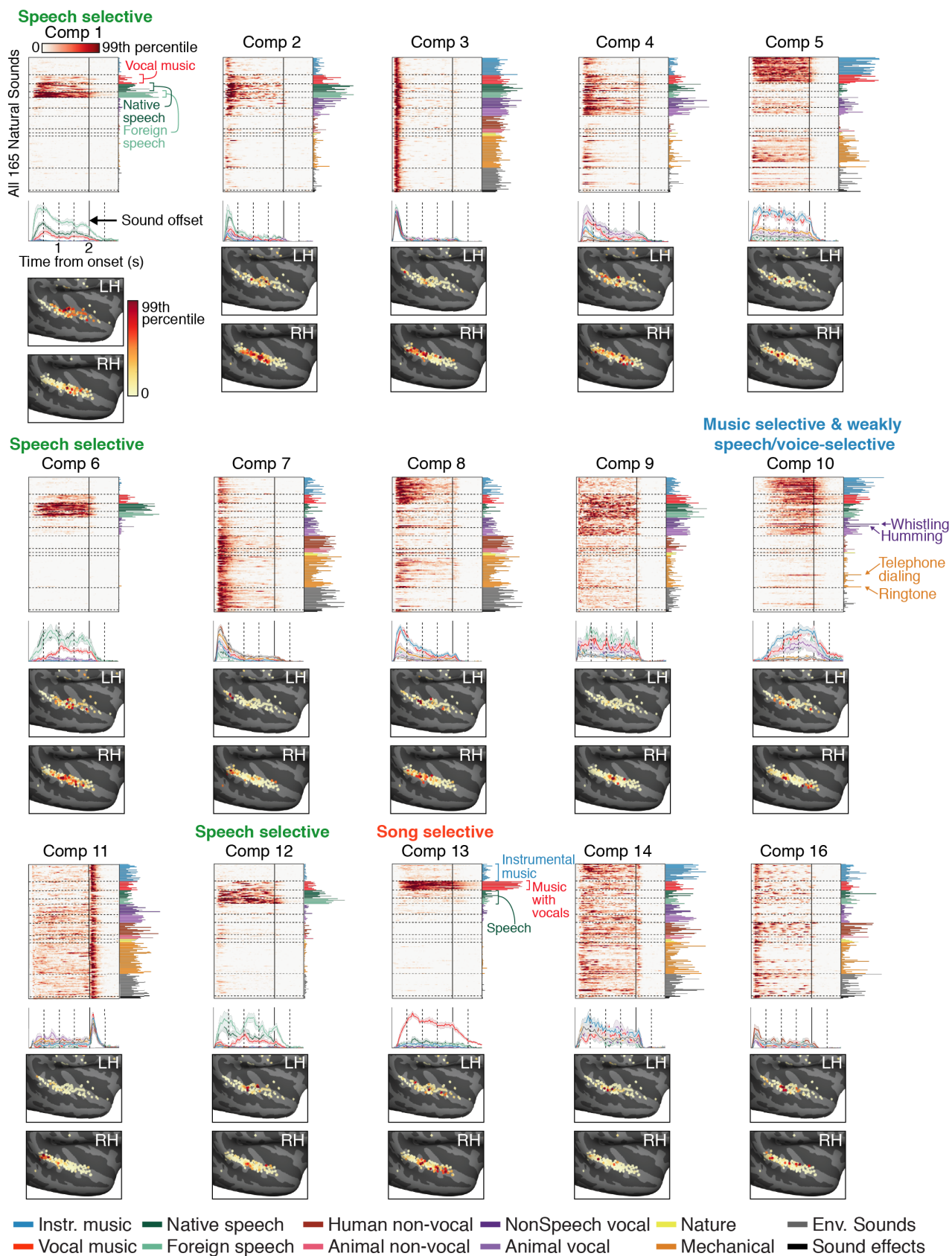


## C Model comparison



1137  
 1138 **Figure S2. Component model and its evaluation via cross-validation.** **A**, Schematic of the “sparse and smooth”  
 1139 component model, which was motivated by the statistical properties shown in **Fig S1**. Each electrode was represented  
 1140 by its response timecourse (broadband gamma) across all sounds (measured relative to silence). This timecourse was  
 1141 modeled as the weighted sum of multiple component timecourses to capture the fact that each electrode is influenced  
 1142 by many neurons and thus might reflect multiple underlying neuronal populations. The component response timecourses  
 1143 were the same across electrodes, but the weights varied to account for different response patterns. Both the component  
 1144 responses and weights were constrained to be positive. To encourage the component response patterns to be sparse  
 1145 and skewed, we modeled each component as the convolution of a set of sparse activations with a smoothing kernel.  
 1146 The activations, weights and smoothing kernel were all learned by minimizing a cost function with two terms: (1) a  
 1147 reconstruction penalty encouraging the components to closely approximate the data; and (2) a sparsity penalty  
 1148 encouraging the activations and weights to be sparse. The smoothing kernel was learned separately for each component  
 1149 to account for variable levels of smoothness in the responses across electrodes. **B**, Average squared correlation between

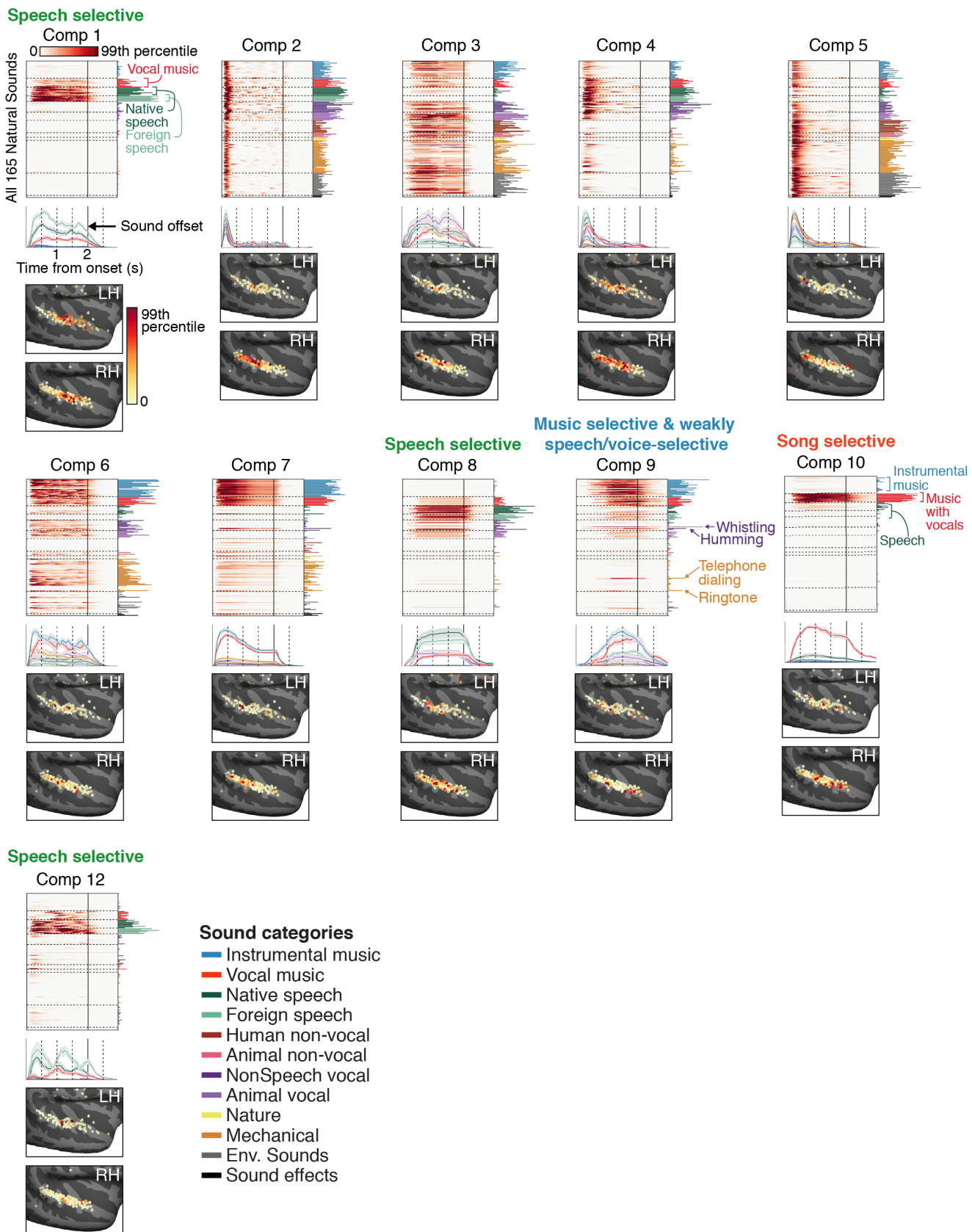
1150 the measured and model-predicted response in test data as a function of the number of components and sparsity penalty  
1151 (the correlation has been noise-corrected; **Fig 1E** shows results for the best sparsity parameter ( $\lambda = 0.33$ )). **C**,  
1152 Comparison of the prediction accuracy (average correlation in test data) of the SSC model with several baseline models  
1153 that did not rely on the convolutional decomposition used by the SSC model: (1) non-negative matrix factorization (NMF)  
1154 where the components and weights were constrained only to be positive; (2) NMF with a sparsity penalty applied directly  
1155 to the responses and weights; (3) NMF with a L2 smoothness penalty applied to the derivative (first-order difference) of  
1156 the component responses; and (4) NMF with both an L1 sparsity and L2 smoothness penalty. Data from independent  
1157 subjects was used to select the hyper-parameters for each model and evaluate prediction accuracy. Error bars show the  
1158 median and central 68 percent of the sampling distribution measured via bootstrapping across subjects.



1159  
1160  
1161  
1162  
1163  
1164  
1165

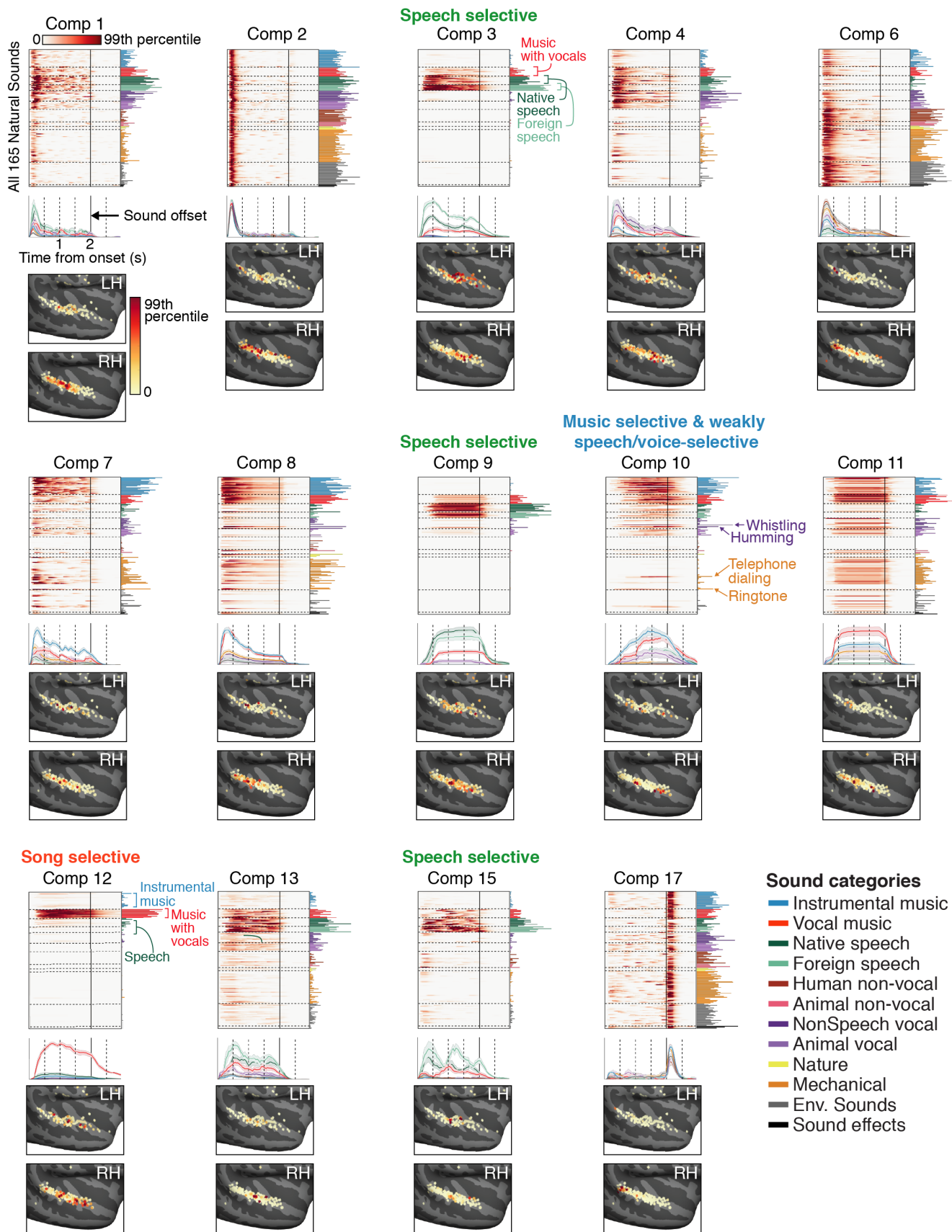
**Figure S3. Components from non-negative matrix factorization (NMF) model.** Component responses and weights from a model that only imposed non-negativity on the responses/weights. Conventions the same as **Fig 2&S5** which show components from the SSC model (which had the best prediction accuracy). 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.





1166  
1167  
1168  
1169  
1170  
1171

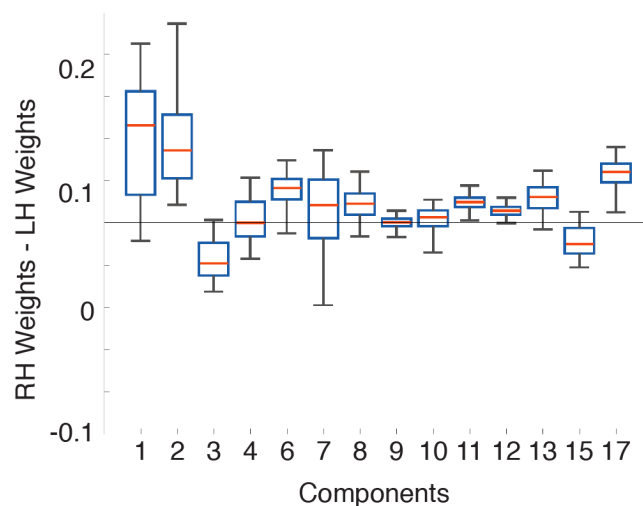
**Figure S4. Results from 15-component model.** Component responses and weights from a model with only 15 components. Conventions the same as **Fig 2&S5**, which show results from a 20-component model. We focus on components that were consistent across subjects and reliable across random re-initializations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



1172  
1173  
1174  
1175

**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.

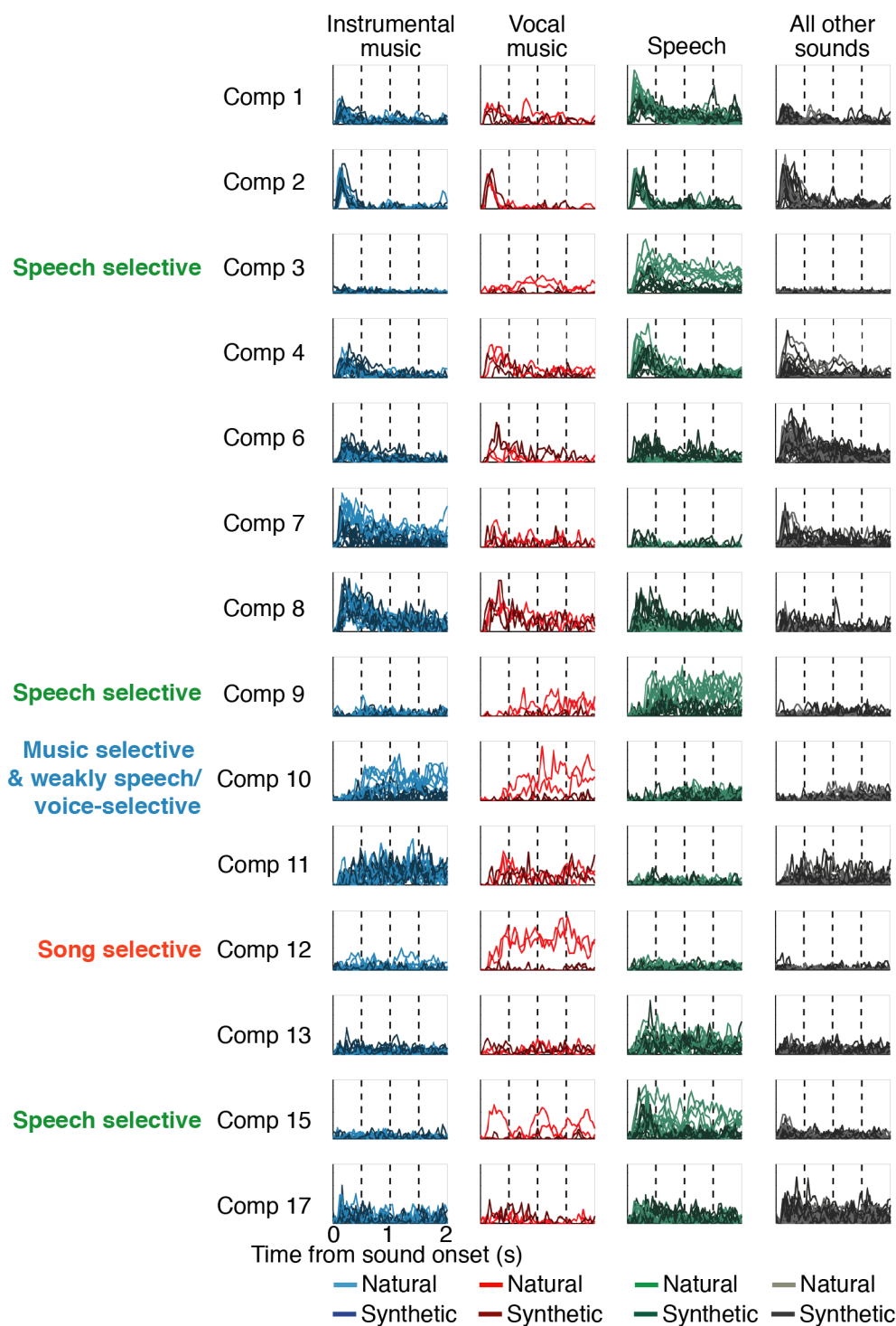
## Laterality of component electrode weights



1176  
1177  
1178  
1179  
1180

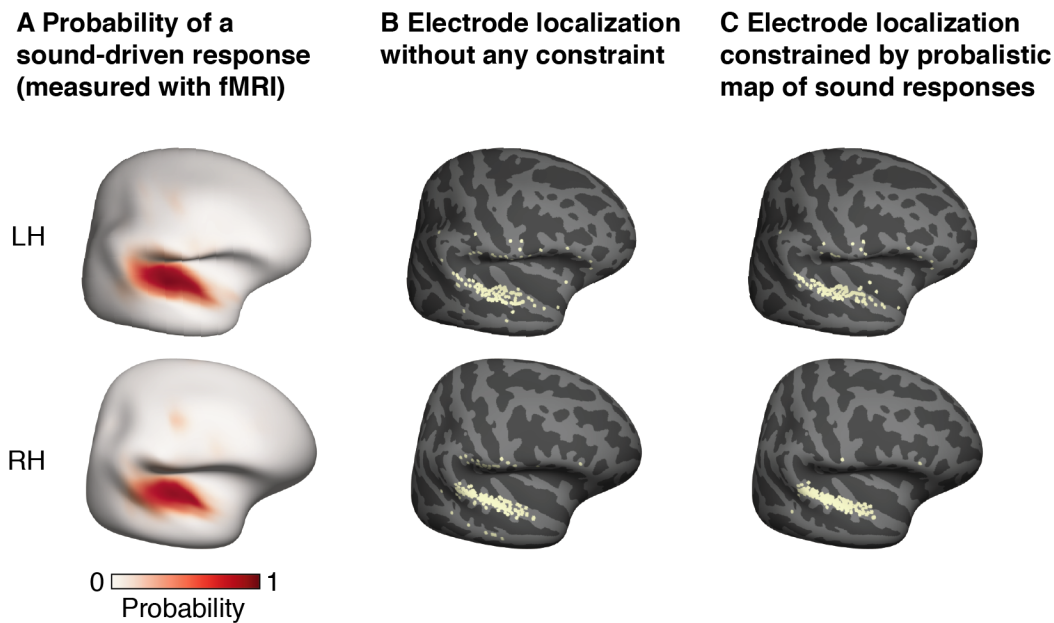
**Figure S6. Laterality of component electrode weights.** For each reliable component, we plot the average difference in the electrode weights between the right and left hemisphere. Bootstrapping across subjects was used to estimate the sampling distribution for each component. Boxes show the central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

## Response of all components to natural and modulation-matched synthetic sounds



1181  
1182  
1183

**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.

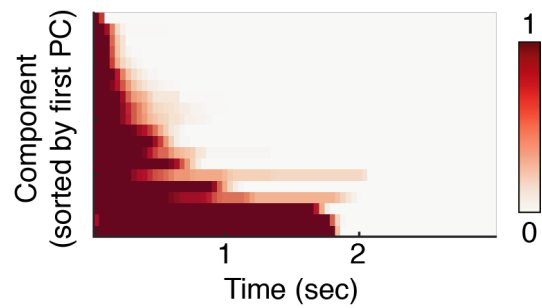


1184  
1185  
1186  
1187  
1188  
1189  
1190  
1191  
1192  
1193  
1194

**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 it to 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.



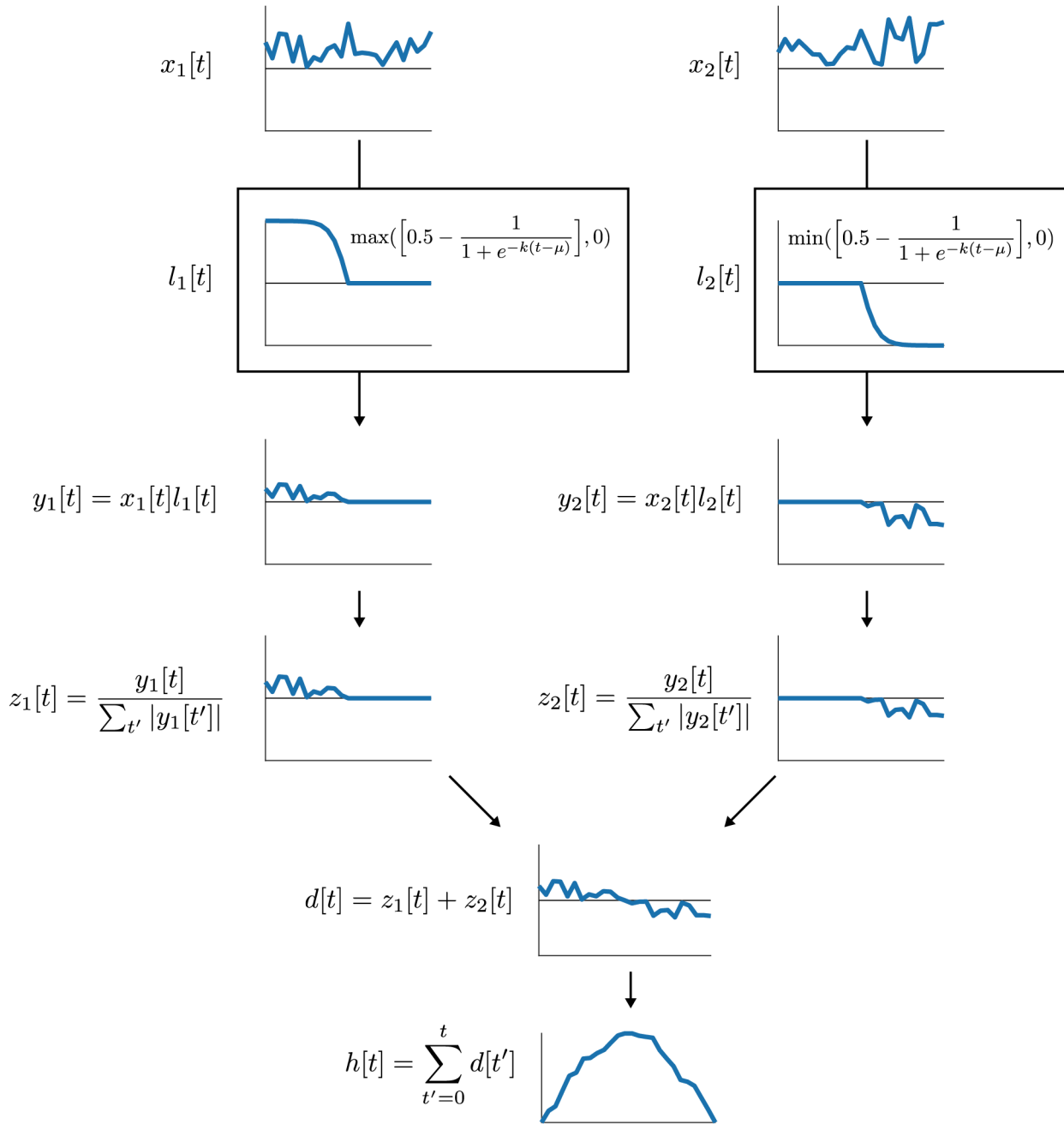
## Learned smoothing kernels



1195  
1196  
1197  
1198  
1199  
1200

**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 were also asymmetric with a fast/instantaneous rise and a slower falloff.

## Constraining the smoothing kernel to be unimodal



1201  
 1202 **Figure S10. Constraining the smoothing kernel to be unimodal.** This plot describes the set of operations  
 1203 (implemented in TensorFlow) that was used to constrain the smoothing kernel to be unimodal. Conceptually, the goal of  
 1204 these operations is to force the derivative to be exclusively positive for the first N time-points and then exclusively  
 1205 negative for the rest of the signal, thus preventing oscillations. We also must force the sum of the derivative to equal  
 1206 zero so that the kernel starts and ends at zero. Two positive vectors (themselves computed as the absolute value of  
 1207 real-valued vectors) were multiplied by a positively or negatively rectified logistic function with the same cross-over point.  
 1208 As a consequence, the first vector has positive values at the start of the signal, followed by zeros, and the second vector  
 1209 has negative values at the end of the signal, preceded by zeros. The two vectors were then normalized so that they sum  
 1210 to 1/-1. Finally, the two vectors were added and cumulatively summed, yielding a unimodal signal. The shape of the  
 1211 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  
 1212 ( $\mu$  and  $k$ ), all of which were learned. The input vectors were initialized with a vector of ones.  $\mu$  was initialized to the value  
 1213 of the middle timepoint, and  $k$  was initialized to the value of 1 (and prevented from taking a value less than 0.001).