Title: **EEG-based classification of natural sounds reveals specialized responses to speech and music**

Abbreviated Title: Enhanced neural responses to speech and music

Authors: Nathaniel J Zuk[1], Emily S Teoh[1,2], Edmund C Lalor[1,2,3,4]

Affiliations: 1) Department of Electronic & Electrical Engineering and Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland; 2) Trinity Centre for Bioengineering, Trinity College Dublin, Dublin 2, Ireland; 3) Department of Biomedical Engineering, University of Rochester, New York 14627, USA; 4) Del Monte Institute of Neuroscience, University of Rochester Medical Center, University of Rochester, New York 14627, USA

Corresponding Author: Nathaniel J Zuk, zukn@tcd.ie

Number of Pages: 24

Number of Figures: 5

Number of words in abstract: 242

Number of words in significance statement: 108

Number of words in introduction: 601

Number of words in discussion: 1090

Conflicts of interest: None

Acknowledgments: The authors would like to thank Sam Norman-Haignere and Josh McDermott for providing the stimuli of natural sounds used in our experiments, and Lauren Szymula for collecting EEG data in Experiment 2. ET was funded by an Irish Research Council Government of Ireland Postgraduate Scholarship (GOIPG/2015/3378), NZ and EL were supported by the Del Monte Institute for Neuroscience at the University of Rochester and an SFI Career Development Award (CDA/15/3316).

Abstract:

Humans can easily distinguish many sounds in the environment, but speech and music are uniquely important. Previous studies, mostly using functional magnetic resonance imaging, have identified separate regions of the brain that respond selectively for speech and music. Yet there is little evidence that brain responses are larger and more temporally precise for human-specific sounds like speech and music, as has been found for responses to species-specific sounds in other animals. We recorded EEG as healthy, adult subjects listened to various types of two-second-long natural sounds. By classifying each sound based on the EEG response, we found that speech, music, and impact sounds were classified better than other natural sounds. But unlike impact sounds, the classification accuracy for speech and music dropped for synthesized sounds that have identical frequency and modulation statistics based on a subcortical model, indicating a selectivity for higher-order features in these sounds. Lastly, the trends in average power and phase consistency of the two-second EEG responses to each sound replicated the patterns of speech and music selectivity observed with classification accuracy. Together with the classification results, this suggests that the brain produces temporally individualized responses to speech and music sounds that are stronger than the responses to other natural sounds. In addition to highlighting the importance of speech and music for the human brain, the techniques used here could be a cost-effective and efficient way to study the human brain's selectivity for speech and music in other populations.

Significance Statement:

In other species, ecologically important sounds such as vocalizations generate larger responses in the auditory cortex than other types of natural sounds, but it is unclear if this is true for humans. We recorded EEG as subjects listened to various natural sounds. Classification analysis based on the EEG responses as well as analysis of the average power of the EEG responses showed increased activation in the brain to speech and music sounds compared to other natural sounds. This study demonstrates the human brain's selectivity to speech and music sounds and suggests a cost-effective and efficient way to study the brain's selectivity for speech and music in other populations.

Introduction:

Sound carries a tremendous amount of information about the world around us, and we can easily decipher an acoustic scene. It is clear, though, that speech and music are especially important for humans. They are present in all cultures with a striking amount of diversity (Pinker and Bloom, 1990; Evans and Levinson, 2009; Savage et al., 2015; Youn et al., 2016; Mehr et al., forthcoming). Given their importance, it seems pertinent that our brain should be able to more easily identify these types of sounds relative to other sounds in the natural world. Relatedly, many studies have demonstrated neural specializations to ecologically relevant sounds in other species, particularly to vocalizations and song (King and Nelken, 2009; Theunissen and Elie, 2014). Since speech and music are arguably the most characteristically human of sounds, it is likely that there exist neural specializations in the human brain for them.

Indeed, there is common agreement that specialized regions of the brain responsive to speech and music sounds do exist (Zatorre et al., 2004; Staeren et al., 2009; Leaver and Rauschecker, 2010; Giordano et al., 2014; Norman-Haignere et al., 2015; Hausfeld et al., 2018; Ogg et al., 2019). These regions appear to be in secondary auditory cortex, which also respond to features beyond those easily captured by spectrotemporal statistics (Norman-Haignere et al., 2015; Kell et al., 2018). However, most of the studies examining specializations for speech and music have used fMRI, and through these studies it is unclear if there are differences in the time course of the responses to speech and music. More recently, an ECoG study showed activation across auditory cortex for speech sounds, and very strong spatially localized neural activation for vocal music in particular (Norman-Haignere et al., 2019). Spatial regions appeared distinct for these activations, but both showed temporally sustained responses throughout the two second stimulus. A few studies have used EEG to compare neural responses to speech and musical instrument sounds (Murray et al., 2006; Cossy et al., 2014) but these EEG studies have not validated if the differences in neural responses are simply due spectrotemporal differences in the sounds, so it unclear if these responses are unique for speech and music (see Giordano et al., 2014; Norman-Haignere et al., 2015, 2019).

Additionally, there is reason to believe that neural responses to speech and music consistently time-lock to particular acoustic features in these sounds. Several studies have shown greater responses or more consistent patterns of activation to speech sounds than various control sounds where the speech is degraded (Ahissar et al., 2001; Luo and Poeppel, 2007; Okada et al., 2010;

Peelle et al., 2013; Zoefel et al., 2018; Nourski et al., 2019) and a similar study found consistent patterns of activation for piano music (Doelling and Poeppel, 2015). Yet auditory cortical neurons can also time-lock to amplitude and frequency modulations in synthetic sounds (deCharms et al., 1998; Liang et al., 2002; Lalor et al., 2009; for review see Joris et al., 2004), and it isn't clear from previous work if this time-locking is comparatively strong for speech and music.

Here we use classification-based analyses with EEG to identify specialized neural responses to speech and music sounds. Furthermore, unlike other natural sounds that evoke unique temporal responses (such as impact sounds), EEG responds less to sounds that have the same acoustic characteristics as speech and music based on a model of subcortical auditory processing (McDermott and Simoncelli, 2011). Our results demonstrate that the specialization of the brain to speech and music sounds can be observed with EEG, where EEG responses are larger and more time-locked than to other natural or spectrotemporally-matched sounds.

Methods:

Experimental paradigm and EEG preprocessing: This study involved two experiments. In both Experiment 1 and 2, 128 channels of scalp EEG data were recorded, along with two additional channels over the mastoid processes using a Biosemi Active Two system at a sampling rate of 512 Hz. In Experiment 1, six subjects (3 female, ages 24-30) participated. Stimuli consisted of 30 different two-second-long sounds that produced the strongest fMRI responses in the independent components of auditory cortical activity found in a previous study (Norman-Haignere et al., 2015). Of the six speech stimuli used in this experiment, four were unintelligible or in a foreign language, and all music stimuli were instrumental (see Figure 2 for the list of stimuli). In each trial, the sounds were presented consecutively such that each sound was repeated at least twice during a trial. Five of those sounds were presented a third time immediately following a presentation of the same sound. During the trial, subjects were asked to detect the consecutive repeats (a "one-back" task) by hitting the spacebar. In total, there were 65 sounds per trial, lasting 2 minutes and 10 seconds. Subjects listened to 43-50 trials.

In Experiment 2, 15 subjects participated (7 female, ages 19-31). Stimuli consisted of a subset of five speech and five music sounds from Experiment 1. Additionally, five "impact sounds" were included based on high within-frequency-channel real modulation correlations as determined by

115 a spectrotemporal model of auditory processing (McDermott and Simoncelli, 2011), all of which
116 were sampled from a larger database of natural sounds (Norman-Haignere et al., 2015). The
117 impact sounds were included because they exhibited high classification accuracies in Experiment
118 1, which were most likely due to the strong, sparse onsets present in these sounds. We also
119 included synthesized versions of these 15 sounds that contained identical time-averaged sound
120 frequency and modulation statistics to the originals (McDermott and Simoncelli, 2011). These
121 were included in order to test if the results observed in Experiment 1 were due to some neural
122 selectivity for the spectrotemporal statistics of the sounds. As in Experiment 1, each sound was
123 repeated twice in a trial, but subjects were asked to detect a sound that was identical to the sound
124 before the previous one (a "two-back" task) (**Figure 1a**). This was more difficult than the one-
125 back task used in Experiment 1, and we changed the task in order to make the experiment more
126 engaging. There were 40 trials in total.

127 *EEG data preprocessing*: The 128 scalp channels were referenced to the average of the two
128 mastoids and then filtered between 1-45 Hz using a zero-phase Chebyshev type II filter with a
129 0.75 to 60 Hz stopband at -60 dB. Eyeblinks and sparse electrical artifacts were identified using
130 runICA from the EEGLAB toolbox (Delorme and Makeig, 2004) in Experiment 1 or the fastICA
131 algorithm (Hyvärinen and Oja, 2000) in Experiment 2 and removed. For the data in Experiment
132 1, channels with standard deviations less than one-third or greater than three times the average
133 standard deviation across all channels were interpolated using spherical spline interpolation
134 (Perrin et al., 1989). For the data in Experiment 2, channels with variance greater than three
135 times the interquartile range plus the median across all 128 electrodes were replaced with a
136 signal equal to the average of the three closest electrodes weighted by the inverse of their
137 distance from the electrode. Despite the differences in the preprocessing, we found comparable
138 results for both experiments (see Results).

139 *Classification analysis*: Each trial of EEG data was spliced into the two-second segments
140 recorded during the sound presentations. Clips that occurred during a target (i.e., repeated)
141 sound were removed from further analysis, resulting in 86-100 EEG clips per stimulus for
142 Experiment 1 and 73-80 EEG clips per stimulus for Experiment 2. In order to reduce the
143 dimensionality of the data, since there were far more dimensions (128 EEG channels x 256 time
144 samples) than data samples for classification, principal components analysis (PCA) was first
145 applied to the spatiotemporal EEG responses for each sound clip and components capturing 95%

146 of the variance of the data (490-1519 components out of 32768 total dimensions) were retained.
147 The stimulus presented during the two second interval was then identified based on the principal
148 components of the EEG data using multi-class linear discriminant analysis with regularization
149 (*fitcdiscr* in Matlab). 75% of the data were randomly selected for training the classifier and the
150 classifier was tested on the remaining 25% of the data. This selection was repeated 100 times to
151 get 100 classification accuracies for the 30 different stimuli. The classification analysis was
152 done separately for each subject. In order to compare classification accuracies across subjects
153 for population analysis, classification accuracies (along the diagonal on the right in **Figure 1b**)
154 were ranked based on the average classification accuracy across the 100 training-testing repeats.

155 In order to understand how classification accuracy varied over time, the classification analysis
156 was performed on 200 ms of the EEG response spaced every 100 ms. Because the
157 dimensionality of each 200 ms of EEG data was still larger than the number of samples (128
158 EEG channels x 26 time samples = 3328 dimensions), PCA was applied to each 200 ms segment
159 of EEG and only the principal components encompassing 95% of the variance of the data were
160 retained for classification (48-496 components).

161 Similarly, to evaluate the contributions of each channel to classification performance, the
162 classification analysis was repeated on each individual EEG channel (256 time samples). To
163 make the procedure comparable to the other analyses, PCA was performed on the data prior to
164 training and testing, but all of the dimensions of the data were retained (256 dimensions).

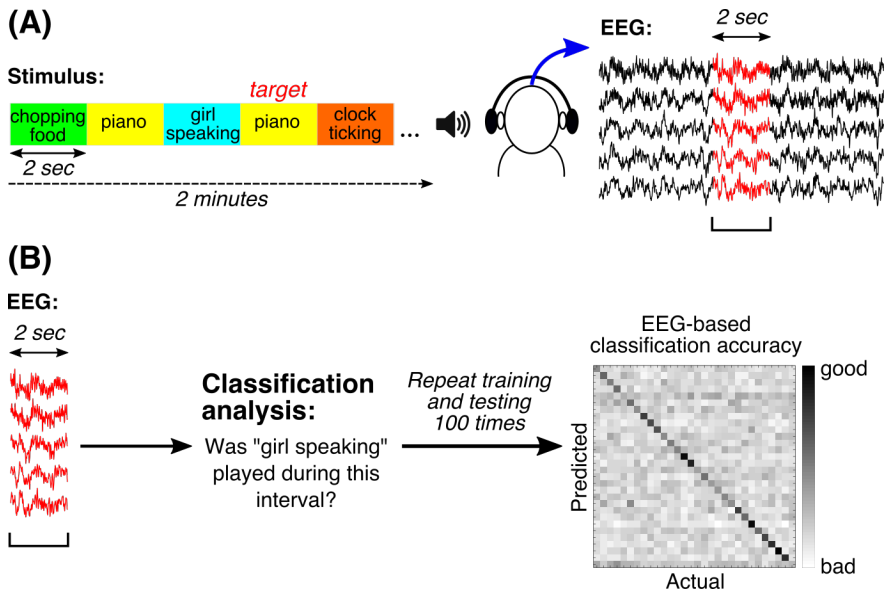

Figure 1: (A) Experiment setup: Stimuli consisted of consecutive two-second sound clips. Subjects were asked to detect repeated sound clips in Experiment 1 (not shown) or detect if a sound clip was identical to the sound clip before the previous one (shown). EEG was recorded and segmented into the corresponding two-second clips after preprocessing. (B) A linear classifier was trained to identify the sound clip presented based on the two-second EEG segments. 75% of the data was used for training and 25% was used for testing. The data was randomly sampled 100 times, and a classification accuracy matrix was constructed based on the average classification accuracy across the 100 repeats of training and testing. For most of the population analyses, the classification accuracies along the diagonal, indicating correct classification, were analyzed.

Speech/music discrimination analysis: Similar to the classification of individual sounds, we also tested how well sounds could be classified as speech or music based on the EEG responses. For each subject, using only the responses to either speech or music sounds, we retained the principal components that captured 95% of the variance in the responses (290-829 dimensions). We then trained a linear discriminant classifier to identify if a response came from a music sound or a speech sound. Crucially, the classifier was trained on the EEG data from the responses to 5 of the 6 speech and music sounds in Experiment 1 and 4 of the 5 speech and music sounds in Experiment 2. Then the classifier was tested on the responses to the remaining pair of sounds, resulting in 36 iterations for Experiment 1 and 25 iterations for Experiment 2. This test ensured that the classifier was identifying speech or music sounds based on a general pattern of the EEG response instead of individualized responses to specific sounds.

Evoked response, global field power (GFP), and phase dissimilarity analyses: In addition to the classification analyses, we examined the median evoked response to each stimulus type across the presentations of non-target stimuli for each subject. As a summary measure of the time course of the magnitude of the evoked response across the scalp, we also examined the global field power (GFP), which is the standard deviation of the evoked response across channels at

each time point (Lehmann and Skrandies, 1980). In order to evaluate the contribution of response magnitude to the pattern of classification results, for each subject, the GFP was averaged over each two-second interval, and the average GFPs were ranked.

Additionally, we expected that speech and music may evoke temporally consistent responses and, relatedly, consistent response phases over multiple repetitions of the same stimulus, as has been shown by prior work (Luo and Poeppel, 2007; Doelling and Poeppel, 2015). We computed the phase dissimilarity by computing the difference between the phase coherence across trials for a single stimulus and the average phase coherence after sampling trials from all 30 different stimuli. Larger values of phase dissimilarity imply temporally consistent responses for a stimulus from trial to trial. First, phase dissimilarity was computed using 4 Hz frequency bands spaced from 0 to 40 Hz in order to identify the frequency range of the EEG producing the most phase consistency across all stimuli and all subjects. Then we ranked the stimulus phase dissimilarities for each subject, like the analysis for classification accuracy and average GFP, and examined differences in phase dissimilarity rankings across.

Results:

Experiment 1: Nearly all stimuli were classified significantly better than chance for all six subjects; at most 5/30 stimuli failed to reach threshold performance (0.045, which is the 95% criterion for a binomial test with $p = 0.033$) for each subject (**Figure 2a**). However, there appeared to be a considerable amount of variation in the accuracies for correct classification, indicating that some stimuli were easier to classify than others (**Figure 2a**). To quantify the relative magnitude of the classification accuracies for the different types of stimuli, the accuracies for correct classification (along the diagonal of the classification matrix in **Figure 2a**, for example) were ranked. Across the six subjects in Experiment 1, speech, music, and "non-vocal human" sounds were classified significantly better than all other natural sounds (**Figure 2b & c**, significance in **Figure 2c** was evaluated using the Dunn & Šidák approach for multiple comparisons, $p < 0.05$).

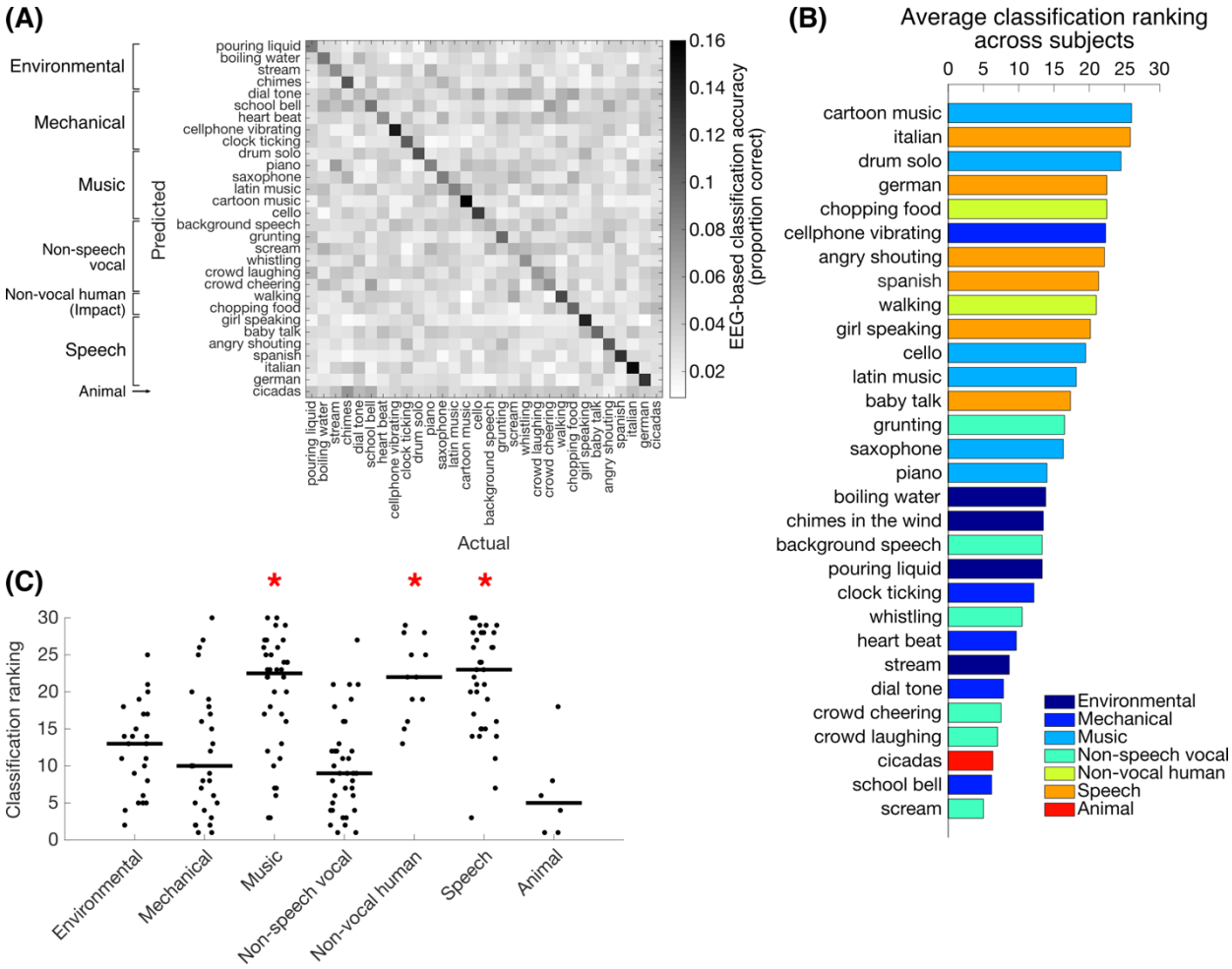

Figure 2: (A) Classification accuracy for an example subject in Experiment 1. Only "cicadas" was classified below chance (where chance is 0.044 based on the 95% confidence interval from a binomial test with p = 0.033). (B) Ranking of classification accuracy (along the diagonal of the classification matrix in A), averaged across subjects. Music, speech, and "non-vocal human" sounds tend to be classified best. (C) Classification rankings for all types of stimuli across 6 subjects recorded in Experiment 1. Red asterisks indicate that the music, speech, and non-vocal human stimuli were classified significantly better than all other stimuli analyzed ($p < 0.05$, Dunn & Šidák approach for multiple comparisons).

The "non-vocal human" sounds were two impact sounds characterized by sharp acoustic transients ("chopping food", "walking on a hard surface"). Using a model that quantified the statistics of the signals within audio frequency and modulation bands based on the physiological stages in subcortical processing (McDermott and Simoncelli, 2011), we identified that these sounds had high within-frequency-channel modulation correlations, for which the real values were notably higher for these two sounds than the other sounds in the dataset. Positive real within-frequency-channel modulation correlations are characteristic of sounds with sharp attacks (McDermott and Simoncelli, 2011). As a result, we hypothesized that these sounds probably evoked transient responses that might have contributed to their increased classification accuracy. In contrast, prior work has shown that speech and music selectivity is sensitive to acoustic

features beyond the spectrotemporal statistics captured by this model (Norman-Haignere et al., 2015, 2019; Kell et al., 2018).

Experiment 2: To validate if the improved classification accuracy for speech, music, and the impact sounds was a result of the statistics captured by the subcortical model, we ran Experiment 2 with speech, music, and impact sounds, as well as model-matched sounds that were generated to have the same frequency and modulation statistics (McDermott and Simoncelli, 2011) (**Figure 3a**). For sounds where the statistics are invariant with time, previous work has shown that model-matched sounds are not discriminable from the originals (McDermott et al., 2013), indicating that the model captures statistics that are sufficient for human perception for some sounds. We then repeated the classification analysis (**Figure 3b**). As in Experiment 1, nearly all stimuli were classified above chance; at most 3/30 stimuli did not pass threshold for each subject. Across all 15 subjects that were recorded during Experiment 2, we found that the model-matched speech and music sounds were classified significantly worse than the originals, while the model-matched impact sounds were classified no differently than the originals (**Figure 3c**). This shows that the neural responses to speech and music producing high classification accuracies were responsive to acoustic features beyond the statistics captured by the model, and therefore cannot be explained by neural specificity to frequency and modulation alone.

Given that previous fMRI work has shown different spatial patterns of selectivity in auditory cortex for speech and music, we also wished to examine the topographies of our EEG classification results for speech and music channel by channel. For all types of the original sounds, classification accuracies were highest in frontal channels. The topography of these classification accuracies may indicate auditory cortical activity (Lalor et al., 2009), but it may also include frontal cortical activity (**Figure 3e**). To test significance of the channel-wise classification accuracies, we focused on 20/128 channels that had the most above-chance classification accuracies for all stimuli and subjects (**Figure 3d**) and compared the classification accuracies averaged across these channels. Both speech and music had significantly better classification accuracies for the original sounds than their model-matched counterparts, while the impact sounds showed no significance (Wilcoxon rank-sum test with Bonferroni correction, significance values are shown in **Figure 3e**).

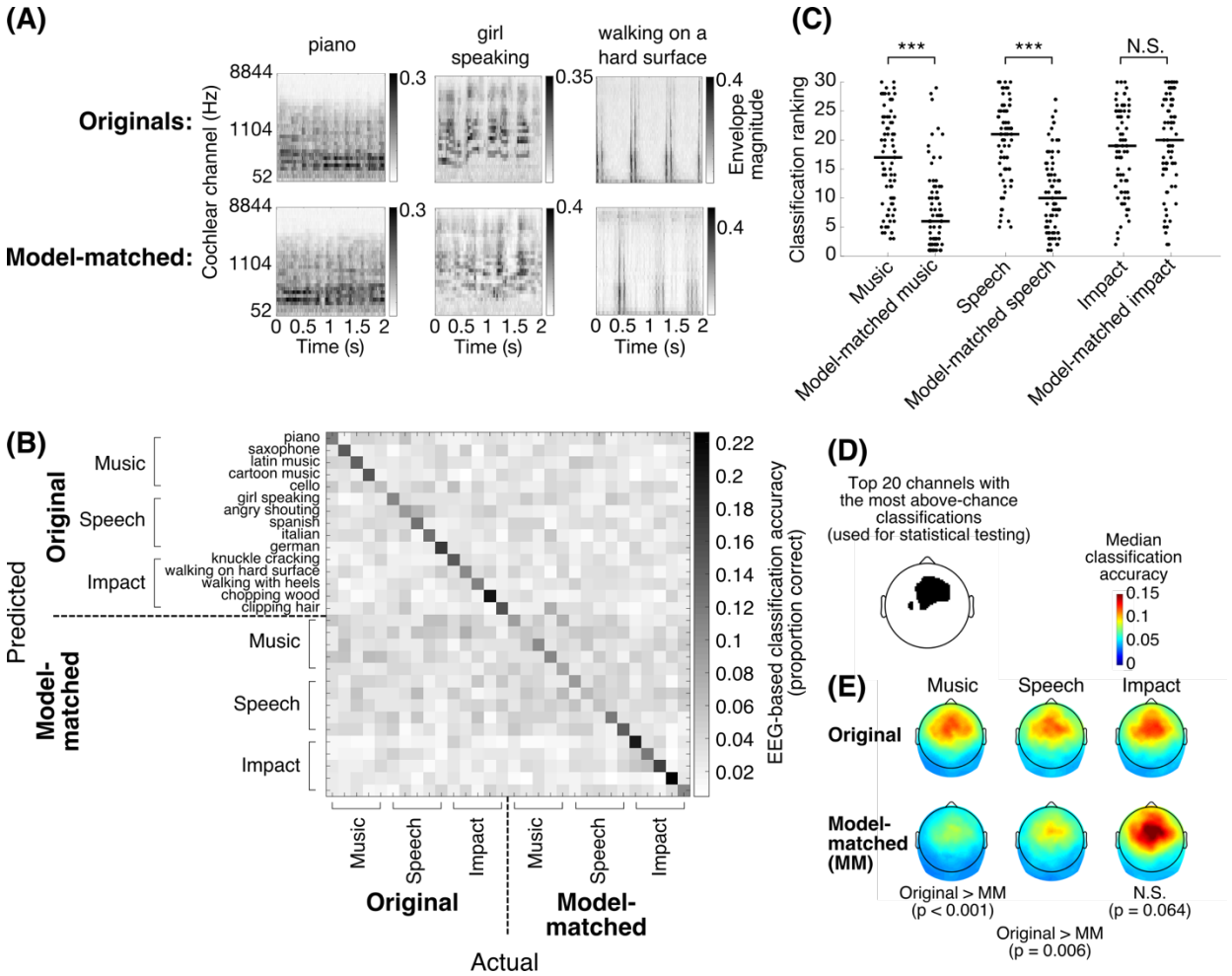

Figure 3: (A) Example cochleograms (see McDermott and Simoncelli, 2011) of the original and model-matched stimulus for an example music ("piano"), speech ("girl speaking"), and impact ("walking on a hard surface") sound. (B) Classification accuracies for the original and model-matched stimuli for an example subject. All stimuli were classified significantly better than chance (0.047 based on the 95% confidence interval from a binomial test with $p = 0.033$). The model-matched music and speech had notably worse classification accuracy (along the diagonal) than the originals. (C) Classification rankings for original and model-matched stimuli in Experiment 2 for all 15 subjects. There is a significant reduction in classification ranking for the model-matched music and speech stimuli relative to the originals (Wilcoxon rank-sum with Bonferroni correction: $p < 0.001$ for both), but no significant difference for the impact sounds. (D) Black designates the top 20 channels that had the most above-chance classifications ($p < 0.001$ based on a Binomial test) across all stimuli and subjects (377-412 out of 450 classifications). (E) Median classification accuracies across subjects and stimuli for the original and model-matched music, speech, and impact sounds. The significance of the difference between the classification accuracies averaged across the region in (D) is displayed below each column of original and model-matched topographies. Significance was determined with a Wilcoxon rank-sum test with Bonferonni correction. Music and speech show significantly better classifications than their model-matched counterparts, while the impact sounds do not.

Time course of classification accuracy: Next, we repeated the analysis of the EEG data in Experiment 2 using 200 ms blocks of time space every 100 ms. This allowed us to examine how classification accuracies and confusion between the stimuli change over the course of the stimulus (**Figure 4a**). Prior work has shown that confusion between images (Cichy et al., 2014; Marti et al., 2015) and phonemes (Khalighinejad et al., 2017) varies with time, indicating different stages of processing for the stimuli. However, for the stimuli used in Experiment 2 we

found no confusion between classes of stimuli (**Figure 4a**, black line). This could be due to unique temporal responses for the different stimuli.

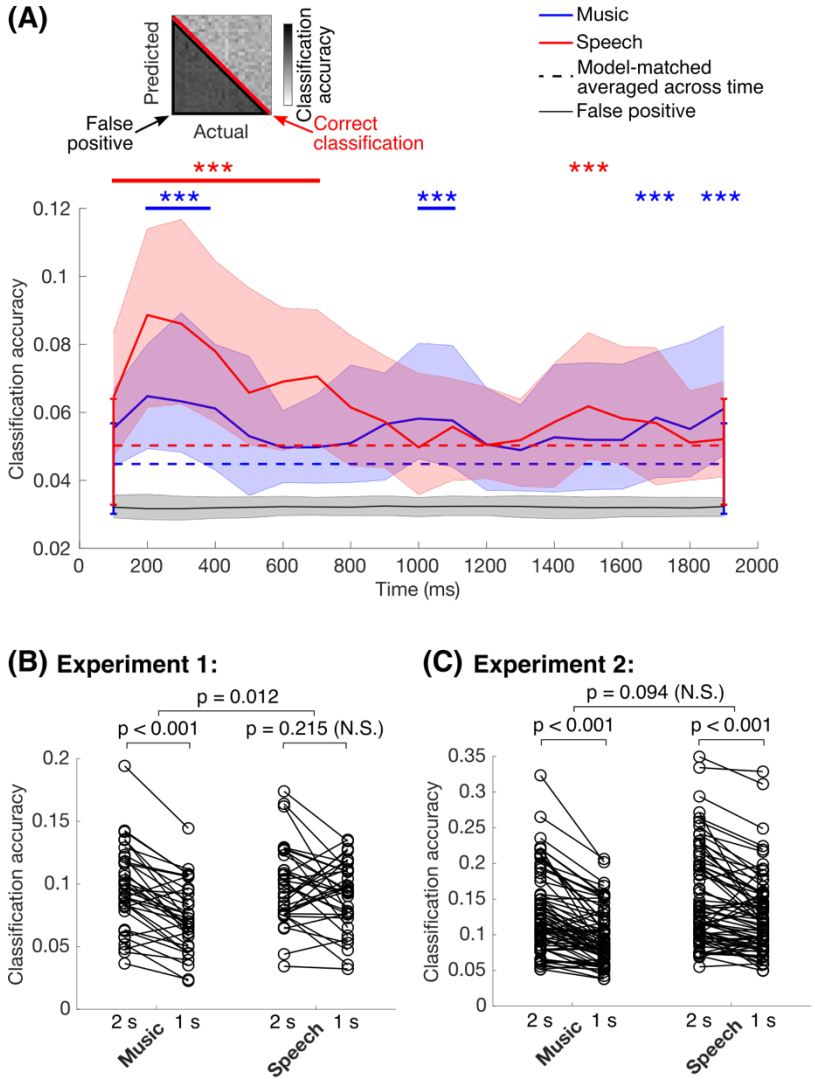

Figure 4: (A) Classification accuracy using 200 ms segments of the EEG response sampled at different points throughout the two-second stimulus was averaged across 15 subjects. We examined the accuracy for correct classification for the different stimulus types (red line), as before, and the off-diagonal classification accuracies indicating the false positive rate (black shaded region). We expected confusion between stimuli, which would result in above-chance false positives. Classification accuracy of the speech (red) and music (blue) stimuli. The solid lines indicate the median values and the shaded regions indicate the interquartile range across all stimuli and subjects. The dashed lines indicate the median and interquartile range of the model-matched stimuli for the speech (red) and music (blue). Above, asterisks label 200 ms intervals where the classification accuracy of the original stimuli exceeds the average model matched accuracy (Wilcoxon rank-sum test, FDR with Benjamini-Yekutieli procedure, $p < 0.001$). The black line and grey region show the median and interquartile range respectively for the false positives. We did not observe confusion between the different stimuli, and the median false positive rate was always slightly but significantly below chance (Wilcoxon signed-rank test relative to chance, with chance probability = 0.033). (C) and (D) show the classification accuracies using the data from Experiment 1 and 2 respectively, when the full two-second response is included ("2 s") and when only the first second of the response is used ("1 s"). Comparisons between 2s and 1s use Wilcoxon signed-rank test, and comparisons between stimulus types use the rank-sum test. Classification accuracy tend to decrease for all stimuli when only the first second of the response is used (this does not reach significance for speech in Experiment 1), but the change is a slightly larger change in accuracy for music than for speech. The difference in accuracy reaches significance for Experiment 1, but it doesn't reach significance for Experiment 2.

302 We then compared the classification accuracies for the original speech and music to the average
303 classification accuracies of the model-matched versions. Classification of speech was
304 significantly better than its model-matched counterpart mainly between 100-700 ms after
305 stimulus onset, and the classification at 1150 ms was also significant (**Figure 4a**). Music
306 similarly showed significantly better classification within the early time range (specifically 200-
307 400 ms), but it also continued to show significantly better classification 1000-1100 ms, 1700 ms,
308 and 1900 ms after stimulus onset.

309 The time points exhibiting significance beyond one second were short, and it was not entirely
310 clear how important these later response times are for representing the stimulus. To validate this,
311 we repeated the classification analysis in Experiments 1 and 2 using only the first second of the
312 response to the stimuli. We expected that the classification accuracy would decrease for both
313 music and speech stimuli when the last second of the response was not included simply because
314 of the reduction of information available for classification, but we hypothesized that this
315 decrease would be larger for music than for speech if the time points after 1000 ms were
316 important. We found a weak but significantly larger decrease for music than for speech in
317 Experiment 1 (**Figure 4b**), and we found a similar pattern in the data for Experiment 2 that did
318 not reach significance (**Figure 4c**). Thus, the response after 1000 ms may be carrying more
319 information for music than for speech, but the contribution appears to be very weak.

320 *EEG-based discrimination between speech and music*: Using both sets of data from Experiments
321 1 and 2, we created classifiers that identified whether a sound clip was speech or music based on
322 the EEG response. Importantly, the classifier was tested on EEG responses for stimuli that were
323 not included in training. This was to ensure that the classifier was not relying on stimulus-
324 specific temporal responses that might not be a reliable representation of more general responses
325 to speech or music. For both sets of data, speech/music discrimination was not as successful as
326 classification of individual sounds: for Experiment 1 stimuli were identified as "speech" above
327 chance for 1/6 subjects, and for Experiment 2 either speech or music exhibited above chance
328 performance for only 2/15 subjects (above chance criterion was 0.56 for the 95% confidence
329 interval based on a binomial test with p=0.5). This supports our expectation that the responses
330 are temporally individualized for each sound. Additionally, it suggests that EEG does not pick
331 up the subtle spatial differences in cortical responsiveness to speech and music sounds, as has
332 been found in prior fMRI literature (Giordano et al., 2014; Norman-Haignere et al., 2015, 2019;

Ogg et al., 2019), which might be observed with many more speech and music samples in a future experiment.

Evoked response, GFP, and phase dissimilarity analyses: One possible explanation for our classification results is that music and speech elicit larger responses than other natural or model-matched sounds, because if the responses are larger, they would be easier to classify based on the EEG. To evaluate this hypothesis, we did a more standard analysis of the evoked responses to each of the different sounds: We looked at the evoked EEG response and the time-varying GFP for each of the different stimulus types (**Figure 5a & b**). Typically, these average responses would be used to identify differences in the neural responses to the different types of sounds, presuming that the responses have a consistent time course within each class of stimuli. There is a clear frontal positivity in the response for all of the different classes of sounds (**Figure 5a & b**), but there were no obvious differences in the time course for these different stimuli. This is expected if the stimulus-specific response was temporally individualized, as we found in our time-based analysis (**Figure 4a**) and our speech/music discrimination analysis.

After computing the evoked responses, we also averaged the GFP over the two second interval of each response. The average GFP values were then ranked within each subject in order to compare them across subjects, like we did for the classification accuracies. Surprisingly, the pattern of average GFP rankings mimicked the pattern of classification accuracy rankings for both experiments, albeit with slightly weaker differences between stimulus types (**Figure 5c & d**, compare to **Figure 2c** and **Figure 3c** for Experiments 1 and 2 respectively). This indicates that the differences in classification accuracy result partly from differences in the magnitude of the brain's response to the different sounds. However, the differences in response magnitudes are subtle, encompassing stimulus-specific responses throughout the full two-second interval of the stimulus, which could not be easily seen by looking at the averaged time course of the GFP.

Prior work has shown that for both speech and music neural responses have consistent low-frequency phases (below 8 Hz) indicating consistent temporal responses across repetitions of the same stimulus (Luo and Poeppel, 2007; Doelling and Poeppel, 2015). We quantified this phase consistency using phase dissimilarity, which is the difference between the within-stimulus phase coherence and the between-stimulus phase coherence (Luo and Poeppel, 2007). We first computed the phase dissimilarity for 4 Hz frequency bands and found that for both experiments the median phase dissimilarity across all stimuli and subjects was largest between 4-8 Hz. The

phase dissimilarities within this frequency band were then ranked. Like the result using average GFP, the phase dissimilarity generally produced a similar pattern of rankings as the classification accuracies for both experiments, but interestingly we found no significant difference between the phase dissimilarity rankings for the speech and model-matched speech stimuli (**Figure 5f & g**). We also validated that the median phase dissimilarity was largest for the speech and music stimuli in both experiments, so the differences in phase dissimilarity rankings we observed were not a consequence of the choice of frequency band. Thus, neural responses between 4-8 Hz appear to be more time-locked to stimulus repetitions than other natural or model-matched stimuli, particularly for music.

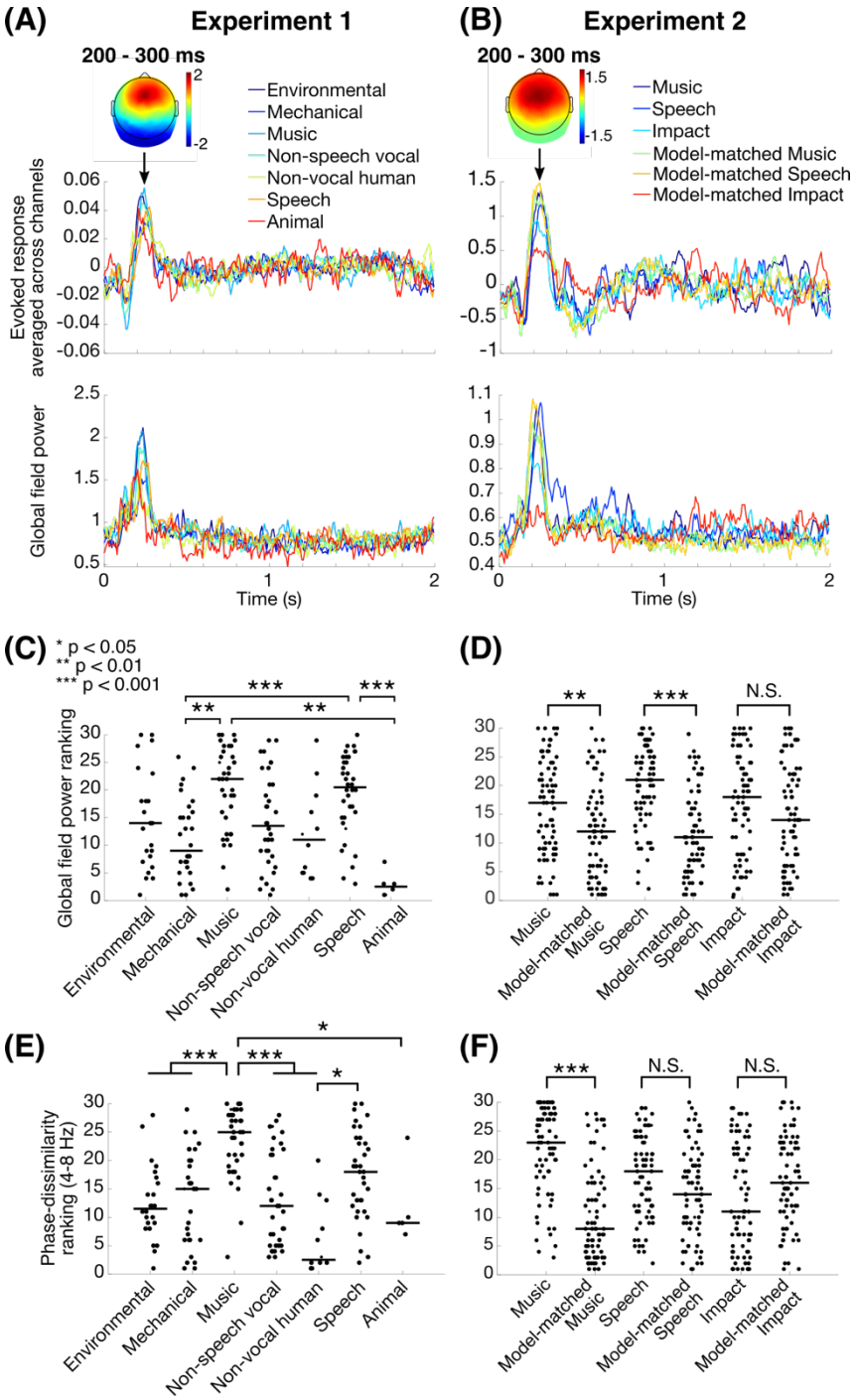

Figure 5: (A) and (B) show the evoked responses for Experiments 1 and 2 respectively. Top plots show the evoked response averaged across all EEG electrodes, where each line is the median electrode-averaged evoked response across stimuli and subjects; for clarity error bars are not shown. Bottom plots show the global field power for each stimulus type, where each line is the median across stimuli and subjects. The topographies of the evoked response between 200 – 300 ms, median values across all stimuli and subjects, are shown at the top. For each stimulus, the global field power of the evoked response was averaged over the two-second duration of the stimulus, and the average global field powers for each of the stimuli were ranked. (C) and (D) are these GFP rankings for all subjects and stimuli for Experiments 1 and 2, respectively. Despite the lack of clear temporal response differences between the evoked responses for the different stimuli, the pattern of GFP rankings are similar to those we observed based on classification accuracy, indicating that the classification accuracy is partly driven by subtle differences in the magnitude of the neural responses to these sounds. (C: asterisks based on Dunn & Šidák multiple comparisons test relative to all

comparisons, speech and music were not significantly different than any other stimuli; D: significance based on Wilcoxon rank-sum test with Bonferroni correction for the three comparisons shown). (E & F) Ranked phase dissimilarity (Luo and Poeppel, 2007) for all stimuli and subjects in Experiment 1 and Experiment 2. To rank the phase dissimilarity, we focused on the 4-8 Hz range because it produced the largest average phase dissimilarity for all stimuli and subjects. Below are the phase dissimilarity rankings for Experiment 1 (E) and Experiment 2 (F). Phase dissimilarity rankings generally showed the same pattern as classification accuracies and average GFP, but speech showed weaker phase dissimilarity relative to other natural sounds and its model-matched counterpart, while the significant differences relative to music were largely retained (significance tests for E and F are identical to the tests for those shown in C and D respectively).

Discussion:

In this study, we showed that EEG-based classification of speech and music sounds is better than the classification of other natural sounds. Additionally, unlike the responses to impact sounds, this responsiveness did not persist for model-matched stimuli generated to have identical frequency and modulation statistics (McDermott and Simoncelli, 2011). The pattern persisted, albeit more weakly, when we examined the average GFP for each of the sounds and their phase dissimilarity between 4-8 Hz. Together, these results indicate that the human brain is especially responsive to naturalistic speech and music sounds.

We also showed that there are temporal differences in the timing of neural responses for speech and music, such that classification accuracy peaked 200 ms after onset for speech but persisted sporadically throughout the stimulus for music. Furthermore, we found that the classification accuracy decreased more for music than for speech when the response from 1000-2000 ms was left out, although this effect was very weak. Both of these results are in line with recent ECoG work (Norman-Haignere et al., 2019). In particular, the component responsive to music gradually increased over the two-second duration of the stimulus, peaking just after one second (Norman-Haignere et al., 2019). Differences we observed in the time course of classification accuracy may relate to subtle differences in the temporal response of auditory cortex to speech and music sounds. Contrary to our expectations, though, we found no confusion between the sounds over time, and we were also largely unable to discriminate between speech and music sounds based on the EEG responses in either experiment. Relatedly, the short, above-model-matched classification accuracies beyond 1000 ms that we observed for music could in part be due to features in the specific stimuli we chose; this could be validated in future studies using other music stimuli. Thus, our results suggest that the specialized responses are also temporally distinct, perhaps relating to an underlying neural response to unique temporal features for these sounds. Indeed, midbrain neuron models with different synaptic parameters optimally capture vowel formant and beat-related information in speech and music respectively (Carney et al.,

2015; Zuk et al., 2018). The specializations for speech and music in the auditory cortex may have underpinnings at earlier processing stages in the auditory system.

Much of our focus was on the speech and music results, but the importance of decoding impact sounds also reflects several earlier studies. In particular, the impact sounds used in our study could all be categorized as "non-human" or "non-living" using terms from prior work (Lewis et al., 2005; Murray et al., 2006; Giordano et al., 2013; Cossy et al., 2014). There may also be a particular processing stream for these types of stimuli as well; not only are these sounds decodable with fMRI (Hjortkjær et al., 2018), but these types of sounds have been shown to evoke motor cortical areas (Lewis et al., 2005). In the absence of any clear evidence of this in the present work, a parsimonious interpretation would be to just assume that, in Experiments 1 and 2, the impact sounds were well classified because of differences in the timing of transient events that produced time-locked evoked potentials.

The spatial organization of EEG scalp electrodes that were involved in decoding these sounds tended to be frontal electrodes. The topography generated by this pattern of classification accuracy is indicative of auditory cortical activity (Lalor et al., 2009). However, we cannot completely rule out the possibility that frontal regions of cortex are involved. Frontal cortex is known to be involved in task performance (Miller and Cohen, 2001; Gold and Shadlen, 2007) and neurons in frontal cortex show a specificity for evoked sounds that are task relevant (Fritz et al., 2010). In our experiments, subjects were required to respond to repeated sounds in our experiment, so the task did require some amount of mental categorization of the sounds, which could have evoked frontal cortical regions, even for non-target sounds. One possible way of reducing the task-related activity, for example, is by presenting the sounds during a completely unrelated visual task. Even then, there is the possibility that frontal cortical neurons are already tuned to ecologically relevant sounds as a result of general experience. All of the natural sounds were recognizable and identifiable (see the behavioral study from Norman-Haignere et al., 2015). Thus, it would be very difficult to completely isolate or rule out the possibility of frontal activity involvement using EEG, and well beyond the scope of this study.

Given that midbrain-level processing of basic acoustic features and frontal cortex selectivity could both be involved in generating speech and music selectivity in auditory cortical regions, it is still an open question as to whether the selectivity we observe here is a result of “bottom-up” processing, due to auditory processing that does not require subject engagement, or “top-down”

factors that depend upon the subject's cognitive state. Disambiguating these two is tricky because specific stimuli with particular acoustic characteristics such as sparsity and roughness could engage a subject's attention (Huang and Elhilali, 2017; Zhao et al., 2019). However, we think it is unlikely that subjects were actively attending to speech and music more than other sounds in the stimulus set because the behavioral task places no special emphasis on these stimuli and all stimuli likely became less interesting over many repeats (>80 presentations). Understanding the exact effects of attention on the speech and music selectivity we have observed here will require further work.

Our work and that of others have identified neural responses specialized for processing music. Yet the definition of music is still debatable. In the field of ethnomusicology the definition of the ability to create music is more often used because of the ambiguity of defining music (Miller and Shahriari, 2012; Rice, 2014). In spite of this others have shown that there is general consensus in identifying "musical" sounds amidst a collection of commonly heard sound recordings, including the ones used in our study (see Norman-Haignere et al., 2015). Our paradigm provides an opportunity to study what defines music from the brain's perspective, what features make the music-selective regions of the brain respond, and how this specialization is affected by culture, experience, and subject expectations (Di Liberto et al., 2019). Additionally, EEG is cheaper and more portable than fMRI, and even though it lacks spatial resolution, we showed that we could still detect specializations for speech and music. With clever experimental design, this opens up possibilities to study the neural processing of music where fMRI is less feasible, such as with infants or populations in remote geographic locations.

References:

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM (2001) Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc Natl Acad Sci 98:13367–13372.

Carney LH, Li T, McDonough JM (2015) Speech coding in the brain: Representation of vowel formants by midbrain neurons tuned to sound fluctuations. eNeuro 2(4) ENEURO.0004-15.2015. doi: 10.1523/ENEURO.0004-15.2015

Cichy RM, Pantazis D, Oliva A (2014) Resolving human object recognition in space and time.

Nat Neurosci 17:455–462.

Cossy N, Tzovara A, Simonin A, Rossetti AO, De Lucia M (2014) Robust discrimination between EEG responses to categories of environmental sounds in early coma. Front Psychol 5:155.

deCharms RC, Blake DT, Merzenich MM (1998) Optimizing sound features for cortical neurons. Science 280:1439–1443.

Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 134:9–21.

Di Liberto GM, Pelofi C, Bianco R, Patel P, Mehta AD, Herrero JL, Cheveigné A de, Shamma S, Mesgarani N (2019) Cortical encoding of melodic expectations in human temporal cortex. bioRxiv 714634. doi: 10.1101/714634

Doelling KB, Poeppel D (2015) Cortical entrainment to music and its modulation by expertise. Proc Natl Acad Sci 112:E6233–E6242.

Evans N, Levinson SC (2009) The myth of language universals: Language diversity and its importance for cognitive science. Behav Brain Sci 32:429–448.

Fritz JB, David S V, Radtke-Schuller S, Yin P, Shamma SA (2010) Adaptive, behaviorally gated, persistent encoding of task-relevant auditory information in ferret frontal cortex. Nat Neurosci 13:1011–1019.

Giordano BL, McAdams S, Zatorre RJ, Kriegeskorte N, Belin P (2013) Abstract Encoding of Auditory Objects in Cortical Activity Patterns. Cereb Cortex 23:2025–2037.

Giordano BL, Pernet C, Charest I, Belizaire G, Zatorre RJ, Belin P (2014) Automatic domain-general processing of sound source identity in the left posterior middle frontal gyrus. Cortex 58:170–185.

Gold JI, Shadlen MN (2007) The neural basis of decision making. Annu Rev Neurosci 30:535–574.

Hausfeld L, Riecke L, Formisano E (2018) Acoustic and higher-level representations of

naturalistic auditory scenes in human auditory and frontal cortex. Neuroimage 173:472–483.

Hjortkjær J, Kassuba T, Madsen KH, Skov M, Siebner HR (2018) Task-modulated cortical representations of natural sound source categories. Cereb Cortex 28:295–306.

Huang N, Elhilali M (2017) Auditory salience using natural soundscapes. J Acoust Soc Am 141:2163–2176.

Hyvärinen A, Oja E (2000) Independent component analysis: algorithms and applications. Neural Networks 13:411–430.

Joris PX, Schreiner CE, Rees A (2004) Neural processing of amplitude-modulated sounds. Physiol Rev 84:541–577.

Kell AJE, Yamins DLK, Shook EN, Norman-Haignere S V., McDermott JH (2018) A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 98:630-644.

Khalighinejad B, Cruzatto da Silva G, Mesgarani N (2017) Dynamic encoding of acoustic features in neural responses to continuous speech. J Neurosci 37:2176–2185.

King AJ, Nelken I (2009) Unraveling the principles of auditory cortical processing: can we learn from the visual system? Nat Neurosci 12:698–701.

Lalor EC, Power AJ, Reilly RB, Foxe JJ (2009) Resolving precise temporal processing properties of the auditory system using continuous stimuli. J Neurophysiol 102:349–359.

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

Lehmann D, Skrandies W (1980) Reference-free identification of components of checkerboard-evoked multichannel potential fields. Electroencephalogr Clin Neurophysiol 48:609–621.

Lewis JW, Brefczynski JA, Phinney RE, Janik JJ, DeYoe EA (2005) Distinct cortical pathways for processing tool versus animal sounds. J Neurosci 25:5148–5158.

Liang L, Lu T, Wang X (2002) Neural representations of sinusoidal amplitude and frequency modulations in the primary auditory cortex of awake primates. J Neurophysiol 87:2237–2261.

Luo H, Poeppel D (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54:1001–1010.

Marti S, King J-R, Dehaene S (2015) Time-resolved decoding of two processing chains during dual-task interference. Neuron 88:1297–1307.

McDermott JH, Schemitsch M, Simoncelli EP (2013) Summary statistics in auditory perception. Nat Neurosci 16:493–498.

McDermott JH, Simoncelli EP (2011) Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. Neuron 71:926–940.

Mehr S, Singh M, Knox D, Ketter D, Pickens-Jones D, Atwood S, Lucas C, Egner A, Jacoby N, Hopkins EJ, Howard RM, O'Donnell TJ, Pinker S, Krasnow M, Glowacki L (forthcoming) Universality and diversity in human song. Science.

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202.

Miller TE, Shahriari A (2012) World Music: A Global Journey, Third Edition. New York, NY: Routledge.

Murray MM, Camen C, Gonzalez Andino SL, Bovet P, Clarke S (2006) Rapid brain discrimination of sounds of objects. J Neurosci 26:1293–1302.

Norman-Haignere S, Feather J, Brunner P, Ritaccio A, McDermott JH, Schalk G, Kanwisher N (2019) Intracranial recordings from human auditory cortex reveal a neural population selective for musical song. bioRxiv 696161. doi: 10.1101/696161

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

Nourski K V., Steinschneider M, Rhone AE, Kovach CK, Kawasaki H, Howard MA (2019)

Differential responses to spectrally degraded speech within human auditory cortex: An intracranial electrophysiology study. Hear Res 371:53–65.

Ogg M, Moraczewski D, Kuchinsky SE, Slevc LR (2019) Separable neural representations of sound sources: Speaker identity and musical timbre. Neuroimage 191:116–126.

Okada K, Rong F, Venezia J, Matchin W, Hsieh I-H, Saberi K, Serences JT, Hickok G (2010) Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech. Cereb Cortex 20:2486–2495.

Peelle JE, Gross J, Davis MH (2013) Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb Cortex 23:1378–1387.

Perrin F, Pernier J, Bertrand O, Echallier JF (1989) Spherical splines for scalp potential and current density mapping. Electroencephalogr Clin Neurophysiol 72:184–187.

Pinker S, Bloom P (1990) Natural language and natural selection. Behav Brain Sci 13:707–727.

Rice T (2014) Ethnomusicology: A Very Short Introduction. New York, NY: Oxford University Press.

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

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

Theunissen FE, Elie JE (2014) Neural processing of natural sounds. Nat Rev Neurosci 15:355–366.

Youn H, Sutton L, Smith E, Moore C, Wilkins JF, Maddieson I, Croft W, Bhattacharya T (2016) On the universal structure of human lexical semantics. Proc Natl Acad Sci 13:1766-1771.

Zatorre RJ, Bouffard M, Belin P (2004) Sensitivity to auditory object features in human temporal neocortex. J Neurosci 24:3637–3642.

Zhao S, Wai Yum N, Benjamin L, Benhamou E, Yoneya M, Furukawa S, Dick F, Slaney M,

Chait M (2019) Rapid ocular responses are modulated by bottom-up driven auditory salience. J Neurosci 0776–19. doi: 10.1523/JNEUROSCI.0776-19.2019

Zoefel B, Archer-Boyd A, Davis MH (2018) Phase entrainment of brain oscillations causally modulates neural responses to intelligible speech. Curr Biol 28:401-408.

Zuk NJ, Carney LH, Lalor EC (2018) Preferred tempo and low-audio-frequency bias emerge from simulated sub-cortical processing of sounds with a musical beat. Front Neurosci 12:349. doi: 10.3389/fnins.2018.00349