High-order areas and auditory cortex both represent the high-level event structure of music

Jamal A. Williams[1], Elizabeth H. Margulis[1,2], Samuel A. Nastase[1], Janice Chen[4], Uri Hasson[1], Kenneth A. Norman[1], Christopher Baldassano[3]

[1]Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ 08544, USA

[2]Department of Music, Princeton University, Princeton, NJ 08544, USA

[3]Department of Psychology, Columbia University, New York, NY 10027, USA

[4]Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD 21218, USA

Abstract

Recent fMRI studies of event segmentation have found that default mode regions represent high-level event structure during movie watching. In these regions, neural patterns are relatively stable during events and shift at event boundaries. Music, like narratives, contains hierarchical event structure (e.g., sections are composed of phrases). Here, we tested the hypothesis that brain activity patterns in default mode regions reflect the high-level event structure of music. We used fMRI to record brain activity from 25 participants (male and female) as they listened to a continuous playlist of 16 musical excerpts, and additionally collected annotations for these excerpts by asking a separate group of participants to mark when meaningful changes occurred in

each one. We then identified temporal boundaries between stable patterns of brain activity using a hidden Markov model and compared the location of the model boundaries to the location of the human annotations. We identified multiple brain regions with significant matches to the observer-identified boundaries, including auditory cortex, mPFC, parietal cortex, and angular gyrus. From these results, we conclude that both higher-order and sensory areas contain information relating to the high-level event structure of music. Moreover, the higher-order areas in this study overlap with areas found in previous studies of event perception in movies and audio narratives, including regions in the default mode network.

Significance Statement

Listening to music requires the brain to track dynamics at multiple hierarchical timescales. In our study, we had fMRI participants listen to real-world music (classical and jazz pieces) and then used an unsupervised learning algorithm (a hidden Markov model) to model the high-level event structure of music within participants' brain data. This approach revealed that default mode brain regions involved in representing the high-level event structure of narratives are also involved in representing the high-level event structure of music. These findings provide converging support for the hypothesis that these regions play a domain-general role in processing stimuli with long-timescale dependencies.

Introduction

Recent work has demonstrated that the brain processes information using a hierarchy of temporal receptive windows, such that sensory regions represent relatively short events (e.g., milliseconds to seconds) and higher-order regions represent longer events (e.g., minutes) while inheriting some of the lower-level structure from sensory regions (Hasson et al., 2015; Chen et al., 2016; Baldassano et al., 2017). For example, Baldassano and colleagues (2017) used a hidden Markov model (HMM) to find transitions between stable patterns of neural activity in BOLD data acquired from participants that watched an episode of the TV series Sherlock. The HMM temporally divides data into “events” with stable patterns of activity, punctuated by “event boundaries” where activity patterns rapidly shift to a new stable pattern. They found that, in sensory regions such as early visual cortex, the data were best-fit by a model with short-lasting chunks, presumably corresponding to low-level perceptual changes in the episode; by contrast, when they applied the model to data from a higher-order area such as posterior medial cortex, the best-fitting model segmented the data into longer-lasting chunks corresponding to more semantically meaningful scene changes. Critically, human annotations of important scene changes most closely resembled the model-identified boundary structure found in frontal and posterior medial cortex, which are key hubs in the brain’s default mode network (DMN) (Shulman et al., 1997; Raichle et al., 2001). Studies have also found that the same event-specific neural patterns are activated in default-mode regions by audiovisual movies and by verbal narratives describing these events (Zadbood et al., 2017; Baldassano et al., 2017, 2018),

69 providing further evidence that these regions represent the underlying meanings of the
70 events and not low-level sensory features.
71
72 Jackendoff and Lerdahl (2006) suggest that music and language are structured into
73 meaningful events that help people comprehend moments of tension and relaxation
74 between distant events. If music resembles language in this way, then the
75 representation of hierarchical event structure in music (e.g., at the level of phrases,
76 sections, and entire songs) and in verbal and audiovisual narratives may be supported
77 by similar neural substrates. Indeed, some evidence already exists for shared neural
78 resources for processing music and language (Tallal and Gaab, 2006; Patel, 2011;
79 Koelsch, 2011; Peretz et al., 2015; Lee et al., 2019). In the current work, we test the
80 hypothesis that posterior medial cortex (PMC) and other default mode network (DMN)
81 regions that represent high-level event structure in narratives also play a role in
82 representing high-level event structure in music.
83
84 In our paradigm, we presented fMRI participants with examples of complex real-world
85 music belonging to genres familiar to our participant population: jazz and classical. A
86 separate group of behavioral participants were asked to annotate meaningful events
87 within each of the excerpts. We applied hidden Markov models using both a whole-brain
88 searchlight and cortical parcellation to measure event structure represented in cortical
89 response patterns throughout the brain. We then used the events provided by the
90 annotators to guide the event segmentation of the brain data. The goal of this analysis
91 was to identify brain regions that chunk the stimuli in a way that matched the human

annotations. By fitting the model at each ROI and then comparing the observed boundary structure to that of the annotators, we show that – in a group of passive listeners – regions in the default mode network and also sensory areas are involved in representing the high-level event structure in music (i.e., these regions show neural pattern shifts that line up with human annotations of event boundaries). We also show that these event representations become more coarse as they propagate up the cortical processing hierarchy.

Materials and Methods

Participants

We collected fMRI data from a total of 25 participants (12 female, ages 21–33). We also recruited 7 human annotators for a separate behavioral task (described below). 13 of the fMRI participants were native English speakers. The experimental protocol was approved by the Institutional Review Board of Princeton University, and all participants gave their written informed consent.

Stimuli

16 musical excerpts were selected based on the criteria that changes between sub-sections would likely be recognized by people without formal music training (e.g., change from piano solo to drum solo). Excerpts also had to be instrumental (i.e. lack vocals). Excerpts were drawn from two different genres (8 classical and 8 jazz). Excerpts were then randomly selected to be truncated (with the introduction kept intact)

to one of four different durations: 90 s, 135 s, 180 s, and 225 s, such that there were four excerpts of each length. Furthermore, two excerpts of each duration were sampled from each genre. For example, only two classical excerpts had a duration of 90 seconds and only two jazz excerpts had a duration of 90 seconds. The total duration of the playlist was approximately 45 minutes and there were no breaks between excerpts.

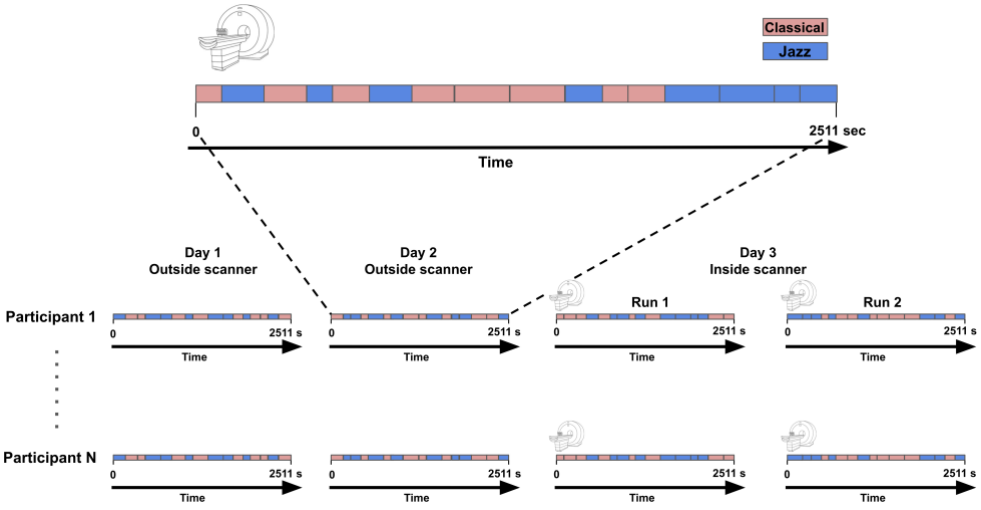


Figure 1. Top: Example 45 minute scanning run with classical excerpts depicted in pink and jazz excerpts in blue. Each block in the timeline represents an excerpt and block lengths reflect excerpt durations. Bottom: Overview of experiment. Participants heard the playlist four times (once on each of the two days prior to scanning and twice on the third day while being scanned). The excerpts were presented in a different order each of the four times that a given participant heard the playlist, but – within a given phase of the experiment (e.g., Run 1 on Day 3) – the order of excerpts was kept the same across participants.

Experimental design and statistical analysis

The experiment took place over three consecutive days (Figure 1): On the first two days participants heard a playlist of sixteen musical excerpts (once for each day) and on the third day they heard the same playlist for two separate runs while we recorded changes in their BOLD activity using fMRI. Altogether, each participant heard the

playlist four times. Each time that a given participant heard the playlist, the excerpts were presented in a different order. However, within a given phase of the experiment (e.g., the first scanner run on day 3), the order of excerpts was kept the same across participants; this way, all participants listened to the same stimuli in the same order in the scanner, which was necessary for the Shared Response Model analyses described below. To promote stable representations of the music, participants listened to the playlist on each of the two days prior to scanning. During these listening sessions, we collected ratings from participants about their enjoyment, engagement, and familiarity with each piece (only familiarity ratings are discussed in this manuscript); these ratings were collected immediately after hearing each piece. Answers for each rating category were given on a 5-point Likert scale where a 5 corresponded to very familiar. We found an increase in average familiarity from day 1 to day 2 ($t(22)= 9.04$, $p < 0.0001$) indicating that participants remembered the music played in the first pre-scan session. Two participants were excluded from this analysis because their day 2 ratings were lost.

After each of these listening sessions, participants took a short recognition test where they heard 32 randomly drawn 3-second clips of a piece that were either from the actual listening session or a lure (i.e. different piece by the same artist) and made a response using a 5-point Likert scale indicating whether they recognized the excerpt as having been presented previously. In addition to the familiarity ratings across the two pre-scan days, this measure helped us determine if participants had learned the music after each behavioral listening session. Participants showed above-chance discrimination (i.e.,

higher recognition scores for presented excerpts vs. lures) on both days (Day 1: $t(24) = 12.2$, $p < 0.0001$; Day 2: $t(24) = 15.1$, $p < 0.0001$; Figure 2).

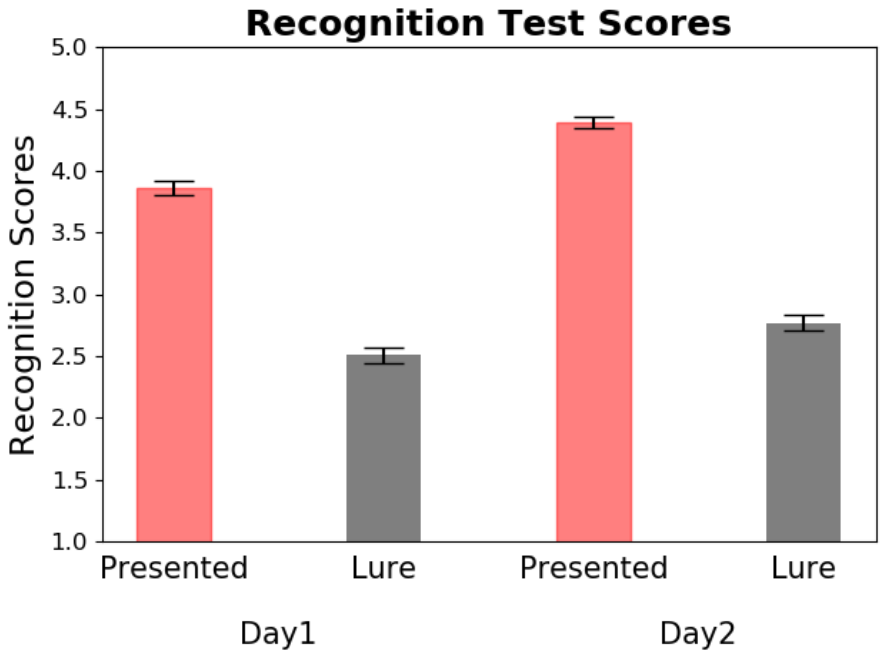


Figure 2. Recognition test scores for both prescan days. Plot shows that presented excerpts were given higher recognition scores than lures. The y-axis represents a 5-point Likert scale where one means not studied and five means studied. Error bars represent standard error of the mean.

On the third day, participants returned for the scanning session in which they listened to the playlist twice (with excerpts played in a different order for the two scanning runs; as noted above, the order of excerpts within a run was the same across participants). During each run, participants were asked to perform a white noise detection task. Specifically, during each excerpt, a brief (1 sec) white noise pulse was played at a randomly chosen time point within the middle 60% of each excerpt. The onset of each noise pulse was also randomized across participants. Participants were told to make a button response to indicate that they heard the noise. This manipulation served to keep

participants attentive throughout each excerpt. Following both scanning runs, participants took a final recognition test and then completed a brief demographic survey.

Event annotations by human observers

In a separate behavioral experiment, we asked seven different raters (only one rater reported having extensive musical training) to listen to our stimuli one at a time with the task of pressing a button when a “meaningful” transition occurred within each piece (similar to the method used by Sridharan et al., 2007). The number of event boundaries identified by the observers varied across excerpts ranging from 3 to 17 boundaries (with a mean of 7.06 and standard deviation of 0.91 across excerpts). It is worth noting that excerpt durations also varied, with a range of 90 sec to 225 sec (durations were either 90, 135, 190, or 225 sec) and an average duration of 157.5 sec and standard deviation of 50.3 sec across excerpts. A timepoint was considered to be an event boundary when at least five annotators marked a boundary within three seconds before or after the given timepoint (method used from Baldassano et al., 2017). The mean number of consensus boundaries across excerpts acquired using this method roughly matched the mean number of boundaries assigned by individual participants across all of the excerpts (with a mean of 7.98 and standard deviation of 2.98 across excerpts).

Scanning parameters and preprocessing

Imaging data were acquired on a 3T full-body scanner (Siemens Prisma) with a 64-channel head coil. Data were collected using a multi-band accelerated T2-weighted

192 echo-planar imaging (EPI) sequence (release R015) provided by a C2P agreement with
193 University of Minnesota (Moeller et al., 2010; Setsompop et al., 2012; Xu et al., 2013;
194 Auerbach et al., 2013; Sotiropoulos et al., 2013; Cauley et al., 2014): 72 interleaved
195 transverse slices; in-plane resolution = 2.0 mm; slice thickness = 2.0 mm with no inter-
196 slice gap; field of view (FOV) = 208 mm; base resolution = 104; repetition time (TR) =
197 1000 ms; echo time (TE) = 37 ms; flip angle (FA) = 60 deg; phase-encoding (PE)
198 direction = anterior to posterior; multi-band acceleration factor = 8. Three spin-echo
199 volume pairs were acquired matching the BOLD EPI slice prescription and resolution in
200 opposing PE directions (anterior to posterior and posterior to anterior) for susceptibility
201 distortion correction: TR/TE = 8000/66.60 ms; FA/refocus FA = 90/180 deg; TA = 32s
202 (Andersson et al., 2003).
203
204 Additionally, a whole-brain T1-weighted volume was collected: 3D magnetization-
205 prepared rapid gradient echo (MPRAGE) sequence; 176 sagittal slices; 1.0 mm^3
206 resolution; FOV = 256 mm; base resolution = 256; TR/TE = 2300/2.88 ms; inversion
207 time (TI) = 900 ms; FA = 9 deg.; PE dir = anterior to posterior; IPAT mode = GRAPPA
208 2X; TA= 5 min 20 sec.
209
210 The EPI volumes were realigned using a 6 parameter rigid-body registration (MCFLIRT;
211 Jenkinson et al., 2002). Given the short effective TR, slice time correction was not
212 performed. Susceptibility-induced distortions were modeled in the opposing spin-echo
213 volume pairs using the FSL ‘topup' tool, and the resulting off-resonance field output was
214 provided as input to distortion correct the time series of fMRI data using the FSL

215 'applywarp' tool (Andersson et al., 2003). The susceptibility distortion correction and
216 realignment were applied in a single interpolation step to minimize blurring. Remaining
217 pre-processing and co-registration steps were performed using FEAT (Woolrich et al.,
218 2001, 2004). This included linear detrending, high-pass filtering (330 s cutoff), and
219 spatial normalization to the MNI152 template released with FSL.
220

221 Whole-brain parcellation and searchlights

222 We conducted our primary analysis across the whole brain by defining ROIs in two
223 parallel ways. In the parcellation approach, we used 300 non-overlapping parcels tiling
224 the whole cortex in MNI space (Schaefer et al., 2018). In the whole-brain searchlight
225 approach, we divided the MNI volume into overlapping spherical searchlights with a
226 radius of 10 voxels and a stride of 5 voxels. This resulted in 2,483 searchlights that
227 spanned the whole cortex in MNI space. Only searchlights containing at least 30 voxels
228 were included in the analysis (matching the number of shared features specified in the
229 shared response model described in the following section) and the mean number of
230 voxels per searchlight was 381.76 voxels with a standard deviation of 168.09 voxels.
231 We assigned the output value for a given searchlight to all voxels within a 5-voxel radius
232 to account for the stride and then averaged the values for voxels where overlap
233 occurred. All analyses below were run separately within each ROI (parcel or
234 searchlight).
235
236
237

Shared response model

Within each ROI, we first reduced the dimensionality of our BOLD data to a set of shared features (30 features) using the shared response model (SRM; Chen et al., 2015), which aligns functional response patterns across participants and provides a low-dimensional representation of the shared variance in the BOLD data (Figure 3B). Since the only information shared across participants is the musical stimulus, shared variance is necessarily *music-related* variance. The SRM procedure thus has the effect of highlighting (shared) music-related variance (see Figure 3 for diagram of analysis pipeline). The “optimal” number of shared features will tend to vary across both stimuli and brain areas (e.g., as determined by cross-validation); here, for the sake of simplicity, we fixed the number of shared features at 30 as a reasonable middle ground based on values used in prior work (Chen et al., 2015; Chen et al., 2017; Nastase et al., 2020). We applied an SRM with parameters estimated from run 1 to transform the run 2 data, and applied an SRM with parameters estimated from run 2 to transform the run 1 data. Next, for each participant’s transformed data, song-specific onset and offset times were used to extract time points corresponding to a particular excerpt; for each of the 32 excerpts (16 songs in each of the two runs), we then averaged (across participants) all of the participant-specific timeseries for that excerpt, resulting in 32 average timeseries. These average timeseries were then analyzed using the hidden Markov model as described in the next section.

Event segmentation analysis

After fitting the SRM to a given ROI, we fit a hidden Markov model (HMM; Baldassano et al., 2017) to the response time series in shared space for each excerpt, setting the number of states in the HMM equal to the number of segments specified by our human annotators for each excerpt. We used a specialized HMM variant developed by Baldassano et al. (2017) that is optimized for event segmentation (i.e., identifying jumps in neural patterns). This HMM variant seeks to model the shared feature timecourse as a set of successive transitions between stable states (i.e., patterns of fMRI activity in shared space), where – in our variant of the HMM – the model is not permitted to return to a state once it leaves that state. Once fit to the data, the model tells us what the stable best-fitting patterns are and when the transitions occurred between these patterns; this HMM variant was implemented using the EventSegment function in BrainIAK (Kumar et al., 2020a, 2020b).

For our primary analysis, we were interested in finding brain regions (parcels or searchlights) whose transition structure most closely resembled the event boundary structure given by our annotators (Figure 3C). After acquiring boundary estimates from the HMM, we evaluated how closely in time the boundaries found by the model matched the boundaries supplied by our annotators. To quantify the degree of match, we counted the number of human-annotated boundaries for which there was a HMM boundary within three TRs of that human-annotated boundary. Note that all human boundaries were shifted forward 5 TRs (5 seconds) to account for the hemodynamic lag. We created a null model by randomly selecting timepoints as boundaries (keeping

the number of events the same, as in Baldassano et al., 2017) and computed the number of matches for these null boundaries, repeating this process 1000 times to produce a null distribution. We computed a z-value of the real result versus the null distribution by subtracting the average of the permuted match scores from the true match score and dividing this difference by the standard deviation of the permuted scores. This procedure was repeated at every ROI. By acquiring z-values at each ROI for all 32 excerpts (16 distinct excerpts x 2 runs), we obtained 32 separate spatial maps of z-values. Next, we averaged the two z-maps corresponding to each distinct excerpt (one from each run), resulting in 16 total z-maps. To summarize across the z-values for the 16 distinct excerpts, we ran a one sample t-test against zero to see which voxels had the most reliable matches across all excerpts. The resulting t-values were converted to p-values and then adjusted for multiple tests to control the false discovery rate (FDR) at a value q (Benjamini & Hochberg, 1995).

To visualize the results, each spatial map of t-values (one corresponding the voxel-wise searchlight procedure and the other corresponding to the parcellation procedure) was displayed on the cortical surface (masked to include only vertices that exhibited a significant effect). Since each analysis was performed in volumetric space, volume data were projected to the cortical surface using the automatic volume to surface rendering algorithm within Surf Ice (https://www.nitrc.org/projects/surfice/).

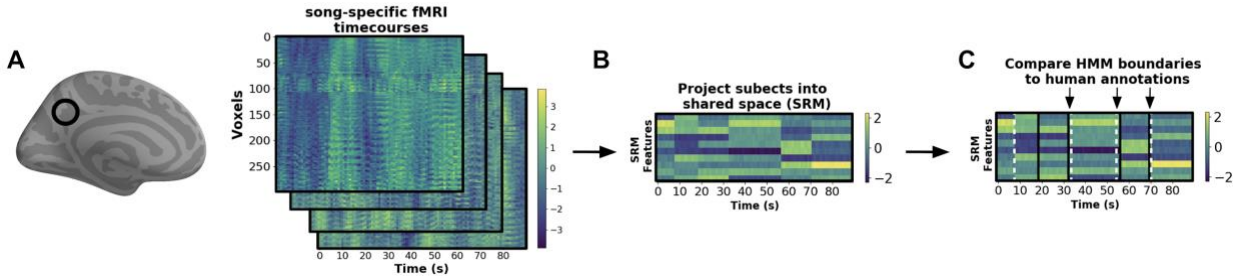


Figure 3. Diagram of analysis pipeline. (From left to right) **Panel A.** For each participant (N=25), voxels from an ROI were selected using pre-defined parcellations or the searchlight method; we then extracted song-specific timecourses (Voxels x TRs) from the selected voxels (black circle). Inflated brain image was created using PySurfer (https://github.com/nipy/PySurfer/). **Panel B.** Selected data were reduced to a set of shared features using the shared response model (SRM), which highlights shared variance (*music-related* variance) across participants. The SRM was applied to the full time series concatenated across all 16 excerpts in a single run, then the full time series was split back into separate excerpts. **Panel C.** HMM boundaries (white dotted lines) and human annotations (black lines) were considered to match (black downward arrows) when HMM boundaries fell within 3 TRs of a human annotation. Then, true match scores were compared to a null distribution constructed by comparing shuffled HMM boundaries to human annotations, resulting in a z-value for each ROI.

Controlling for low-level music features

To further determine whether brain areas of the DMN represent high-level musical event structure, as opposed to low-level sensory information, we repeated the searchlight analysis, this time regressing out musical features extracted from each auditory stimulus prior to fitting the HMM. These features consisted of mel-frequency cepstral coefficients (MFCCs; i.e. timbre information), chromagrams (harmonic information), and tempograms (rhythmic information). For MFCCs, the top 12 channels were extracted since these lower order coefficients contain most of the information about the overall spectral shape of the source-filter transfer function (Poorjam, 2018). Chromagrams

consisted of 12 features, each corresponding to a distinct key in the chromatic scale. Tempograms initially consisted of 383 features, each representing the prevalence of certain tempi (in beats per minute; BPM) at each moment in time. Since most of the tempo-related variance was explained by a much smaller set of features, we reduced the 383 features to 12 features using PCA in order to match the number of features used for MFCCs and chromagrams. Feature extraction on music stimuli was performed using Librosa (McFee et al., 2015), a Python package developed for audio and music analysis.

Identifying preferred event timescales

After identifying brain regions with neural event boundaries that match human annotations (using the procedures described in *Event segmentation analysis*, above), we ran a follow-up analysis to further probe the properties of three such regions (right auditory cortex, left angular gyrus, and left precuneus). Specifically, the goal of this follow-up analysis was to assess the preferred timescale of these regions. In contrast to our primary event segmentation analysis (which used a fixed number of events for each excerpt, matching the number of human-annotated events for that excerpt), here we tried models with different numbers of events and assessed how well model fit varied as a function of the number of events. The measure of model fit we used was the average pattern similarity between pairs of timepoint-specific multivoxel patterns falling *within* the same event, minus the average pattern similarity between patterns falling *across* events (Baldassano et al., 2017). We call this measure the *WvA score* (short for “Within versus Across”); higher WvA scores indicate a better fit of the event boundaries to the data.

The ROIs for this analysis (right auditory cortex, left angular gyrus, and left precuneus) were defined by selecting voxels within the parcellation corresponding to right auditory cortex, left angular gyrus, and left precuneus, and then (for extra precision) intersecting these parcels with voxels that were also significant in our primary searchlight analysis looking for neural boundaries that matched human-annotated boundaries ($q < 0.01$). For each ROI, we fit HMMs to each song with differing numbers of events ranging from 3 to 45. For each HMM fit, we measured the maximum event duration, and then identified all pairs of timepoints whose temporal distance was less than this duration. The constraint of using timepoints whose distance was less than the maximum event duration was used so that the number of within- and across-event pairs would be roughly equal (regardless of the number of events). The WvA score was computed as the average spatial pattern correlation for pairs of timepoints falling in the same (HMM-derived) event minus the average correlation for pairs of timepoints falling in different events. We then averaged the results across excerpts. Note that, since the excerpts are different lengths, a given number of events might correspond to different average event lengths for different excerpts (e.g., a 3-event model applied to a 180 second excerpt has an average event length of 60 seconds, but a 3-event model applied to a 90 second excerpt would have an average event length of 30 seconds). Because our goal was to find each area's preferred event length, we converted our WvA results for each excerpt to be a function of the average event length (in seconds) rather than the number of events, and averaged these results across excerpts. Finally, we computed a preferred event length for each ROI, operationalized as the weighted average of all possible event lengths (2 seconds to 75 seconds), with the averaged WvA scores as the weights; put

another way, preferred event length was computed as the center-of-mass of the plot of WvA as a function of event length.

To test whether the preferred event length in auditory cortex was shorter than that of angular gyrus and precuneus, we performed a bootstrap analysis, repeating the above analysis (including re-running SRM) 1000 times for different bootstrap resamples of the original dataset. At each iteration of the bootstrap, we applied the analysis to a sample of participants drawn randomly with replacement from the original data. We computed p-values by finding the proportion of bootstraps where the preferred length for auditory cortex was greater than the preferred length for angular gyrus or precuneus.

Results

Neural boundary match to behavioral annotations

We wanted to test the hypothesis that behaviorally-defined event boundaries could be identified in higher-order cortical regions, especially those overlapping with the DMN. For this analysis, we compared the HMM boundaries to the specific timepoints labeled as boundaries by the annotators. According to both parcellation and searchlight results, we found that the highest number of matches between model boundaries and human annotations were in auditory cortex, left angular gyrus, left precuneus, bilateral medial prefrontal cortex, and motor cortex (Figure 4).

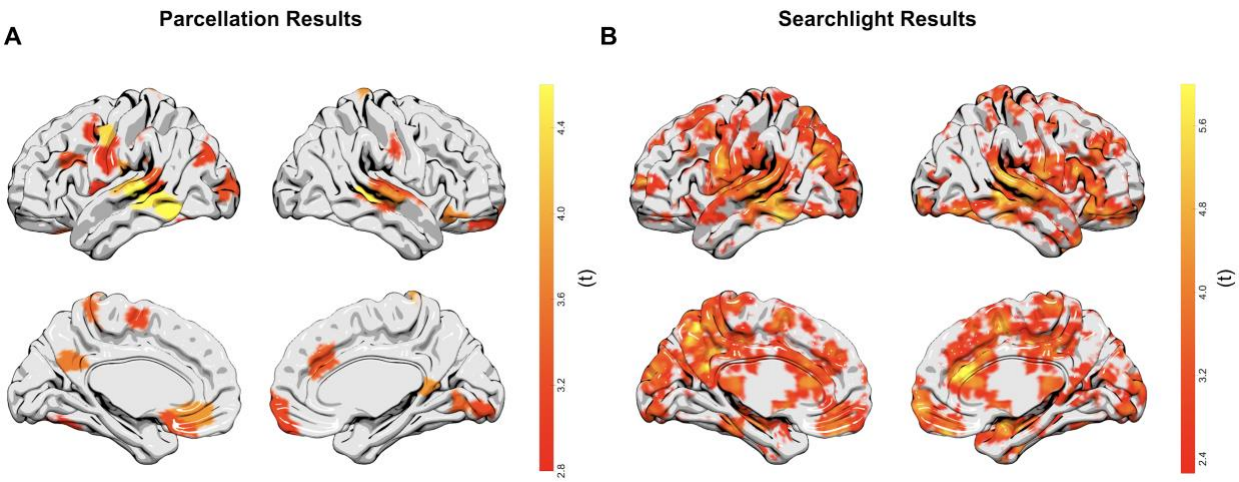


Figure 4. Panel A. Parcellation Results. For each of 300 parcels, we tested whether the average match between HMM-defined neural boundaries and human annotations across all songs was significantly greater than zero. Significant parcels included auditory cortex as well as several regions in the default mode network (DMN): angular gyrus (AG), posterior medial cortex (PMC), and medial prefrontal cortex (mPFC). Parcels in left motor, premotor, and supplementary motor cortex were also significant. Results are thresholded via FDR, $q < 0.05$. **Panel B. Searchlight Results.** For 2,483 searchlights across the whole cortex, we again tested whether the average match between neural and annotated boundaries across all songs was significantly greater than zero. Significant voxels overlapped with those from the parcellation results, and additionally showed effects in right inferior frontal gyrus and hippocampal regions. Results are thresholded via FDR, $q < 0.01$.

Influence of low-level music features

To determine whether the neural event boundaries were driven by low-level acoustic features, we also performed a version of the searchlight analysis in which we controlled for timbral, harmonic, and rhythmic features. Overall, this reduced the number of searchlights passing the significance threshold (Figure 5) compared to the original searchlight analysis. However, some higher-order areas overlapping with the DMN (precuneus and mPFC) did pass the significance threshold ($q < 0.01$), which suggests

that event boundaries found in these regions do not directly correspond to low-level music features and may instead be related to more abstract representations of the event structure. Notably, some significant searchlights in auditory cortex were also observed, indicating that – even in sensory areas – the event boundaries were being driven (at least in part) by high-level music features.

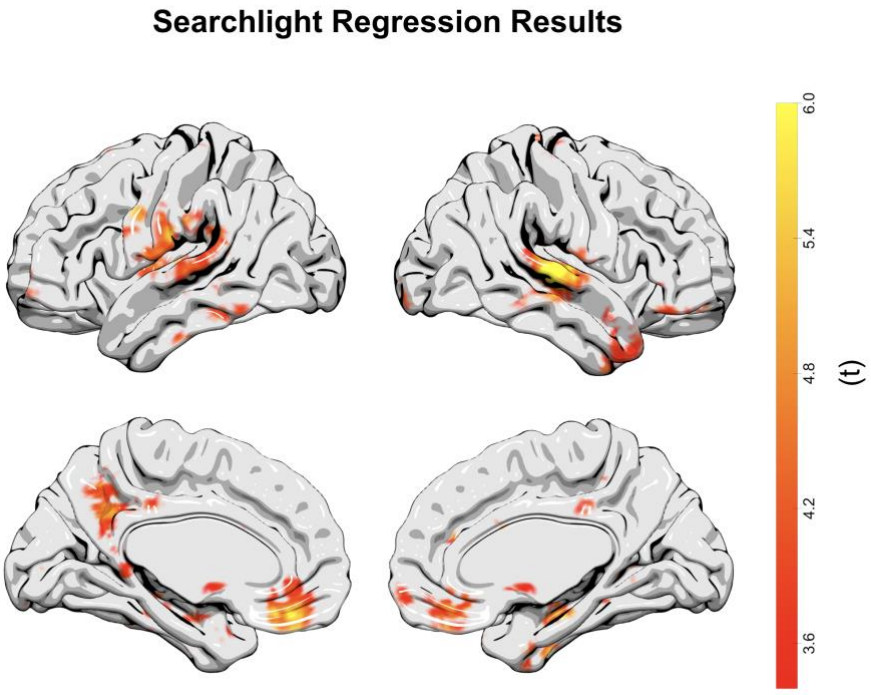


Figure 5. Searchlight results accounting for low-level music features. We recomputed the match between HMM-derived neural boundaries and human annotations after regressing out low-level acoustic features from each participant's BOLD data prior to fitting the HMM. Significant effects were still observed in higher-order areas overlapping with the DMN as well as several auditory areas, suggesting that boundaries detected in these areas do not necessarily depend on low-level musical features. Results are thresholded via FDR ($q < 0.01$).

Preferred event lengths across ROIs

How do we reconcile the role of auditory cortex in high-level event representation (as shown in the above analyses) with its well-known role in representing low-level auditory features? Importantly, these claims are not mutually exclusive. Our analyses, which set the number of event states in the model to equal the number of human-annotated boundaries, show that auditory cortex has some (statistically-reliable) sensitivity to high-level events, but this does not mean that this is the *only* event information coded in auditory cortex or that it is the *preferred* level of event representation. Figure 6 shows timepoint-by-timepoint similarity matrices for three ROIs, in which spatial pattern correlations were computed for each pair of timepoints in an excerpt (Kamasi Washington's *Change of the Guard*). Blocks of high correlation around the diagonal indicate periods of time in which patterns stayed relatively constant. This block structure in default-mode regions (angular gyrus and precuneus) is at a similar scale as the human annotations (black squares), while the smaller blocks in auditory cortex suggest that it represents finer event structure than is present in human annotations.

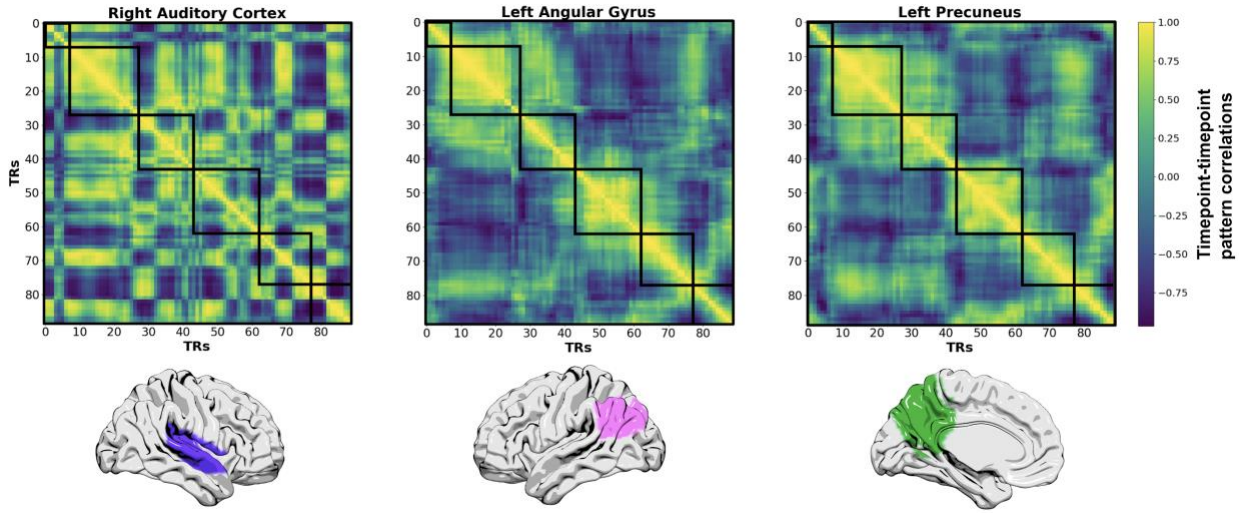
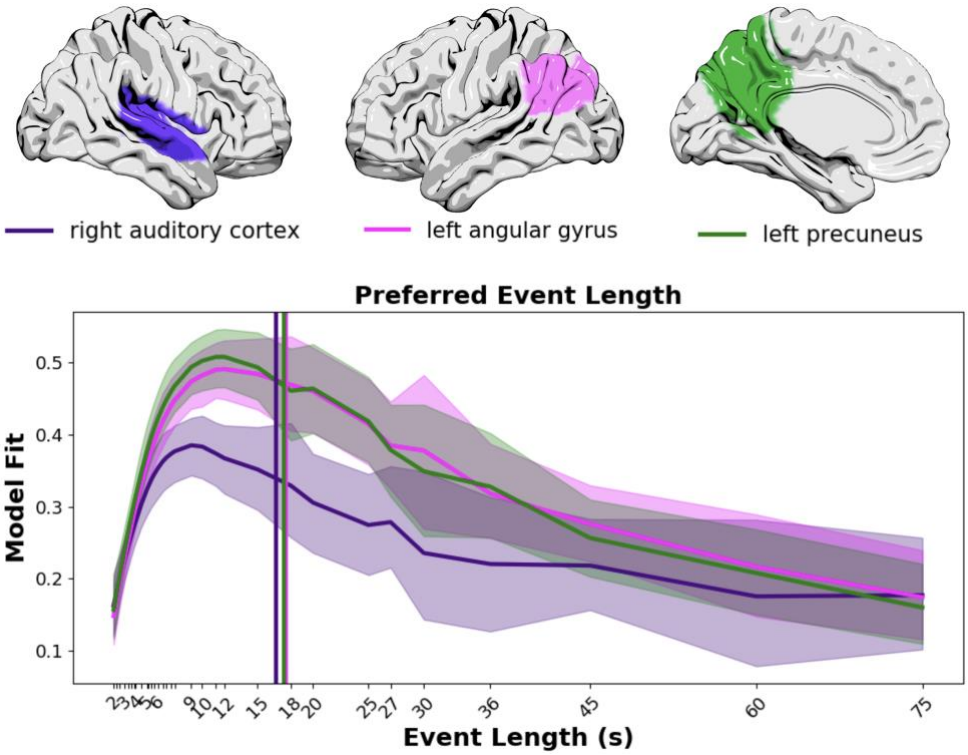


Figure 6. Timepoint-by-timepoint spatial pattern similarity matrices for Kamasi Washington's *Change of the Guard* in right auditory cortex (left), left angular gyrus (middle), and left precuneus (right). Black squares overlaid on the similarity matrices represent events defined behaviorally by our human annotators.

We defined the preferred timescale of each region by running HMMs with different numbers of event states, and finding the average event length (in seconds) that produced the best model fits across songs (Figure 7). Using a bootstrap analysis, we found that auditory cortex's preferred event length (average center of mass across bootstraps was 16.67 seconds) was significantly shorter than the preferred event lengths of both angular gyrus ($p=0.012$; average center of mass was 17.57 seconds) and precuneus ($p=0.036$; average center of mass was 17.34 seconds).



462

463 **Figure 7.** To test the hypothesis that higher-order areas (precuneus and AG) exhibit event-structured
464 activity for longer periods than auditory cortex, we computed a measure of model fit (within-event vs.
465 across-event pattern correlations; WvA score) for HMMs with the number of events ranging from 3 to 45.
466 We found that auditory cortex's preferred length (purple vertical line) was shorter on average than the
467 preferred lengths of angular gyrus and precuneus (pink and green vertical lines, respectively).

468

469 Discussion

470

471 In this study, we sought to demonstrate that brain areas that have been implicated in
472 representing high-level event structure for narrative-based stimuli are also involved in
473 representing the high-level event structure of music in a group of passive listeners. We
474 provide evidence that regions in the default mode network (DMN) are involved in
475 representing the event structure of music as characterized by human annotators. The

durations of these human-annotated events lasted on the order of a few seconds up to over a minute.

Our results indicate that high-level structure is represented in both high-level DMN regions but also in auditory cortex. Auditory cortex, however, may not explicitly represent high-level events at the level of human annotators; that is, the behaviorally-identified event boundaries are likely a subset of the finer-grained event boundaries encoded in auditory cortex. When we force the HMM to match the number of human-annotated boundaries, the HMM finds them, demonstrating that coding in auditory cortex is modulated by high-level event structure. However, when we remove this constraint and allow the number of events to vary, auditory cortex prefers shorter events on average relative to higher-order brain areas (Figure 7). Additionally, we showed that—when we regress out low-level music features relating to timbre, harmony, and rhythm and re-run the analysis—higher-order areas (several overlapping with DMN), as well as auditory cortex, still significantly match with the annotations. These results provide additional evidence that event boundaries in DMN, as well as auditory areas, are not purely driven by low-level acoustic changes in the music, but are also tracking more abstract event structure in musical pieces. It is possible that boundaries marking the shift between large-scale segments within DMN and auditory regions could be driven by a complex shift in a combination of the acoustic properties or possibly emotional changes within the excerpts (Daly et al., 2015). Notably, our findings of high-level coding in auditory cortex converge with other recent work demonstrating that hierarchical neural representations of music are distributed across primary and non-

primary auditory cortex (Landemard et al., 2020) and that higher-order representations of music in these areas support human performance on a genre recognition task (Kell et al., 2018).

Our findings that left precuneus, bilateral mPFC, and angular gyrus were involved in representing event structure at the level of phrases and sections contrast with those in Farbood et al. (2015), who found that these regions responded reliably to stories but did not respond reliably to music. Furthermore, in their study, there was minimal overlap between voxels in posterior medial cortex that responded to stories and voxels that responded to music. In our study, we show that, at a regional level, these “verbal narrative” areas are indeed involved in representing the high-level event structure in music. One major way in which our studies differed was our use of an HMM to detect evidence of musical event structure in higher-order areas. The HMM is optimized to detect periods of relative stability punctuated by shifts in response patterns, which one would expect for an area encoding high-level event structure (i.e., there should be stability within events and changes across events). Temporal ISC (inter-subject correlation analysis; the analysis method used in the study by Farbood) is designed to pick up on *any* kind of reliable temporal structure and is not specifically designed to detect the "stability punctuated by shifts" structure that we associate with event cognition, making it less sensitive to this kind of structure when it is present. This highlights one of the advantages of using HMMs for detecting meaningful brain activity related to the temporal dynamics of naturalistic stimuli, such as music.

In our study, we showed the involvement of frontal brain areas (specifically vmPFC and rIFG) in representing high-level musical event structure. The recruitment of vmPFC during music processing has been found in a previous study (Blood and Zatorre, 2001). Specifically, Blood and Zatorre showed that activity in vmPFC was correlated with pleasure response ratings to music. This effect was also found in supplementary motor areas (SMA) as well as anterior cingulate cortex (ACC). Interestingly, these same areas appear to represent the high-level event structure for music in our study, suggesting that regions that represent long-timescale structure also play a role in affective responses to music. This makes sense given that pleasure in response to music is likely derived from the accumulation of information, leading to the buildup of expectations that are eventually resolved over time (Margulis, 2005; Lehne et al., 2013; Gingras et al., 2016).

We also showed that rIFG was significantly involved in representing the high-level event structure of real-world music, particularly in the searchlight analysis. Several studies have reported the role of rIFG in the processing of musical syntax (Koelsch et al., 2002; Tillman et al., 2003; Minati et al., 2008) but it was unclear from these studies whether hierarchical processing was occurring in this region. In a recent fMRI study, however, Cheung et al. (2018) found that rIFG increased in activity when violations to non-local dependencies in nested musical sequences occurred, suggesting that rIFG is involved in the hierarchical processing of event structure in music, consistent with the findings reported here.

Both searchlight and parcellation analyses revealed the involvement of motor areas in high-level event representation. Specifically, we see the recruitment of bilateral primary motor cortex (M1), bilateral premotor cortex (PMC), bilateral supplementary motor areas (SMA), and primary somatosensory cortex. These findings are not surprising given the large number of studies that find motor areas to be active during passive music listening (Gordon et al., 2018). Research suggests that the motor system is involved in a wide range of music-related processes such as sequencing, timing, and spatial organization (Zatorre et al., 2007). Furthermore, predictive coding accounts of passive musical listening such as the ASAP (Action Simulation for Auditory Prediction) hypothesis (Patel and Iversen, 2014) suggest that the motor cortex sends signals to auditory cortex that help the auditory system make predictions about upcoming beats. In our study, consensus boundaries were often found in locations that were on a beat within a developed beat pattern. In other words, meaningful changes tended to occur on an expected beat. Importantly, the fMRI participants in our study never performed a manual event segmentation task (the annotations were made by a separate group of participants) and were not explicitly told to notice these changes, but motor areas nonetheless appear to be involved in representing them.

Conclusion

In this study, we sought to determine whether certain regions in the default mode network, which have been shown to be involved in representing the high-level event structure in narratives, were also involved in representing the high-level event structure in real-world music. Recent fMRI work, not using music, has shown that hidden Markov

models (HMMs) can help us understand how the brain represents large-scale event structure (e.g., Baldassano et al., 2017; Baldassano et al., 2018; Antony et al., 2020). By using HMMs to segment fMRI response patterns over time according to the event structure provided by a separate group of human annotators, we found that areas of the DMN were indeed involved in representing the high-level event structure (e.g., phrases, sections) in music in a group of passive listeners. We also showed that our event segmentation model was capable of detecting high-level event structure in auditory cortex when the model was constrained by the number of human annotations. However, when we allowed the number of events to vary, we found that the auditory cortex data was best fit by a model with shorter events, compared to data from precuneus and angular gyrus. Results in a subset of regions, including higher-order regions and some auditory regions, remained significant even after controlling for low-level musical features, suggesting that sensitivity to high-level event structure in higher-order brain areas may reflect contributions from memory or the perception of more abstract musical structure.

References

Alexander, S., Ru, K., Gordon, E.M., Laumann, T.O., Zuo, X., Holmes, A.J., Eickhoff, S.B., & Yeo, B.T.T. (2018) Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, *28*(9), 3095–3114, https://doi.org/10.1093/cercor/bhx179

588 Andersson, J.L., Skare, S., & Ashburner, J. (2003) How to correct susceptibility
589 distortions in spin-echo echo-planar images: application to diffusion tensor imaging.
590 *NeuroImage*. *20*(2), 870-88.
591
592 Antony, J.W., Hartshorne, T.H., Pomeroy, K., Hasson, H., McDougle, S.D., & Norman,
593 K.A. (2020) Behavioral, physiological, and neural signatures of surprise during
594 naturalistic sports viewing. *Neuron*. *109*(2), 377-390.
595 https://doi.org/10.1016/j.neuron.2020.10.029
596
597 Auerbach, E.J., Xu, J., Yacoub, E., Moeller, S., & Uğurbil, K. (2013) Multiband
598 accelerated spin-echo echo planar imaging with reduced peak RF power using time-
599 shifted RF pulses. *Magnetic Resonance in Medicine*. *69*, 1261-1267.
600
601 Baldassano, C., Chen, J., Zadbood, A., Pillow, J.W., Hasson, U., & Norman, K.A. (2017)
602 Discovering event structure in continuous narrative perception and memory. *Neuron*,
603 *95*(3), 709-721.
604
605 Baldassano, C., Hasson, U., & Norman, K.A. (2018) Representation of real-world event
606 schemas during narrative perception. *Journal of Neuroscience*, *38*(45), 9689-9699. doi:
607 10.1523/JNEUROSCI.0251-18.2018
608

609 Blood, A.J. & Zatorre, R.J. (2001) Intensely pleasurable responses to music correlate
610 with activity in brain regions implicated in reward and emotion. *PNAS*, *98*(20), 11818-
611 11823. doi: 10.1073/pnas.191355898
612
613 Cauley, S.F., Polimeni, J.R., Bhat, H., Wald, L.L., & Setsompop, K. (2014) Interslice
614 leakage artifact reduction technique for simultaneous multislice acquisitions. *Magnetic*
615 *Resonance in Medicine*, *72*(1), 93-102.
616
617 Cheung, V.K.M., Meyer, L., Friederici, A.D., & Koelsch, S. (2018) The right inferior
618 frontal gyrus processes nested non-local dependencies in music. *Scientific Reports*, *8*,
619 3822. https://doi.org/10.1038/s41598-018-22144-9
620
621 Chen, J., Leong, Y.C., Honey, C.J., Yong, C.H., Norman, K.A., & Hasson, U. (2017)
622 Shared memories reveal shared structure in neural activity across individuals. *Nature*
623 *Neuroscience*, *20*, 115-125.
624
625 Chen, P., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J.V., & Ramadge, P.J. (2015) A
626 reduced-dimension fMRI shared response model. *Advances in Neural Information*
627 *Processing Systems (NIPS)*.
628
629

630 Daly, I., Williams, D., Hallowell, J., Hwang, F., Kirke, A., Malik, A., Weaver, J., Miranda,
631 E., & Nasuto, S.J. (2015) Music-induced emotions can be predicted from a combination
632 of brain activity and acoustic features. *Brain and Cognition*, *101*, 1-11.
633 https://doi.org/10.1016/j.bandc.2015.08.003
634
635 Farbood, M. M., Heeger, D. J., Marcus, G., Hasson, U., & Lerner, Y. (2015) The neural
636 processing of hierarchical structure in music and speech at different timescales.
637 *Frontiers in Neuroscience*, *9*, 157. doi: 10.3389/fnins.2015.00157
638
639 Gingras, B., Pearce, M.T., Goodchild, M., Dean, R.T., Wiggins, G., & McAdams, S.
640 (2016) Linking melodic expectation to expressive performance timing and perceived
641 musical tension. *Journal of Experimental Psychology: Human Perception and*
642 *Performance*, *42*(4), 594-609.
643
644 Gordon, C.L. (2018) Recruitment of the motor system during music listening: An ALE
645 meta-analysis of fMRI data. *PLoS One*, *13*(11).
646 https://doi.org/10.1371/journal.pone.0207213
647
648 Hasson, U., Chen, J., & Honey, C.J. (2015). Hierarchical process memory: memory as
649 an integral component of information processing. *Trends in Cognitive Science*, *19*, 304-
650 313.
651

652 Jackendoff, R. & Lerdahl, F. (2006) The capacity for music: what is it and what's special
653 about it? *Cognition*, *100*, 33–72.
654
655 Jenkinson, M., Bannister, P.R., Brady, J.M., & Smith, S.M. (2002) Improved optimisation
656 for the robust and accurate linear registration and motion correction of brain images.
657 *NeuroImage*, *17*(2), 825-841.
658
659 Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S.V., McDermott, J.H.
660 (2018) A task-optimized neural network replicates human auditory behavior, predicts
661 brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630-644.
662 https://doi.org/10.1016/j.neuron.2018.03.044.
663
664 Koelsch, S., Gunter, T.C., Cramon, D., Zysset, S., Lohmann, G., Friederici, A.D. (2002).
665 Bach speaks: A cortical 'language-network' serves the processing of music.
666 *NeuroImage*, *17*, 956–966.
667
668 Koelsch, S. (2011) Toward a neural basis of music perception - a review and updated
669 model. *Frontiers in Psychology*, *2*, 110.
670
671 Kumar, M., Ellis, C.T., Lu, Q., Zhang, H., Capota, M., Willke, T.L., Ramadge, P.J., Turk-
672 Browne, N.B., Norman, K.A. (2020a) BrainIAK tutorials: User-friendly learning materials
673 for advanced fMRI analysis. *PLOS Computational Biology*, *16*(1).
674 https://doi.org/10.1371/journal.pcbi.1007549

675 Kumar, M., Anderson, M.J., Antony, J.W., Baldassano, C., Brooks, P.P., Cai, M.B.,
676 Chen, P.H.C., Ellis, C.T., Henselman-Petrusek, G., Huberdeau, D., Hutchinson, J.B., Li,
677 P.Y., Lu, Q., Manning, J.R., Mennen, A.C., Nastase, S.A., Hugo, R., Schapiro, A.C.,
678 Schuck, N.W., Shvartsman, M., Sundaram, N., Suo, D., Turek, J.S., Vo, V.A., Wallace,
679 G., Wang, Y., Zhang, H., Zhu, X., Capota, M., Cohen, J.D., Hasson, U., Li, K.,
680 Ramadge, P.J., Turk-Browne, N.B., Willke, T.L. & Norman, K.A. (2020b). BrainIAK: The
681 Brain Imaging Analysis Kit. *OSF Preprints*.
682
683 Landemard, A., Bimbard, C., Demene, C., Shamma, S., Norman-Haignere, S.,
684 Boubenec, Y. (2020). Distinct higher-order representations of natural sounds in human
685 and ferret auditory cortex. *bioRxiv.*
686
687 Lee, D.J., Jung, H., & Loui, P. (2019). Attention modulates electrophysiological
688 responses to simultaneous music and language syntax processing. *Brain Sciences*,
689 *9*(11), 305. https://doi.org/10.3390/brainsci9110305
690
691 Lehne, M., Rohrmeier, M., & Koelsch, S. (2013). Tension-related activity in the
692 orbitofrontal cortex and amygdala: an fMRI study with music. *Social Cognitive and*
693 *Affective Neuroscience*, *9*(10), 1515-1523. doi: 10.1093/scan/nst141
694
695 Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of
696 a hierarchy of temporal receptive windows using a narrated story. *Journal of*
697 *Neuroscience*, 31, 2906–2915. doi: 10.1523/JNEUROSCI.3684-10.2011

698 Margulis, E.M. (2005). A model of melodic expectation. *Music Perception*, *22*(4), 663-
699 714.
700
701 McFee, B., Colin R., Dawen, L., Daniel, E.P.W., Matt, M., Eric, B., and Oriol, N. (2015)
702 "librosa: Audio and music signal analysis in python." *Proceedings of the 14th Python in*
703 *Science Conference*, 18-25.
704
705 Minati, L., Rosazza, C., D'Incerti, L., Pietrocini, E., Valentini, L., Scaioli, V., Loveday, C.,
706 & Bruzzone, M.G. (2008). fMRI/ERP of musical syntax: comparison of melodies and
707 unstructured note sequences. *Neuroreport*, *19*, 1381–1385.
708
709 Moeller, S., Yacoub, E., Olman, C.A., Auerbach, E., Strupp, J., Harel, N., & Uğurbil, K.
710 (2010) Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial
711 parallel imaging with application to high spatial and temporal whole-brain fMRI.
712 *Magnetic Resonance in Medicine*, *63*(5), 1144-1153.
713
714 Nastase, S.A., Liu, Y., Hillman, H., Norman, K.A., & Hasson, U. (2020). Leveraging
715 shared connectivity to aggregate heterogeneous datasets into a common response
716 space. *NeuroImage*, *217*. https://doi.org/10.1016/j.neuroimage.2020.116865.
717
718 Patel, A.D. (2011) Why would musical training benefit the neural encoding of speech?
719 The OPERA hypothesis. *Frontiers in Psychology*, *2*, 142. doi:10.3389/fpsyg.2011.00142
720

721 Patel, A.D. & Iversen, J.R. (2014) The evolutionary neuroscience of musical beat
722 perception: the Action Simulation for Auditory Prediction (ASAP) hypothesis. *Frontiers in*
723 *Systems Neuroscience*, *8*, 57.
724
725 Peretz, I., Vuvan, D., Lagrois, M., & Armony, J. L. Neural overlap in processing music
726 and speech. *Philosophical Transactions of the Royal Society B*, *370*.
727 http://doi.org/10.1098/rstb.2014.0090.
728
729 Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., & Shulman,
730 G.L. (2001) A default mode of brain function. *PNAS*, *98*(2), 676-682.
731
732 Poorjam, A. H. (2018). Re: Why we take only 12-13 MFCC coefficients in feature
733 extraction?. Retrieved from: https://www.researchgate.net/post/Why_we_take_only_12-
734 13_MFCC_coefficients_in_feature_extraction/5b0fd2b7cbdfd4b7b60e9431/citation/dow
735 nload.
736
737 Setsompop, K., Gagoski, B.A., Polimeni, J.R., Witzel, T., Wedeen, V.J., & Wald, L.L.
738 (2012) Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo
739 planar imaging with reduced *g*-factor penalty. *Magnetic Resonance in Medicine*. *67*(5),
740 1210-1224.
741

742 Shulman, G. L., Fiez, J. A., Corbetta, M., Buckner, R. L., Miezin, F. M., Raichle, M. E. &
743 Petersen, S. E. (1997) Common blood flow changes across visual tasks: II. Decreases
744 in cerebral cortex. *Journal of Cognitive Neuroscience*, 9, 648–663.
745
746 Sotiropoulos, S.N., Moeller, S., Jbabdi, S., Xu, J., Andersson, J.L., Auerbach, E.J.,
747 Yacoub, E., Feinberg, D., Setsompop, K., Wald, L.L., Behrens, T.E.J, Uğurbil, K., &
748 Lenglet, C. (2013) Effects of image reconstruction on fiber orientation mapping from
749 multichannel diffusion MRI: Reducing the noise floor using SENSE. *Magnetic*
750 *Resonance in Medicine*, *70*(6),1682-1689.
751
752 Sridharan, D., Levitin, D.J., Chafe, C.H., Berger, J., & Menon, V. (2007) Neural
753 dynamics of event segmentation in music: converging evidence for dissociable ventral
754 and dorsal networks. *Neuron*, *55*(3), 521-532.
755
756 Tallal, P. & Gaab, N. (2006) Dynamic auditory processing, musical experience and
757 language development. *Trends in Neuroscience*, *29*(7), 382-390.
758
759 Tillmann, B., Koelsch, S., Escoffier, N., Bigand, E., Lalitte, P., Friederici, A.D., &
760 Cramon, D.Y. (2006) Cognitive priming in sung and instrumental music: Activation of
761 inferior frontal cortex. *NeuroImage*, 31, 1771–1782.
762

763 Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal
764 autocorrelation in univariate linear modeling of fMRI data. *NeuroImage*, *14*(6), 1370–
765 1386. http://doi.org/10.1006/nimg.2001.0931
766
767 Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F., Jenkinson, M., & Smith, S. M.
768 (2004). Multilevel linear modelling for fMRI group analysis using Bayesian inference.
769 *NeuroImage*, *21*(4), 1732–1747. http://doi.org/10.1016/j.neuroimage.2003.12.023
770
771 Xu, J., Moeller, S., Auerbach, E.J., Strupp, J., Smith, S.M., Feinberg, D.A., Yacoub, E.,
772 & Uğurbil, K. (2013) Evaluation of slice accelerations using multiband echo planar
773 imaging at 3 T. *NeuroImage*, *83*(0), 991-1001.
774
775 Zadbood, A., Chen, J., Leong, Y.C., Norman, K.A., & Hasson, U. (2017) How we
776 transmit memories to other brains: constructing shared neural representations via
777 communication. *Cerebral Cortex*, *27*(10), 4988-5000. 10.1093/cercor/bhx202
778
779 Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in*
780 *Psychological Science*, *16*(2), 80–84. http://doi.org/10.1111/j.1467-8721.2007.00480.x
781
782 Zatorre, R.J., Belin, P., & Penhune, V.B. (2002) Structure and function of auditory
783 cortex: music and speech. *Trends in Cognitive Sciences*, *6*(1), 37-46.
784

785 Zatorre, R., Chen, J. & Penhune, V. (2007) When the brain plays music: auditory–motor
786 interactions in music perception and production. *Nature Reviews Neuroscience*, *8*, 547–
787 558. https://doi.org/10.1038/nrn2152.

788

789

790 **Acknowledgments**

791 We thank Mark A. Pinsk for contributing to the *Scanning parameters and preprocessing*
792 section of the manuscript, Benson Deverett for helping with the stimulus presentation
793 script in python, Elizabeth McDevitt for suggestions on the figures, Sara Chuang for
794 helping with stimulus selection, and the members of the Hasson, Pillow, and Norman
795 labs for their comments and support. This work was supported by NIMH R01
796 MH112357-01 to UH and KAN and NINDS D-SPAN award F99 NS118740-01 to JW.