Music-selective cortex is sensitive to structure in both pitch and time

Dana L. Boebinger[1,2,3], Sam V. Norman-Haignere[5,6,7]
Josh H. McDermott[1,2,3,4], Nancy G. Kanwisher[2,3,4]

[1] *Speech and Hearing Bioscience and Technology, Harvard University, Cambridge MA 02138 USA*
[2] *Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge MA 02139 USA*
[3] *McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge MA 02139 USA*
[4] *Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge MA 02139 USA*
[5] *Zuckerman Institute for Brain Research, Columbia University, New York NY 10027 USA*
[6] *Department of Biostatistics & Computational Biology, University of Rochester, Rochester NY 14642 USA*
[7] *Department of Neuroscience, University of Rochester, Rochester NY 14642 USA*

ABSTRACT

Converging evidence suggests that neural populations within human non-primary auditory cortex respond selectively to music. These neural populations respond strongly to a wide range of music stimuli, and weakly to other natural sounds and to synthetic control stimuli matched to music in many acoustic properties, suggesting that they are driven by high-level musical features. What are these features? Here we used fMRI to test the extent to which musical structure in pitch and time contribute to music-selective neural responses. We used voxel decomposition to derive music-selective response components in each of 15 participants individually, and then measured the response of these components to synthetic music clips in which we selectively disrupted musical structure by scrambling either the note pitches and/or onset times. Both types of scrambling produced lower responses compared to when melodic or rhythmic structure was intact. This effect was much stronger in the music-selective component than in the other response components, even those with substantial spatial overlap with the music component. We further found no evidence for any cortical regions sensitive to pitch but not time structure, or vice versa. Our results suggest that the processing of melody and rhythm are intertwined within auditory cortex.

Keywords: auditory cortex, fMRI, melody, music, rhythm

INTRODUCTION

Growing evidence suggests the existence of neural populations in bilateral non-primary auditory cortex that respond selectively to music and thus seem likely to figure importantly in musical perception and behavior (Leaver and Rauschecker, 2010; Rogalsky et al., 2011; Fedorenko et al., 2012; Tierney et al., 2013; LaCroix et al., 2015; Norman-Haignere et al., 2015, 2021). The clearest initial evidence for these neural populations came from a component analysis of fMRI responses. This analysis revealed a response component located in anterior and posterior regions of the superior temporal gyrus (STG) that responded more strongly to music than to other real-world sounds (Norman-Haignere et al., 2015). The finding of music selectivity and its anatomical distribution has since been replicated in two other participant groups (Boebinger et

al., 2021). In addition, the music component's response is much lower to synthetic control stimuli matched to music in many acoustic properties, even though these synthetic stimuli produce very similar responses in primary auditory cortex (Norman-Haignere and McDermott, 2018). These results have recently been further confirmed by intracranial recordings, which show individual electrodes with clear selectivity for music (Norman-Haignere et al., 2021). However, the musical features that drive this response remain unknown.

Melody and rhythm are often assumed to be distinct organizational dimensions of music, both colloquially and within the field of music theory (Jackendoff and Lerdahl, 2006). Consistent with this idea, numerous examples exist of neuropsychological patients with focal brain damage who show selective deficits in either melody or rhythm perception (Grant-Allen, 1878; Peretz, 1990; Peretz and Kolinsky, 1993; Peretz et al., 2002, 1994; Liégeois-Chauvel et al., 1998; Ayotte et al., 2000; Piccirilli et al., 2000; Peretz and Coltheart, 2003; Peretz and Hyde, 2003; Di Pietro et al., 2004; Hyde and Peretz, 2004). While these findings in "amusia" patients suggest that distinct neural mechanisms underlie processing of these two domains, this conclusion is complicated by the fact that many of these patients display deficits in other non-musical domains (Brust, 1980; Mavlov, 1980; Mazzucchi et al., 1982; Tanaka et al., 1987; Mendez and Geehan, 1988; Tramo et al., 1990, 2002; Eustache et al., 1990; Fries and Swihart, 1990; Hofman et al., 1993; Peretz et al., 1994; Johkura et al., 1998; Piccirilli et al., 2000; Di Pietro et al., 2004; Hattiangadi et al., 2005; Sihvonen et al., 2016), and in some cases their musical deficits can be explained in terms of more basic perceptual deficits (Tanaka et al., 1987; Fries and Swihart, 1990; Johannes et al., 1998; Johnsrude et al., 2000; Tramo et al., 2002; Wilson et al., 2002; Hattiangadi et al., 2005). To definitively establish a neural dissociation between the processing of high-level musical structure in pitch versus time, it would be necessary to demonstrate a selective deficit in one but not the other domain that is limited to music, and which cannot be explained by lower-level perceptual deficits or other cognitive domains. Although some cases come close to meeting these criteria (e.g. patient C.N. described in Peretz and Kolinsky, 1993; Peretz et al., 1994; Peretz, 1996; Dalla Bella and Peretz, 1999), to our knowledge no such fully unambiguous cases have been reported.

Indeed, a large body of behavioral research has shown that processing of melodic and rhythmic structure are not fully separable, but rather are integrated to some extent in both perception and memory (Jones et al., 1982, 1987; Kidd et al., 1984; Boltz, 1999; Prince et al., 2009; Prince, 2011; but see e.g. Palmer and Krumhansl, 1987). Consistent with this idea, prior neuroimaging studies have found similar brain regions in the temporal lobes to be involved when making same-different judgements about either the pitch or duration of random tone sequences (Griffiths et al., 1999), as well as when detecting melodic or rhythmic errors in short tonal melodies (Krumhansl and Zatorre, 2003a). Similarly, brain regions that respond more to intact than scrambled music show sensitivity to both pitch and rhythm scrambling (Fedorenko et al., 2012). On the other hand, multiple studies have found that different cortical regions are sensitive to aspects of spectrotemporal modulation that may be related to melody and rhythm: anterior regions show selectivity for lower temporal and higher spectral modulation rates, and posterior regions show selectivity for high temporal and low spectral modulation rates (Schönwiesner and Zatorre, 2009; Santoro et al., 2014; Norman-Haignere et al., 2015; Hullett et

al., 2016; Hamilton et al., 2018). However, modulation tuning is unable to explain non-primary cortical responses like those selective for music (Norman-Haignere and McDermott, 2018), and thus the relationship between modulation tuning and higher-order melodic and rhythmic structure remains unclear.

While a small handful of neuroimaging studies suggest some degree of integrated processing of the pitch and temporal structure of music, those studies did not establish whether the brain regions involved were truly selective for music. Thus, it is unclear whether the observed effects are related specifically to the processing of music, or whether they instead reflect more general acoustic features or cognitive processes. One challenge is that true music selectivity is generally weak when measured using standard fMRI methods, likely as a consequence of the coarse spatial resolution of fMRI and the fact that voxels pool responses from hundreds of thousands of neurons that often have different tuning properties. The component analysis from our previous study (Norman-Haignere et al., 2015) attempts to overcome these challenges by unmixing overlapping response patterns within voxels in order to reveal a response component with clear selectivity for music.

In the current study, we examined the extent to which cortical music selectivity reflects sensitivity to pitch and temporal structure in music. We measured fMRI responses to synthetic MIDI-generated music clips in which either the pitch or temporal structure of the music was corrupted. Then, using methods from our previous study (Norman-Haignere et al., 2015), we measured the effect of these scrambling manipulations on the music-selective component of auditory cortical responses.

METHODS

Participants

Fifteen young adults (7 self-identified female, 8 self-identified male, 0 self-identified non-binary; mean age = 25.6 years, SD = 4.5 years) participated in the experiment. Of these participants, 12 had some degree of musical training (mean = 10.1 years of training, SD = 6.7). All participants were fluent English speakers and reported normal hearing. The study was approved by the Committee on the Use of Humans as Experimental Subjects (COUHES) at MIT, and written informed consent was obtained from all participants. Participants were compensated with an hourly wage for their time.

Overview of experimental design

For the current study, participants completed multiple scanning sessions, which included a set of fMRI “localizer” scans to estimate each participant’s voxel weights for the components from our previous study (Norman-Haignere et al., 2015), and the main fMRI experiment in which responses were measured to different music conditions. For the component localizer, participants were scanned while they listened to a set of 30 natural sounds. Based on the voxel

responses to these sounds, we then estimated each participant's voxel weights that best approximated the previously inferred response profiles from the original study (Norman-Haignere et al., 2015). We then applied these weights to participants' voxel responses in the main experiment, in which they listened to synthetic music clips that had been manipulated by scrambling either the note pitches and/or onset times. This process allowed us to examine how the response components from our previous study (Norman-Haignere et al., 2015) respond to the new stimulus conditions, and thus the extent to which music selectivity reflects pitch vs. temporal structure in music.

Component localizer

Overview of voxel decomposition

Because a single fMRI voxel pools responses from hundreds of thousands of neurons, voxel responses may be modeled as the sum of the responses of multiple underlying neuronal populations. In our previous work, we found that voxel responses to natural sounds in auditory cortex can be approximated by a weighted sum of six canonical response profiles, or "components:"

$$\boldsymbol{v}_i \approx \sum_{k=1}^{6} \boldsymbol{r}_k w_{k,i}$$

(Eq. 1)

where $\boldsymbol{v}_i$ is the response (a vector) of a single voxel to the sound set, $\boldsymbol{r}_k$ represents the kth component response profile across sounds, which is shared across all voxels, and $w_{k,i}$ represents the weight (a scalar) of component k in voxel i. If we concatenate the responses of many voxels into a sound x voxel data matrix, D, this approximation corresponds to matrix factorization:

$$D \approx RW$$

(Eq. 2)

where R is a sound x component response matrix and W is a component by voxel weight matrix. The component responses and weights were jointly inferred in our prior studies by maximizing the non-Gaussianity of the weights across voxels, akin to standard algorithms for independent components analysis (ICA). The voxel decomposition method is described in detail in our previous paper (Norman-Haignere et al., 2015) and code is available online (https://github.com/snormanhaignere/nonparametric-ica).

Deriving components from scratch requires a large amount of data because both the component responses (R) and weights (W) need to be inferred, which requires using statistics computed across thousands of voxels (requiring multiple participants). However, if the component responses are known (e.g. from a prior study), the voxel weights can be inferred in a single new participant using linear regression (separately for each voxel). In addition, once the component

voxel weights are known, the component responses to a new set of sounds, measured in the participants for whom the weights are known, can also be easily inferred using linear regression (separately for each sound). This latter procedure is analogous to identifying a set of voxels that respond to a particular 'localizer' contrast, and then measuring their response to a new set of stimuli.

Stimuli

To reduce the amount of scan time needed to infer the component weights, we chose a subset of 30 sounds from our original 165-sound set that were best able to identify the six components. These sounds were selected by greedily discarding sounds so as to minimize the expected variance of the inferred component weights. Assuming isotropic voxel noise, the expected variance of the weights, inferred by ordinary least-squares (OLS) regression, is proportional to (Dale, 1999):

$$\mathrm{trace}((R^T R)^{-1})$$

(Eq. 3)

where R is the sound x component response matrix. Intuitively, the optimization chooses sounds that have high response variance for each component (e.g., lots of music and non-music sounds for the music-selective component), and where that response variance is relatively uncorrelated across components. Each column/component of R was z-scored prior to performing our greedy optimization (which amounts to iteratively discarding rows of R to minimize the equation above).

Stimulus presentation and scanning procedure

The data for the component localizer came from previous studies on the same participants (n = 9) and from new participants scanned for this study (n = 6). As a result, there were three versions of the component localizer, differing in the minor details of stimulus presentation and scanning parameters (see **Table 1**), but which otherwise were very similar. All three versions contained the same set of 30 2-second natural sounds. In all versions, stimuli were presented during scanning in a "mini-block design," in which each 2-second sound was repeated multiple times in a row. For versions 1 and 2, stimuli were repeated 3 times in a row, and 5 times for version 3. During scanning, stimuli were presented over MR-compatible earphones (Sensimetrics S14) at 75 dB SPL. Each stimulus was presented in silence, with a single fMRI volume collected between each repetition (i.e. "sparse scanning"; Hall et al., 1999). To encourage participants to pay attention to the sounds, either the second or third repetition in each "mini-block" was 8dB quieter (presented at 67 dB SPL), and participants were instructed to press a button when they heard this quieter sound. Data pre-processing, denoising, and initial GLM analyses were the same as in the main experiment (see "*fMRI data acquisition and analysis*" below).

Component weight estimation

To estimate each participant's component weights (W), we multiplied each participant's data matrix of voxel responses to the 30 natural sound stimuli (D_{30}; sound x voxel) by the

pseudoinverse of the known component response matrix for those same 30 sounds (R_{30}; sound x component):

$$W = ({R_{30}}^T R_{30})^{-1} {R_{30}}^T D_{30}$$

(Eq. 4)

Multiplying by the pseudoinverse is equivalent to performing ordinary least-squares regression. For visualization purposes (e.g. **Figures 2B** and **3A**), component weights were averaged across participants in standardized anatomical coordinates (Freesurfer's FsAverage template), and plotted with a color scale that spans the central 95% of the weight distribution for each component.

Localizer	Version 1	Version 2	Version 3
MRI scanner	Trio	Prisma	Trio
Head coil	32-channel	32-channel	32-channel
TR	3.59	3.4	3.4
TA	1.19	1.02	1
TE	30ms	33ms	30ms
Flip angle	90	90	90
Number of slices	45 (whole-brain)	48 (whole-brain)	15 (partial coverage)
Slice thickness	2.8 mm	3 mm	4 mm
Slice gap	10%	10%	10%
Matrix	96 x 96	96 x 96	96 x 96
Voxel size	2 x 2 x 2.8 mm	2.1 x 2.1 x 3 mm	2.1 x 2.1 x 4 mm
Simultaneous multi-slice acceleration (SMS)	3	4	none
Number of stimuli	30	192	165
Number of stimulus rep. in each "mini-block"	3	3	5
Number of stimulus blocks per run	30	24	15
Run duration	6.5 min	5.5 min	5.4 min
Number of runs	10	48	33
Number of stimulus repetitions	10	6	3
Number of participants who completed this version	6	7	2

Table 1. Scan parameters for the three versions of the component localizer scan sessions.

Component response estimation

The goal of the current study was to then use participants' inferred component weights to estimate the component responses to each of the stimulus conditions in the main experiment. To do this, we multiplied participants' data matrix of voxel responses to the experimental stimuli (D_{exp}) by the pseudoinverse of the estimated component weights:

$$R_{exp} = D_{exp} W^T (W W^T)^{-1}$$

(Eq. 5)

Unless otherwise noted in the text, we included all voxels in a large anatomical mask region encompassing bilateral superior temporal and posterior parietal cortex (shown as colored regions in **Figures 2B** and **3A**, and white outlines in **Figures 3A** and **5A** & **C**). For analyses comparing different cortical regions (e.g. right and left hemispheres), we estimated the component response for each cortical region separately, using the voxel responses and estimated component weights from the corresponding subset of voxels.

Main experiment

Stimuli

The stimuli for the main experiment consisted of 10-second clips of intact and scrambled versions of synthetic music, as well as several non-music control conditions. Specifically, we created versions of each tonal melody in which we manipulated (1) the note pitches that made up each melody, (2) the note onset times within the melody, and (3) both the note pitches and onset times. So that we could examine the effect of manipulating rhythmic structure in the absence of pitch structure, we also included a set of synthetic drum stimuli and scrambled versions of each of those stimuli in which we manipulated the note onset times.

To construct these stimuli, we first selected a set of 20 melodies from the stimulus set used in a previous study by our group (Fedorenko et al., 2012). These melodies originated from polyphonic musical pieces from a variety of genres common in modern Western musical traditions, each with complex melodic and rhythmic content, and were rendered in MIDI (musical instrument digital interface) format. Each piece contained multiple instrumental tracks, which were all converted to a piano timbre for use in the experiment. The keys of the melodies were assigned by drawing (without replacement) from a pool of all 12 major and 12 minor keys (preserving the mode, i.e. "major" vs. "minor"), and the key of each melody was then changed by adding/subtracting a constant (corresponding to the difference in semitones between the original and desired keys) to the note pitches. As a result, the 14 major melodies included in the stimulus set spanned all 12 major keys (with C# and G# major occurring twice), and the 6 minor melodies were all in different keys.

We then created three additional versions of each melody (20 unique melodies x [1 intact + 3 scrambled versions] = a total of 80 MIDI melodies), in which the musical structure was manipulated by scrambling either the pitches and/or timing of the notes (**Figure 1**). MIDI represents music as a matrix of notes, with different columns specifying the onset, duration,

pitch, and instrumentation of the notes, which makes it possible to manipulate these features independently and to scramble across-note structure without altering acoustic properties of individual notes. In all cases, within-chord structure was preserved by performing the scrambling operation on clusters of notes with a similar onset time (defined for our purposes as within 0.1 beats of each other), rather than on individual notes (**Figure 1A** & **B**). This is a more conservative scrambling procedure than was used in our previous study using related methods (Fedorenko et al., 2012). For the pitch-scrambled melodies, each note cluster was shifted either up or down by 0-2 semitones (sampled from a discrete, uniform distribution), which disrupted both melodic and across-chord harmonic structure (**Figure 1C** & **D**). For time-scrambled melodies, the inter-onset intervals between note clusters were jittered by adding an amount drawn from a uniform distribution spanning ±0.25 beats (the tempi of the MIDI stimuli ranged from 72 to 220 bpm, so 0.25 beats corresponded to between 68-208ms depending on the stimulus example), but with the constraint that no inter-onset interval could be shorter than 0.1 beats. To ensure that the total duration of the time-scrambled melody was the same as the original version, we divided the vector of jittered inter-onset intervals by its mean and then multiplied it by the mean of the original vector of inter-onset intervals. Next, the jittered inter-onset intervals were randomly reordered, and the note onset times of the scrambled piece were derived from these reordered inter-onset intervals (via cumulative summation). This process effectively eliminated any isochronous beat or metrical hierarchy (**Figure 1E**). For the melodies that were both pitch- and time-scrambled, they first went through the time-scrambling procedure and then the pitch-scrambling procedure. This set of 80 MIDI melodies was supplemented by a set of 20 MIDI drum stimuli, chosen to represent a variety of genres, and versions of these melodies that were time-scrambled following the procedure described above (**Figure 1F**), for a total of 40 MIDI drum stimuli. Each MIDI file was converted into audio using AppleScript and QuickTime Player 7 (Ellis, 2014).

To complement these 120 MIDI music stimuli, we selected 16 additional non-music stimuli to serve as controls for assessing selectivity. Specifically, we chose 4 English speech excerpts (taken from audiobook recordings and radio broadcasts), 4 foreign speech excerpts (German, Hindi, Italian, Russian), 4 animal vocalizations (chimpanzees, cow, sheep, wolf), and 4 environmental sounds (footsteps, ping-pong ball, power saw, copy machine). All stimuli were resampled to 20 kHz, trimmed to be 10 seconds long, windowed with a 25ms half-Hanning-window ramp, and normalized to have the same RMS.

Because the MIDI scrambling manipulations were intended to be subtle and to preserve some surface aspects of musical structure (e.g. consonance/dissonance amongst simultaneous notes within a chord), an online experiment (via Amazon's Mechanical Turk) was used to ensure that the manipulation produced detectable changes in musical structure. In this experiment, participants listened to the same 120 MIDI stimuli used in the full experiment (n = 46; 16 self-identified female, 27 self-identified male, 0 self-identified non-binary, and 3 did not answer; mean age = 42.9 years, SD = 11.9 years; an additional 7 participants completed the experiment but were excluded due to performance of <60% in every condition). Stimuli were presented in pairs, with each trial containing an intact and scrambled version of the same MIDI stimulus. Participants were instructed that one of the stimuli was "corrupted by having the melody and/or

the note timing disrupted," and were told to choose which of the two melodies had been "corrupted." Stimuli were blocked by condition, so that each participant completed four blocks of 20 trials each (i.e. intact vs. pitch-scrambled, intact vs. time-scrambled, intact vs. pitch-and-time-scrambled, and intact vs. time-scrambled drums); participants were told which type of "corruption" to listen for in each block. Prior to completing the experiment participants completed a headphone check to help ensure they were wearing headphones or earphones (Woods et al., 2017); we have found that this headphone check is typically sufficient to obtain performance on par with that of experiments conducted in the lab under more controlled listening conditions (Woods and McDermott, 2018; McPherson et al., 2020; Traer et al., 2021).

Performance was well above chance in all four blocks, indicating that the scrambling manipulation was clearly detectable in all conditions (pitch-scrambling: median = 87.5%, interquartile range or IQR = 20%; time-scrambling: median = 100%, IQR = 5%; pitch-and-time-scrambling: median = 100%, IQR = 5%; time-scrambling drums: median = 100%, IQR = 5%; one-sample Wilcoxon signed rank tests comparing percent correct to chance performance of 50%; all p's < 1e-08, two-tailed).

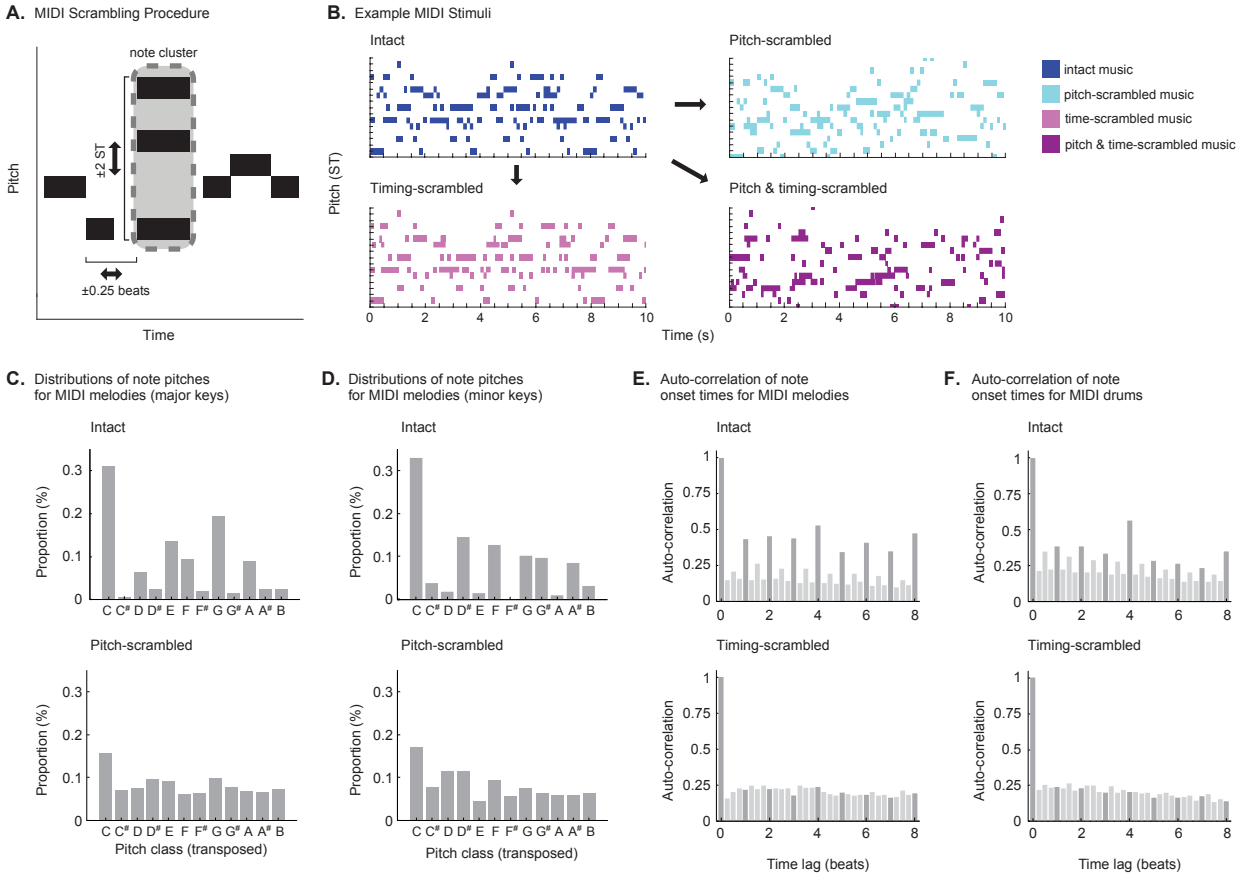

Figure 1. Scrambling procedure for MIDI music stimuli. **A.** Schematic of scrambling procedure. First, notes with the same onset (within 0.1 beats) were grouped together, and scrambled as a unit. For pitch-scrambling, each note cluster was shifted by up to ±2 semitones. For timing-scrambling, the inter-onset intervals (IOIs) between note clusters were jittered by adding an amount drawn from a uniform distribution spanning ±0.25 beats (with the

constraint that no IOI could be shorter than 0.1 beats). The jittered IOIs were normalized to ensure the total duration of the timing-scrambled melody was the same as the original intact version and were randomly reordered. For the melodies that were both pitch- and timing-scrambled, they first went through the timing-scrambling procedure and then the pitch-scrambling procedure. **B.** Intact and scrambled versions of an example MIDI melody stimulus. **C.** Histograms showing the distribution of note pitch classes averaged across the 14 MIDI melodies in major keys, in both their intact (top) and pitch-scrambled (bottom) versions. For the purpose of this analysis, all melodies were transposed to C major. Pitch class distributions were weighted by note durations, as computed using the *"pcdist1"* function in the MATLAB MIDI Toolbox (Eerola and Toiviainen, 2004). For the intact melodies (top), certain pitch classes are more common than others, reflecting the tonal hierarchy typical of major keys in Western music. The pitch-scrambling procedure (bottom) disrupts melodic and tonal structure, leveling out the distribution of pitch classes to a certain extent. **D.** Same as **C**, but for the 6 MIDI melodies in minor keys. **E.** Histograms showing the auto-correlation of note onset times in units of beats averaged across all 20 intact (top) MIDI melodies and their timing-scrambled versions (bottom). Onset times were weighted by note durations, as computed using the *"onsetacorr"* function in the MATLAB MIDI Toolbox (Eerola and Toiviainen, 2004). For the intact melodies (top), peaks can be seen on every beat (darker gray bars) that reflect the isochronous beat and metrical structure typical of Western music. The timing-scrambling procedure (bottom) disrupts this metrical structure, resulting in no clear peaks in the auto-correlation function at multiples of the beat. **F.** Same as **E**, but for the 20 MIDI drum stimuli.

Stimulus presentation and scanning procedure

During the scanning session for the main experiment, auditory stimuli were presented over MR-compatible earphones (Sensimetrics S14) at 80 dB SPL. The 10-second stimuli were presented back-to-back with a 500ms inter-stimulus interval. To encourage participants to pay attention to the stimuli, each stimulus either ramped up (by 6 dB) or down (by 9 dB) in level over 1 second starting at the 5s point, and the participant indicated whether each stimulus got "louder" or "quieter" via button press. The increment and decrement values were chosen to approximately equate the subjective salience of the level change. Participants' average performance on this task was 97.1% correct (SD = 2%), and performance never fell below 85% correct for any participant on any run.

The scanning session for the main experiment consisted of twelve 7.3-minute runs (all participants completed 12 runs, except for one participant who only completed 8 runs), with each run consisting of thirty-four 10-second stimuli and six 10-second silent periods during which no sound was presented. These silent blocks were the same duration as the stimuli, and were distributed randomly throughout each run, providing a baseline. The thirty-four stimuli presented in each run consisted of five stimuli from of each of the MIDI music conditions ("intact," "pitch-scrambled," "time-scrambled," "pitch-and-time-scrambled," "intact drums," and "time-scrambled drums"), and one stimulus each from the four non-music control conditions ("English speech," "foreign speech," "animal sounds," "environmental sounds"). To maximize the temporal interval between a given MIDI music stimulus and its scrambled versions, only one version of a given MIDI stimulus (i.e. intact, pitch-scrambled, time-scrambled, pitch-and-time-scrambled) was presented per run for all but three participants (the remaining three participants were run on an earlier version of the experiment that was identical except without the constraint maximizing the temporal interval between intact and scrambled versions of a given MIDI stimulus, and with an extra null block in each run). The full set of 136 stimuli was repeated every four runs, meaning that each individual stimulus was presented a total of three times during the

course of the 12 experimental runs (twice for the subject who only completed 8 runs). The entire scanning session lasted approximately 2 hours.

fMRI data acquisition and analysis

MRI data were collected at the Athinoula A. Martinos Imaging Center of the McGovern Institute for Brain Research at MIT, on a 3T Siemens Trio with a 32-channel head coil. Because the sound stimuli were long (10 seconds), we used continuous instead of sparse scanning (TR = 2.1 sec, TE = 30ms, 90 degree flip angle, 3 discarded initial acquisitions). Each functional acquisition consisted of 46 roughly axial slices (oriented parallel to the anterior-posterior commissure line) covering the whole brain, with voxel size 2 x 2 x 2.8mm (96 x 96 matrix, 10% slice gap). A simultaneous multi-slice (SMS) acceleration factor of 2 was used in order to minimize acquisition time. To localize functional activity, a high-resolution anatomical T1-weighted MPRAGE image was obtained for each participant (1 mm isotropic voxels).

Preprocessing and data analysis were performed using FSL software and custom Matlab scripts. Functional volumes were motion-corrected, slice-time-corrected, skull-stripped, linearly detrended, and aligned to each participant's anatomical image (using FLIRT and BBRegister; Greve & Fischl, 2009; Jenkinson & Smith, 2001). Motion correction and functional-to-anatomical registration was done separately for each run. Each participant's preprocessed data were then resampled to the cortical surface reconstruction computed by FreeSurfer (Dale et al., 1999), registered to the FsAverage template brain, and then smoothed on the surface using a 3mm FWHM kernel to improve the signal-to-noise ratio (SNR) of the data by removing local, uncorrelated noise.

GLM-denoise (Kay et al., 2013) was used to further improve SNR and then estimate the response to each of the stimulus conditions using a general linear model (GLM). A separate boxcar regressor was used for each of the 10 experimental conditions, which were convolved with a hemodynamic response function (HRF) that specifies the shape of the BOLD response. GLM-denoise assumes a common HRF that is shared across all conditions and voxels, the shape of which is estimated from the data using an iterative linear fitting procedure (for more details, see Kay et al., 2013).

To remove large-scale correlated noise across voxels, GLM-denoise also includes noise regressors based on principal components analysis (PCA). These PCA-based regressors were derived from the time series of a "noise pool" consisting of voxels with <0% of variance explained by the task regressors (R^2 was measured using leave-one-run-out cross-validation, which is negative when the null hypothesis that all beta weights equal 0 outperforms the model on left-out data). The optimal number of noise regressors was determined by systematically varying the number of principal components included in the model, and estimating the R^2 of each model using leave-one-run-out cross-validation.

GLM-denoise then performs a final fit of the model, and estimates of beta weight variability are obtained by bootstrapping across runs 100 times. The final model estimates are taken as the median across bootstrap samples, and these beta weights were converted to units of percent

BOLD signal change by dividing by the mean signal intensity in each voxel. For analyses involving component responses (see "*Music-selective component is sensitive to both pitch and temporal structure*" section of the Results), GLM-denoise was applied to the data from all runs for each participant. For the standard methods (see "*Standard functional region-of-interest (fROI) analyses also show no dissociation of sensitivity to pitch vs. temporal structure in music*" section of the Results), each participant's data was split into four 75%/25% partitions and GLM-denoise was applied to each partition independently (because the fROI analyses require independent subsets of data to select voxels and analyze responses).

The resulting beta weights were downsampled to a 2mm isotropic grid on the Freesurfer-flattened cortical surface. To accommodate the partial-brain coverage for two of the participants (see "version 3" in **Table 1**), we limited all analyses to voxels within a large anatomical mask region encompassing bilateral superior temporal and posterior parietal cortex (shown as colored regions in **Figures 2B** and **3A**, and white outlines in **Figures 4C** & **D**). All subsequent analyses were conducted in this downsampled 2D surface space, and for ease of description we refer to the elements as "voxels" throughout the paper.

Anterior and posterior auditory cortex anatomical ROIs

To quantify the effect of scrambling musical structure separately for different subregions of auditory cortex, we divided voxels into anterior and posterior anatomical ROIs. We divided auditory cortex in half by drawing a straight line on the downsampled 2D cortical, approximately halfway along Heschl's gyrus (HG) and roughly perpendicular to STG (see white dashed line in **Figure 2B**).

Functional region-of-interest analyses

To complement the component methods, we used more standard methods to select the most music-selective voxels and examine their response. To define fROIs in each individual participant, we partitioned their data into four splits (each consisting of 3 runs, except for one participant who only completed 8 total runs and whose splits each consisted of 2 runs). Then, leaving out one split each time, we estimated beta weights to each of the stimulus conditions in the other 75% of the data using a GLM (Kay et al., 2013), and computed a "music vs. non-music" contrast between the "intact" MIDI condition and the four non-music conditions. We chose this particular contrast because it was unclear how to handle the scrambled MIDI conditions, but we note that the results did not change appreciably if a different "music vs. non-music" contrast (e.g. the 6 MIDI music conditions vs. the 4 non-music conditions) was used. Then, for each split of the data separately, we selected the top 10% most significant voxels from within the large anatomical mask region described previously (see *"fMRI data acquisition and analysis"* section of Methods) to create a binary activation map. We further required that voxels responded to intact MIDI music significantly above baseline, using a lenient threshold of $p < 0.01$ (uncorrected).

We then derived group-level anatomical constraint parcels within which the most music-selective voxels would lie for most individual participants. To do this, we applied a method similar to that used in many previous studies from our group and others (Fedorenko et al., 2010; Julian et al., 2012; Nieto-Castañón and Fedorenko, 2012; Fischer et al., 2016). Specifically, we overlaid individual participants' binary activation maps on top of each other, and then spatially smoothed this overlap map with a Gaussian filter (8 mm FWHM) and thresholded it so that the map contained only those voxels with at least 10% overlap across participants. This group-level parcel contained an average of 496 voxels (averaged across the four splits of data; SD = 29 voxels) (**Figure 4A**). Note that the large amount of smoothing and the lenient overlap threshold were chosen in order to find an anatomical parcel within which most individual participants' most significant voxels will be located.

Next, we selected each participant's top 10% most significant voxels (in the music versus nonmusic contrast) from within this anatomical group-level parcel, using the same data from that individual that was used to derive the group anatomical parcel. This method allows the exact voxels selected to vary across participants, but ensures that they are located in the same general region of cortex. Finally, we then measured the response of these voxels in independent data (the 25% of each participant's data that was not used to select voxels), and measured the sensitivity of these voxels to scrambling in both pitch and time (see "*Statistics"* section for details).

We also ran two fROI analyses to look in a more targeted way for a dissociation between sensitivity to pitch and temporal musical structure, using contrasts that would maximally isolate the effect of the pitch-scrambling and time-scrambling manipulations. For the sake of brevity, throughout this paper we will refer to these as the "melody-sensitive fROI" and the "rhythm-sensitive fROI," cognizant that the contrasts may not isolate "melody" and "rhythm" given that note timing likely contributes to the perception of melody/harmony, and that pitch structure likely influences the perception of rhythm (Jones et al., 1982; Kidd et al., 1984; Schmuckler and Boltz, 1994; Crowder and Neath, 1995; Prince et al., 2009; Prince, 2011). Because of the conditions used in this experiment, there were many pairwise contrasts that could be used for this purpose. In this case, we used "time-scrambled vs. pitch-and-time-scrambled" to isolate voxels that were maximally sensitive to musical pitch structure, and "intact vs. time-scrambled drums" to isolate voxels that were maximally sensitive to temporal structure in music without any potential interference from the presence of melodic information. However, we note that similar results were observed for a variety of other pairwise contrasts that also attempt to isolate scrambling effects in one domain or the other. As before, we also required that voxels responded to intact MIDI music significantly above baseline ($p < 0.01$, uncorrected). The group-level parcels for these two contrasts contained of 1017 voxels (averaged across the four splits of data; SD = 36 voxels) for the pitch-scrambling contrast (**Figure 4C**, left), and 845 voxels (SD = 36 voxels) for the time-scrambling contrast (**Figure 4C**, right). We then measured these voxels' response in independent data, and determined the extent to which voxels chosen to be maximally sensitive to musical structure in one domain (i.e. "pitch" or "timing") also showed an effect of scrambling in the opposite domain.

The Dice similarity coefficient (DSC) (Dice, 1945) was used to quantify the extent of overlap between the two group parcels, as well as between the voxels selected for the two contrasts within individual participants. This metric corresponds to twice the number of overlap voxels divided by the total number of voxels, and thus ranges from 0 to 1. It is expressed as:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

(Eq. 6)

where $|X|$ and $|Y|$ represent the number of voxels selected for each of the two localizer contrasts.

Because we found there to be substantial overlap between the voxels selected for these two fROIs in individual participants (Dice coefficient = 0.42, SD = 0.11), we ran two additional fROI analyses in which we examined only voxels that were selected for one of the contrasts (top 10% most significant within the group parcel) but not the other. This was intended to maximize our chances of finding voxels that were *selectively* sensitive to one type of musical structure. This left an average of 61 voxels that were significant for "pitch-scrambling" but not "time-scrambling" (averaged across the four splits of data and then participants; SD = 14 voxels), and 45 voxels that were significant for "time-scrambling" but not "pitch-scrambling" (SD = 11 voxels). As before, we measured these voxels' response in independent data, and determined the extent to which voxels chosen to be maximally *selective* for musical structure in one domain (i.e. "pitch" but not "timing") still showed an effect of scrambling in the opposite domain.

Statistics

For both the component analyses and fROI analyses, in order to quantify the effect of scrambling on neural responses, we ran 2 x 2 repeated measures ANOVAs on the four MIDI melody conditions with factors "pitch-scrambling" and "time-scrambling". In all cases, we used the Shapiro-Wilk test to verify that the standardized residuals for every combination of factors were normally distributed, and the results of this test were non-significant ($p > 0.05$) unless otherwise noted in the text. When the assumption of normality was not met, the significance of the F-statistic was evaluated using approximate permutation tests. To do this, we randomized the assignment of the data points for each participant across the conditions being tested 10,000 times, recalculated the F-statistic for each permuted sample, and then compared the observed F-statistic to this null distribution to determine significance. Note that there was no need to check for violations of sphericity, because sphericity necessarily holds for repeated-measures factors with only two levels. For both parametric and non-parametric ANOVAs, effect sizes were quantified using partial eta-squared (η_p^2).

To compare responses across two conditions, such as intact vs. scrambled drums, two-tailed paired t-tests were used when data distributions were normal (evaluated using the Shapiro-Wilk test), and effect sizes were calculated using Cohen's *d*. When the assumption of normality was not met, Wilcoxon signed-rank tests were used and effect sizes were calculated as:

$$r = Z/\sqrt{N}$$

(Eq. 7)

where Z is the z-score output of the Wilcoxon signed rank test, and N is the number of pairs (i.e. the number of participants).

For all figures, we used "within-subject" scatter plots and error bars (Loftus and Masson, 1994), which remove between-subject variance by subtracting each participant's mean response across conditions. The global mean across participants is then added to this de-meaned data before plotting and computing the standard error (SEM).

RESULTS

The goal of the current study was to test which properties of music drive music-selective responses in human non-primary auditory cortex. To that end, we used voxel decomposition techniques to isolate these music-selective responses in a set of 15 participants, and then measured the effect of scrambling different types of music structure on the responses.

Because music selectivity is weak when measured using standard voxel-wise fMRI analyses, we first utilized techniques from our prior work (Norman-Haignere et al., 2015) to isolate music-selective responses from other, potentially spatially overlapping brain responses. This method entails decomposing voxel responses into a weighted sum of a small number of response profiles, each of which plausibly reflects a different neural subpopulation in auditory cortex. In our previous study using this method, we found that voxel responses to a wide variety of natural sounds could be approximated by a set of six canonical response profiles, or "components." Each component is defined by its response profile across a large set of natural sounds, as well as a pattern of weights across voxels, specifying the degree to which that component contributes to the response of each voxel. In the original study, four of the six components were found to reflect acoustic properties of the sound set (e.g. frequency, spectrotemporal modulation) and were concentrated in and around primary auditory cortex (PAC), consistent with prior results (Schönwiesner and Zatorre, 2009; Humphries et al., 2010; Da Costa et al., 2011; Herdener et al., 2013; Santoro et al., 2014; Hullett et al., 2016; Norman-Haignere and McDermott, 2018). The two remaining components responded selectively to speech and music (**Figure 2A**), respectively, and were not well accounted for using standard acoustic properties alone.

We used our component localizer data to estimate a set of voxel weights (**Figure 2B**) for each participant that best approximated these previously inferred component response profiles. We then used these weights to infer the component responses to the stimuli in the main experiment (by multiplying the pseudoinverse of the weights with the participant's voxel responses from the main experiment). This process is conceptually analogous to standard functional region-of-interest (fROI) methods, in which a set of voxels are selected based on their profile of activity

across stimulus conditions, and then the response of these voxels is further queried in independent data. The difference is that instead of treating each voxel as homogenous, with binary assignments to an fROI, we allow each voxel to reflect several component responses and use the component weights to determine the voxel's contribution to each component's response.

The music-selective component response estimated in this way passes several sanity checks. As expected, we found that the component responded most strongly to intact music, and minimally to the non-music control stimuli (**Figure 2C**). Further, the response of the music component to drum stimuli was significantly higher than the non-music sounds (Z = 3.35, p = 0.0008, r = 0.87) but significantly lower (Z = 3.35, p = 0.0008, r = 0.87) than the more melodic music stimuli, replicating previous findings (Boebinger et al., 2021).

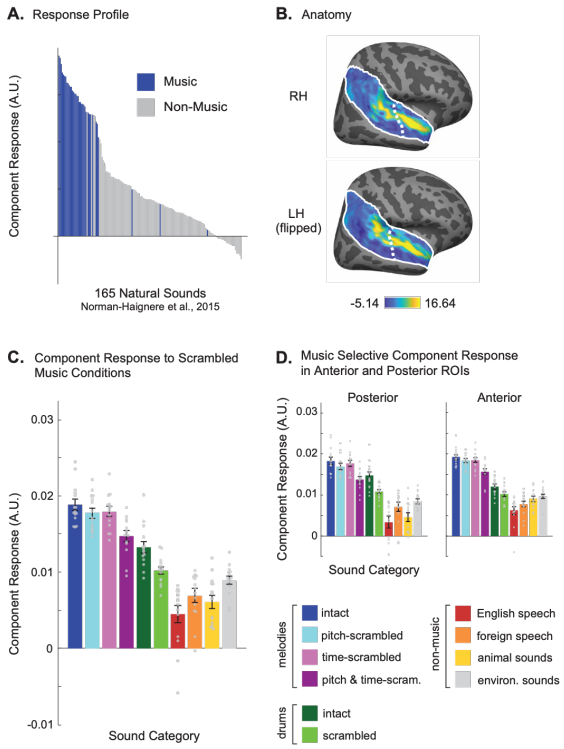

Figure 2. Music component response to MIDI music and non-music control stimuli. **A.** Music component response profile across all 165 natural sounds from Norman-Haignere et al. (2015) study. Sounds categorized as both "instrumental music" and "vocal music" (as determined by raters on Amazon Mechanical Turk) are shaded blue to highlight the degree of music selectivity. **B.** Spatial distribution of music component voxel weights for participants in the current study, averaged across participants in standardized anatomical coordinates (FreeSurfer's FsAverage template). Color scale spans the central 95% of the weight distribution. The white dashed line shows the dividing line between anterior and posterior ROIs. This line was defined by drawing a straight line on the downsampled 2D cortical surface. **C.** Music component response to the stimulus conditions in the main experiment, averaged across participants. Each gray dot corresponds to one participant's component response to a given condition. Scatter plots are within-subject, such that differences between participants are removed by subtracting each participant's mean response across conditions and then adding the global mean across participants. Error bars indicate ±1 within-subject SEM. **D.** Same as **C**, but for the posterior (left) and anterior (right).

Music-selective component is sensitive to both pitch and temporal structure

Our main question was whether manipulating musical structure in pitch and time would affect the response of the music component. To address this question, we extracted the music component response to each of our experimental conditions (**Figure 2C**) and ran a 2 x 2 repeated-measures ANOVA on the component response to the four MIDI melody conditions. We observed significant main effects of scrambling both note pitches ($F(1, 14) = 29.12$, $p = 9.43e\text{-}05$, $\eta_p^2 = 0.68$) and onset times ($F(1, 14) = 14.64$, $p = 0.002$, $\eta_p^2 = 0.51$), as well as a significant interaction, such that the effect of scrambling in one domain was larger if the other domain was scrambled as well ($F(1, 14) = 28.29$, $p < 1.08e\text{-}04$, $\eta_p^2 = 0.67$). A separate Wilcoxon signed-rank test comparing the music component response to synthetic drum stimuli also showed a large effect of time-scrambling ($Z = 3.41$, $p = 0.0007$, $r = 0.88$). These results indicate that the music-selective component is sensitive to both pitch and temporal structure in music, consistent with the idea that these types of information may be jointly processed within auditory cortex.

No evidence for lateralization of sensitivity to pitch vs. temporal structure

Motivated by proposals of hemispheric differences related to music (Zatorre et al., 2002; Albouy et al., 2020), we repeated this entire analysis using component weights and voxel responses from one hemisphere at a time in order to obtain the music component response for voxels in each hemisphere separately. We then compared the component responses from each hemisphere using a 2 x 2 x 2 repeated-measures ANOVA including “hemisphere” as a factor. Contrary to music-related lateralization hypotheses, we found no significant main effect of hemisphere ($F(1, 14) = 0.55$, $p = 0.47$, $\eta_p^2 = 0.04$), nor did we observe any significant interactions involving the “hemisphere” factor (2-way and 3-way interaction p’s > 0.10).

Weak evidence for greater sensitivity to temporal structure in posterior regions

Another proposed organizational principle of auditory cortex is tuning to spectrotemporal modulation rates (Schönwiesner and Zatorre, 2009; Santoro et al., 2014; Norman-Haignere et al., 2015; Hullett et al., 2016; Hamilton et al., 2018), such that regions anterior to Heschl’s gyrus respond preferentially to sounds with fine spectral modulation energy and a strong sense of pitch (Norman-Haignere et al., 2013), and posterior regions respond preferentially to sounds with rapid temporal modulation and to sound onsets (Hamilton et al., 2018). As an initial test of whether these tuning properties might manifest in differential sensitivity to pitch vs. temporal structure in music, we conducted an exploratory analysis in which we divided auditory cortex along the middle of Heschl’s gyrus, roughly perpendicular to STG (see white dashed line in **Figure 2B**), and measured the music component response in anterior and posterior auditory cortex separately (**Figure 2D**). A 2 x 2 x 2 repeated-measures ANOVA found no significant main effect of region ($F(1, 14) = 1.11$, $p = 0.31$, $\eta_p^2 = 0.07$), nor any significant interactions involving the “region” factor (2-way and 3-way interaction p’s > 0.05). However, a separate 2 x 2 repeated-measures ANOVA focused only on the two drum conditions did find a significant interaction, such that the effect of time-scrambling drum stimuli was significantly larger in

posterior compared to anterior auditory cortical regions ($F(1, 14) = 13.60$, $p = 0.006$, $\eta_p^2 = 0.49$; because data was non-normal, the significance of the F-statistic was determined using a non-parametric permutation test, see "*Statistics*" section of Methods for more details). While responses to the drum stimuli were higher overall in posterior regions compared to anterior regions, this difference was not statistically significant ($F(1, 14) = 2.22$, $p = 0.16$, $\eta_p^2 = 0.14$; significance of F-statistic determined using a non-parametric permutation test). Together, these mixed results (a significant effect of "region" for the MIDI drum but not MIDI melody stimuli) are only partially consistent with the notion that posterior auditory cortical regions might play a more central role in representing temporal musical structure than anterior auditory cortical regions, at least in the absence of pitch structure.

Sensitivity to musical structure is strongest in the music-selective component

We also examined the sensitivity of the other five auditory cortical response components (**Figure 3A**) to manipulations in musical structure. Of particular note is the response component that has previously been shown to be sensitive to pitch (Component 4, **Figure 3**) (Norman-Haignere et al., 2015; Boebinger et al., 2021), which responded strongly to the melodic MIDI music as well as speech and animal sounds, but only weakly to MIDI drums (which do not have a strong fundamental frequency in the range of audible pitch). Despite its strong response to music, we did not observe a significant effect of scrambling musical structure in the response of this pitch-sensitive component (repeated-measures ANOVA with factors "pitch-scrambling" and "time-scrambling" produced no significant main effects or interaction; all p's > 0.05, see **Table 2**). Even though the spatial distribution of the pitch-sensitive and music-selective components substantially overlap, their response properties dissociate, highlighting the ability of voxel decomposition to isolate neural subpopulations within voxels that show different response characteristics.

As in previous studies (Norman-Haignere et al., 2015; Boebinger et al., 2021), the speech-selective component (Component 5, **Figure 3B**) responded strongly to both English and foreign speech but only weakly to music. The speech-selective component did show a statistically significant main effect of pitch-scrambling ($F(1,14) = 48.39$, $p = 6.69e\text{-}04$, $\eta_p^2 = 0.78$) and a significant pitch-scrambling x time-scrambling interaction ($F(1, 14) = 20.07$, $p = 0.0005$, $\eta_p^2 = 0.60$). However, a 2 x 2 x 2 repeated-measures ANOVA (with factors "pitch-scrambling," "time-scrambling," and "component") showed that the effect of scrambling was significantly greater in the music-selective component than the speech-selective component (2-way interaction between "component" x "pitch-scrambling": $F(1, 14) = 13.56$, $p = 0.002$, $\eta_p^2 = 0.49$; 3-way interaction: $F(1, 14) = 13.01$, $p = 0.003$, $\eta_p^2 = 0.48$).

Other components that reflect standard acoustic properties of sound (Components 1-4, **Figure 3A** & **B**) also showed occasionally significant (though modest) effects of music scrambling (see **Table 2** for full ANOVA results). One potential explanation is that these scrambling effects are driven by low-level acoustic differences between the intact and scrambled stimulus conditions. Another potential explanation is that the rotation matrix used in our original study to maximize the statistical independence of the components was unable to fully separate them, such that

small amounts of the music-selective component were either added (in the case of the speech-selective component) or subtracted from (in the case of the high-frequency component) the other components' responses. Despite these occasional scrambling effects in other components, a series of 3-way repeated-measures ANOVAs comparing each component to the music-selective component indicated that the effects of scrambling were greatest for the music-selective component (see **Table 3** for full ANOVA results; all 2-way main effects and 3-way interaction were significant: p's < 0.02). We also note that none of the other 5 components both (1) responded to the MIDI music conditions above baseline, and (2) showed a pattern of greater response to intact structure as compared to scrambled structure. Together, these results suggest that sensitivity to high-level melodic and/or rhythmic structure is specific to music-selective neural populations.

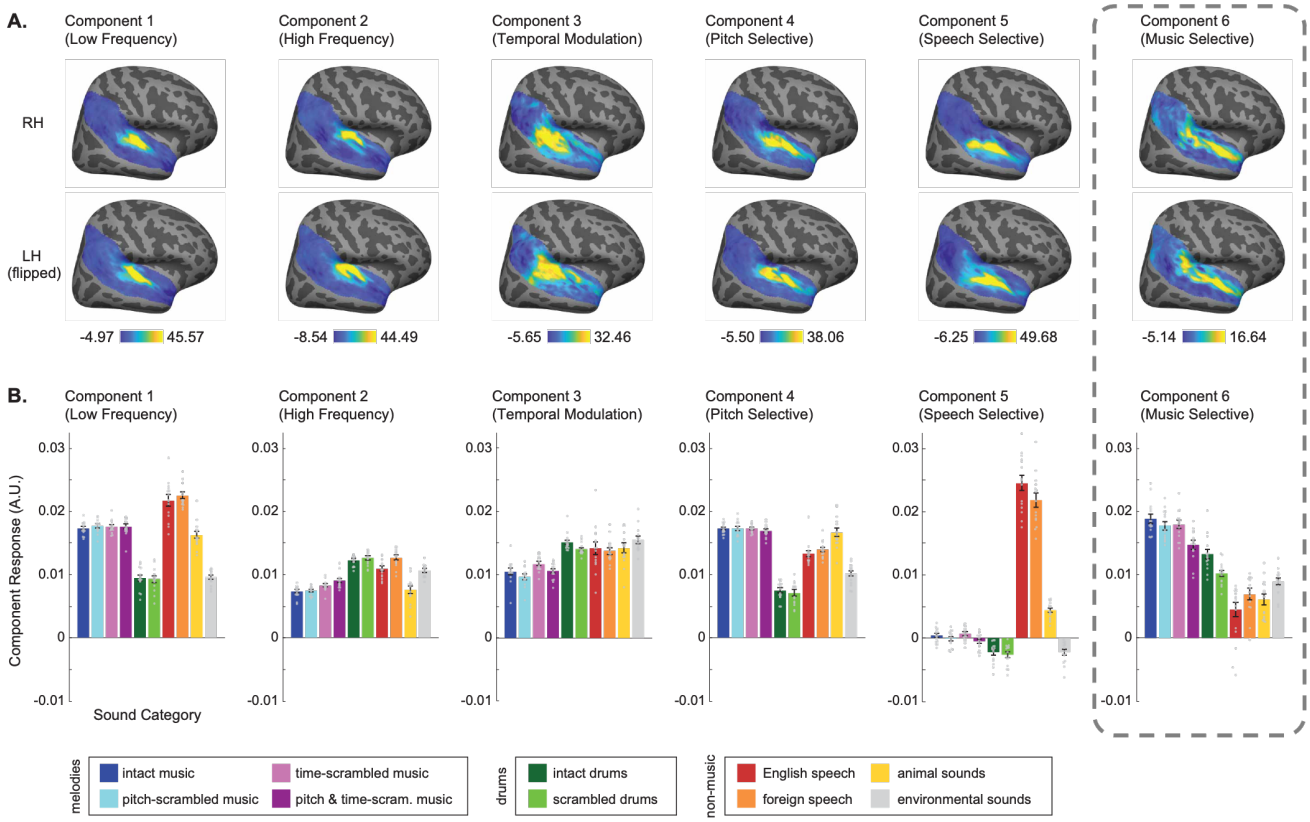

Figure 3. Response of all components to MIDI music and non-music control stimuli**. A.** Spatial distribution of component voxel weights for participants in the current study, inferred using the response components from Norman-Haignere et al. (2015). Voxel weights are averaged across participants in standardized anatomical coordinates. Color scale spans the central 95% of the weight distribution for each component. Music component voxel weights are the same as depicted in **Figure 2B**. **B.** Component responses to the stimulus conditions in the main experiment, averaged across participants. Each gray dot corresponds to one participant's component response to a given condition. Scatter plots are within-subject, such that differences between participants are removed by subtracting each participant's mean response across conditions and then adding the global mean across participants. Error bars indicate ±1 within-subject SEM. Music component response is the same as depicted in **Figure 2C**.

Comp.	Pitch-scrambling main effect			Timing-scrambling main effect			Pitch x timing interaction		
	F-value	p-value	η_p^2	F-value	p-value	η_p^2	F-value	p-value	η_p^2
Low freq.	0.88	0.36	0.06	0.04	0.85	0.003	0.97	0.34	0.06
High freq.	12.32*	0.003	0.47	33.02***	5.06e-05	0.70	7.76*	0.015	0.36
Temp. mod.	21.09***	5.0e-04‡	0.60	16.64**	0.002‡	0.54	1.05	0.33‡	0.007
Pitch	1.19	0.29	0.08	1.77	0.20	0.11	4.19	0.06	0.23
Speech	48.39***	6.69e-06	0.78	0.23	0.63	0.59	20.07***	5.19e-04	0.59

Table 2. ANOVA results for all auditory cortical response components. ‡ = Shapiro-Wilk test indicated that the standardized residuals for one or more combinations of factors was non-normal ($p < 0.05$), so the significance of the F-statistic was evaluated using a non-parametric permutation test randomizing the assignment of the data points across the relevant conditions 10,000 times (see "*Statistics*" section of Methods). * = Significant at $p < 0.05$, ** = significant at $p < 0.01$, *** = significant at $p < 0.001$.

		Low freq.	High freq.	Temp. mod.	Pitch	Speech
Pitch-scrambling main effect	F-value	15.82**	23.13***	39.35***	25.55***	40.91***
	p-value	0.001	0.0003	<0.0001‡	0.0002	1.67e-05
	η_p^2	0.53	0.95	0.74	0.65	0.75
Timing-scrambling main effect	F-value	7.39*	2.75	1.51	18.09***	12.17**
	p-value	0.017	0.12	0.24‡	0.0008	0.004
	η_p^2	0.35	0.16	0.10	0.56	0.47
Component main effect	F-value	0.028	31.84***	13.84**	0.013	359.24***
	p-value	0.87	6.07e-05	0.001‡	0.91	2.23e-11
	η_p^2	0.002	0.69	0.50	0.0009	0.96
Pitch x timing interaction	F-value	16.99**	12.96**	22.30***	21.68***	37.97**
	p-value	0.001	0.003	0.0004‡	0.0004	0.002
	η_p^2	0.55	0.48	0.61	0.61	0.49
Pitch x comp. interaction	F-value	34.65***	29.92***	7.93*	23.07***	13.56**
	p-value	3.96e-05	8.25e-05	0.004‡	0.0003	0.002
	η_p^2	0.71	0.68	0.36	0.62	0.49
Timing x comp. interaction	F-value	26.09***	23.76**	59.05***	9.90**	15.75**
	p-value	0.0002	0.0002	<0.0001‡	0.007	0.001
	η_p^2	0.65	0.63	0.81	0.41	0.53
Pitch x timing x comp. interaction	F-value	20.72***	30.51***	16.95**	24.25***	13.01**
	p-value	0.0005	7.50e-05	<0.0001‡	0.0002	0.003
	η_p^2	0.60	0.69	0.55	0.63	0.48

Table 3. Results for 3-way repeated-measures ANOVAs comparing the music component to each of the other auditory cortical response components. ‡ = Shapiro-Wilk test indicated that the standardized residuals for one or more combinations of factors was non-normal ($p < 0.05$), so the significance of the F-statistic was evaluated using a non-parametric permutation test randomizing the assignment of the data points across the relevant conditions 10,000 times (see "*Statistics*" section of Methods). * = Significant at $p < 0.05$, ** = significant at $p < 0.01$, *** = significant at $p < 0.001$.

Standard functional region-of-interest (fROI) analyses also show no dissociation of sensitivity to pitch vs. temporal structure in music

To look in a complementary way for dissociations between the processing of musical structure in pitch vs. time, we conducted a standard functional region of interest (fROI) analysis that does not rely on response components inferred using voxel decomposition. We first identified the voxels that are most selective for music, and then tested the extent to which these voxels were sensitive to the pitch-scrambling or time-scrambling manipulations. Specifically, we used an *intact MIDI music* > *non-music* contrast and selected the most significant voxels in each participant individually (**Figure 4A**), and then quantified the response of those voxels to each of the stimulus conditions in held-out data from the same participant (see Methods for additional details).

As can be seen in **Figure 4B**, the response of these maximally music-selective voxels was affected by both pitch-scrambling ($F(1, 14) = 49.30$, $p = 6.04e\text{-}06$, $\eta_p^2 = 0.78$) and time-scrambling ($F(1, 14) = 13.46$, $p = 0.0025$, $\eta_p^2 = 0.49$), and showed a significant interaction between pitch-scrambling and time-scrambling ($F(1, 14) = 49.20$, $p = 6.11e\text{-}06$, $\eta_p^2 = 0.78$). While we also observed a significant effect of time-scrambling in the drum stimuli within this fROI ($Z = 3.41$, $p = 6.55e\text{-}04$, $r = 0.85$), the overall response to drums was no higher than the response to the non-music conditions, unlike what we saw in the music-selective component (**Figure 2C**). The weaker music selectivity of this fROI is consistent with our previous finding that music selectivity overlaps within voxels with other components that respond to lower-level acoustic features (e.g. pitch sensitivity) that are likely to differ between MIDI and drum music (Norman-Haignere et al., 2015). This observation underscores the utility of the voxel decomposition method in isolating music selectivity.

Next, we conducted more targeted fROI analyses to determine whether voxels selected to be sensitive to musical structure in one domain (i.e. pitch or time) were also sensitive to structure in the other domain. We used two contrasts to identify voxels that were most sensitive to either the "pitch-scrambling" or the "time-scrambling" manipulation (**Figure 4C**). These two sets of voxels overlapped substantially within individual participants (mean Dice coefficient across subjects = 0.42, SD = 0.11), and at the group level (Dice coefficient = 0.81; **Figure 4D**; see "*Functional region-of-interest analyses"* section of the Methods for details concerning the creation of group parcels). This indicates that the voxels that were most sensitive to pitch structure in music also tended to be sensitive to its temporal structure, which is consistent with the idea that these two types of information are jointly analyzed in auditory cortex. As a stronger test for dissociation, we omitted these overlap voxels, selecting only voxels within individual participants that were

part of one fROI (e.g. significantly melody-sensitive) but not the other (e.g. not significantly rhythm-sensitive), and then ran a 2 x 2 repeated-measures ANOVA on the response of each resulting set of voxels separately (**Figure 4E**). If these sets of voxels were to have different patterns of selectivity, we would expect the same effect to be present in independent data. However, in both cases, we found that the voxels that were selected to be sensitive to one type of structure also showed a significant effect of scrambling in the other domain (all main effects and interactions were significant, p's < 0.05; see **Table 4** for full ANOVA results; results were nearly identical for the fROIs including the overlap voxels that were both significantly melody-sensitive and rhythm-sensitive).

Note that even though these contrasts were intended to isolate specific aspects of musical structure, the voxel responses within the resulting fROIs were not music-selective, as evidenced by a high response to the English and foreign speech conditions (**Figure 4E**). The voxels in these fROIs plausibly include music-selective neural populations as well as other neural populations that respond strongly to speech (e.g. speech-selective neural populations), which cannot be unmixed using standard fROI analyses like these.

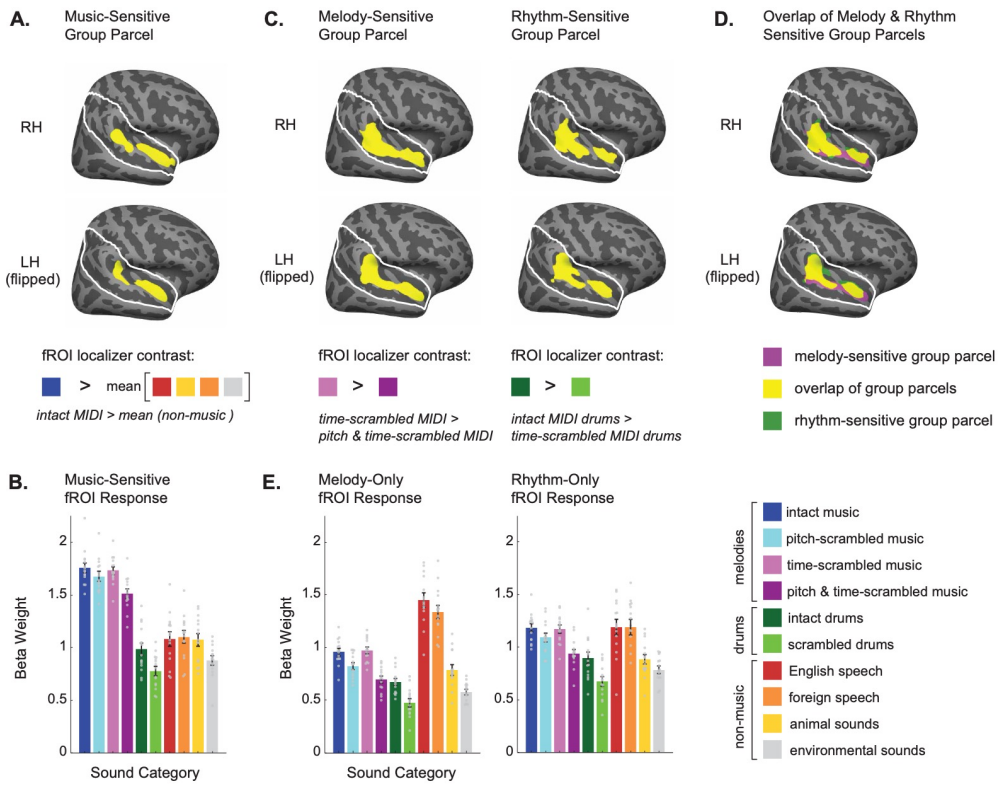

Figure 4. fROI responses to MIDI music and non-music control stimuli. **A.** Group anatomical parcel used to constrain the "Music-sensitive" fROI. To identify the broader anatomical region within which most participants have voxels sensitive to musical structure, we overlapped individual participants' contrast significance maps (p < 0.01, uncorrected) for the "intact-MIDI vs. non-music" contrast, spatially smoothing this overlap with a Gaussian filter (8 mm FWHM), and then thresholding the smoothed overlap map to contain only those voxels with at least 10% overlap across participants. White line indicates the anatomical region from which voxels were selected (same as colored

regions in **Figure 2B** and **Figure 3A**). **B.** Response of selected voxels to the stimulus conditions in the main experiment. For each participant, the 10% most significant voxels for were selected from within the group parcel. The response of these voxels was then measured in independent data. fROI responses are averaged across participants, with each gray dot corresponding to one participant's fROI. Scatter plots are within-subject, such that differences between participants are removed by subtracting each participant's mean response across conditions and then adding the global mean across participants. Error bars indicate ±1 within-subject SEM. **C.** Same as **A**, but for the "Melody-sensitive" (left) and "Rhythm-sensitive" (right) fROIs, using contrasts intended to identify voxels sensitive to pitch structure ("timing-scrambled vs. pitch-and-timing-scrambled," left) or temporal structure ("intact vs. scrambled drums," right). **D.** Overlap of "Melody-sensitive" (purple) and "Rhythm-Sensitive" (green) group parcels. Note that substantial overlap was also present in individual participants. **E.** Same as **B**, but for the voxels in individual participants that were members of one fROI but not the other, e.g. "Melody-sensitive" but not "Rhythm-sensitive" (left), and vice versa (right).

DISCUSSION

Our results show that the music-selective response component of human auditory cortex is sensitive to both the pitch and temporal structure of music, rather than being driven exclusively by structure in one domain or the other. Our multiple attempts to explicitly isolate neural responses that show sensitivity to structure in one domain consistently showed that these regions were sensitive to structure in the other domain as well. These results are consistent with the idea that melodic and rhythmic information are processed jointly within auditory cortex.

Relation to prior work

Within the field of music cognition, there is longstanding interest in the degree to which the processing of pitch and temporal information is integrated. Corpus studies have confirmed that a reliable relationship exists between tonal and metrical hierarchies in music, such that important pitches occur on prominent beats (Prince and Schmuckler, 2014), and that listeners are sensitive to the joint distribution of pitch and temporal information (Prince et al., 2020). Consistent with this idea, some studies have found that listeners treat melody and rhythm as a unified dimension in perception and memory, such that changes to one dimension affect melodic expectancy and goodness-of-fit judgements in the other dimension (Schmuckler and Boltz, 1994; Prince et al., 2009; Prince, 2011), as well as performance on change detection and same/different tasks (Jones et al., 1982; Kidd et al., 1984) and even basic judgements of duration (Crowder and Neath, 1995). However, other studies have argued for the perceptual separability of pitch and temporal information, based on the failure to observe joint effects of pitch and rhythm on judgements of melody completion, pleasantness, or similarity (Palmer and Krumhansl, 1987; Pitt and Monahan, 1987; Makris and Mullet, 2003), or on performance on matching, change detection, or recall tasks (Thompson, 1994; Thompson et al., 2001; Schellenberg et al., 2014). One hypothesis that attempts to reconcile these divergent results is that pitch and temporal information might be initially processed independently and integrated at later processing stages in the auditory hierarchy (Pitt and Monahan, 1987; Peretz and Kolinsky, 1993; Thompson et al., 2001; Peretz and Coltheart, 2003).

The literature on acquired amusia does not fit as naturally with this. Some amusia patients present with deficits in high-level music perception that are seemingly confined to either melody (Grant-Allen, 1878; Peretz, 1990, 1996; Peretz and Kolinsky, 1993; Peretz et al., 1994; Liégeois-Chauvel et al., 1998; Ayotte et al., 2000; Piccirilli et al., 2000; Peretz and Hyde, 2003; Hyde and Peretz, 2004) or rhythm (Mavlov, 1980; Fries and Swihart, 1990; Peretz, 1990; Wilson et al., 2002; Di Pietro et al., 2004; Phillips-Silver et al., 2011). But the heterogeneity of patients with brain damage complicates the interpretation of these results, as does the diversity of experimental tasks used to characterize musical deficits and rule out alternative explanations (see Supplemental Tables 1 and 2 in Stewart, 2006 for example cases and test batteries). For example, in some cases seemingly melody-specific deficits might actually reflect more basic deficits in discriminating pitch differences or direction of pitch change for pairs of tones (Tanaka et al., 1987; Johannes et al., 1998; Johnsrude et al., 2000; Tramo et al., 2002; Wilson et al., 2002; Hattiangadi et al., 2005), as has been found to be the case in congenital amusia (Liu et al., 2010; Peretz et al., 2015). In other cases, it isn't clear that impairments are confined to the perception of music, with some patients also having difficulty interpreting speech prosody or recognizing and discriminating voices (Peretz, 1993; Peretz and Kolinsky, 1993; Peretz et al., 1994; Patel et al., 1998), or presenting with aphasia or more general cognitive impairments (Brust, 1980; Mavlov, 1980; Mazzucchi et al., 1982; Tanaka et al., 1987; Mendez and Geehan, 1988; Tramo et al., 1990, 2002; Eustache et al., 1990; Fries and Swihart, 1990; Hofman et al., 1993; Peretz et al., 1994; Johkura et al., 1998; Piccirilli et al., 2000; Di Pietro et al., 2004; Hattiangadi et al., 2005; Sihvonen et al., 2016). On the other hand, the amusia literature does provide some support for a dissociation between neural representations of lower-level pitch and timing information versus more abstract representations of musical structure, because some patients with acquired amusia have preserved lower-level auditory perceptual abilities, e.g. a preserved ability to discriminate the pitches of pairs or short sequences of tones but impaired processing of tonality (Zatorre, 1985; Peretz, 1993; Peretz et al., 1994; Warrier and Zatorre, 2004) and vice versa (Tramo et al., 1990), or intact discrimination of rhythmic sequences but impaired processing of meter (Liégeois-Chauvel et al., 1998) or vice versa (Di Pietro et al., 2004). Taken together, evidence from neuropsychological patients suggests at least some degree of separation between the representations of pitch and temporal structure in music, but the mapping between specific brain regions and the perceptual consequences of brain damage remain underspecified.

Our results are consistent with other neuroimaging studies in normal listeners that have found largely overlapping regions of auditory cortex to be responsive to pitch and temporal structure in music (Griffiths et al., 1999; Krumhansl and Zatorre, 2003b, 2003a; Alluri et al., 2012; Fedorenko et al., 2012). Most aligned with the current study is that of Fedorenko et al. (2012), which functionally identified brain regions sensitive to musical structure and found a similar magnitude of response in those regions to pitch-scrambled and time-scrambled music. But while the ROIs in that study did indeed respond significantly more strongly to intact than scrambled music and showed no sensitivity to linguistic structure (Fedorenko et al., 2012), that study did not test whether those responses were selective for music above other types of sounds. In the current study, we show similar effects in a component of cortical responses that has been shown to be selective for high-level musical features (Norman-Haignere et al., 2015; Norman-

Haignere and McDermott, 2018; Boebinger et al., 2021). It is also worth noting that the scrambling manipulation from Fedorenko et al. (2012) was relatively coarse compared to the method used in the current study. In Fedorenko et al. (2012), the pitches of individual notes were randomized, which disrupted the consonance of simultaneous notes and completely randomized the melodic contour. The technique used in the current experiment scrambled clusters of notes with the same onset as a single unit, and preserved the global shape of the melodic contour by jittering pitches by up to 2 semitones rather than completely randomizing them. This method enabled us to show that a truly music-selective response component is sensitive to relatively subtle violations of musical structure, while nearby (and sometimes spatially overlapping) response components do not show this sensitivity.

On the other hand, other prior findings about the functional organization of human auditory cortex supported a prediction that we might observe a dissociation between sensitivity to pitch and temporal structure in different subregions of auditory cortex. For example, much of auditory cortex is tuned to particular spectrotemporal modulation rates, with regions anterior to Heschl's gyrus preferring fine spectral modulations and slow temporal modulations, and vice versa for regions posterior to Heschl's gyrus (Schönwiesner and Zatorre, 2009; Santoro et al., 2014; Norman-Haignere et al., 2015; Hullett et al., 2016; Hamilton et al., 2018). Because fine spectral modulations are characteristic of harmonic sounds with a strong sense of pitch, and fast temporal and broad frequency modulations are present in sound onsets, it might seem intuitive to map these selectivities onto high-level musical structure (Patel and Iversen, 2014). Specifically, these findings might suggest that anterior regions would be more sensitive to melodic structure and posterior regions would be more sensitive to rhythmic structure. We found some evidence supporting this hypothesis. While we did not observe a significant difference between anterior and posterior regions in their sensitivity to the pitch and temporal structure in melodic music, we did observe a stronger effect of scrambling drum stimuli in posterior auditory cortex. Thus, posterior regions like planum temporale, which preferentially respond to the rapid temporal modulation and broadband frequency content inherent in sound onsets, might also be sensitive to the temporal patterning of onset events ("rhythm") or to the presence of an isochronous beat ("meter").

Further, some prior studies have proposed an asymmetry between the two hemispheres arising from differences in spectrotemporal resolution, which some argue might have implications for musical processing such that the right hemisphere is specialized for pitch-based aspects of music and the left hemisphere for rhythmic information (Zatorre and Belin, 2001; Albouy et al., 2020). However, we find no evidence for lateralization of music-selective responses in either the current study or in our previous work (Norman-Haignere et al., 2015, 2021; Boebinger et al., 2021), nor have we found lateralization of pitch-selective responses (Norman-Haignere et al., 2013, 2015).

What is music selectivity, then?

In this experiment, we found that the music-selective component inferred using voxel decomposition was sensitive to both pitch and temporal structure in synthetic MIDI music. This

finding is consistent with the results of a previous experiment in which we found this component to respond strongly to drum rhythms with minimal melodic content in addition to real-world music with a strong sense of melody (Boebinger et al., 2021). Further, the sensitivity of the music component to the scrambling manipulations used in the current study suggest that it represents relatively abstract properties of melody and rhythm. This conclusion is consistent with previous experiments that have shown that the music component is driven by a set of features that are relatively specific to music, given that it responds strongly to a wide range of music, including unfamiliar musical genres from non-Western cultures (Boebinger et al., 2021) and weakly to virtually all other natural sounds (Norman-Haignere et al., 2015, 2021; Boebinger et al., 2021), and that the response of the music component does not simply represent the audio or modulation frequencies that are prevalent in music (Norman-Haignere and McDermott, 2018).

While results of the current study add to our understanding of cortical music selectivity, they do not fully specify what aspect of music is driving responses. In addition to the constraints mentioned previously, the relevant features must unfold over relatively short timescales, given that the stimuli used to characterize music selectivity in the component localizer and in previous studies from our group were only 2 seconds long (Norman-Haignere et al., 2015; Boebinger et al., 2021). Possible candidate features include stable and sustained pitch organized into discrete note-like elements or synchronous groups of notes that partly form the basis of harmony. Indeed, these features would largely be preserved in the scrambled conditions from this experiment, which would explain why the scrambling manipulation had only a modest effect on the response of the music component. However, the scrambling effects we observed indicate that the music component is at least somewhat sensitive to how note-like elements are combined, suggesting that it might reflect aspects of musical structure like melodic contour, pitch intervals, or temporally regular beats. Music-selective neural populations in auditory cortex might thus be responsible for extracting temporally local features that are assembled elsewhere into more abstract representations of music, including key, meter, groove, event structure, etc. (Janata et al., 2002; Brett and Grahn, 2007; Lee et al., 2011; Fedorenko et al., 2012; Matthews et al., 2020; Williams et al., 2021). It is also plausible that responses might be further modulated by top-down inputs from brain regions like frontal cortex, perhaps reflecting the important role of expectation when it comes to music perception (Koelsch et al., 2018).

Limitations

The failure to find a dissociation between responses to the pitch and temporal structure of music could simply reflect the limitations of fMRI. Using fMRI we seem to be able to infer only six reliable response components across all of auditory cortex before overfitting to noise in the data, whereas more recent work using methods with higher spatial and temporal resolution has been able to infer a larger number of reliable response components that reflect finer-grained patterns of selectivity across sounds. For example, a recent study using ECoG was able to infer a component selective for song that was distinct from response components reflecting selectivity for music and speech more generally (Norman-Haignere et al., 2021). High temporal resolution

might in fact be required to observe dissociations between processing of pitch and temporal structure in music, as suggested by another recent ECoG study that found differences in the temporal dynamics of expectations associated with note pitches vs. onset-times (Di Liberto et al., 2020), which would not be detectable with fMRI. This potential importance of temporal dynamics is further underscored by the finding that, in animals, neurons in primary auditory cortex have been shown to change their spike timing, but not overall firing rate, in response to rhythmic compared to random sound sequences (Asokan et al., 2021). Thus, it remains possible that there are indeed distinct representations of pitch and temporal musical structure within auditory cortex, but that these are not resolvable with fMRI, either using standard voxel-wise analyses or component methods. Future work with intracranial electrodes in human neurosurgery patients could answer this question.

The lack of dissociation between sensitivity to pitch and temporal structure that we observed in this experiment could also be a consequence of details of our experimental design. Several behavioral studies have shown that the degree to which melodic and rhythmic information are integrated in perception and memory depends on factors like attention and task demands (Monahan and Carterette, 1985; Thompson, 1994; Prince, 2011). For example, it has been proposed that tasks that require listeners to make judgements about local pitch and/or temporal features lead to independent representations of melody and rhythm, whereas integration is observed when listeners attend to stimulus features on a longer time-scale (Jones and Boltz, 1989; Bigand et al., 1999; Tillmann and Lebrun-Guillaud, 2006). Thus, while we didn't observe any dissociation in this experiment, in which participants listened relatively passively while performing an (intentionally simple) intensity discrimination task, it is conceivable that the use of a task that encourages different modes of listening might reveal independent neural representations of pitch and temporal structures.

It is also possible that the scrambling manipulations we employed in this experiment were too subtle to resolve differential sensitivity to one type of musical structure over another. While we did verify that listeners were readily able to detect both pitch- and time-scrambling in our stimulus set (see the behavioral experiments described in the *"Main Experiment: Stimuli"* section of the Methods), the scrambled stimuli are still clearly identifiable as "music." Further, previous studies have shown that only a minimal amount of information is needed for listeners to be able to fully instantiate complex tonal (Dowling, 1978; Cuddy and Badertscher, 1987; Oram et al., 1995; Smith and Schmuckler, 2004) and metrical structure (Povel and Okkerman, 1981; Palmer and Krumhansl, 1990; Brochard et al., 2003; Desain and Honing, 2003). Listeners might therefore be able to draw on their vast amount of experience with Western tonal music in order to form expectations for what the scrambled stimuli "should" sound like. This perceptual "unscrambling" might have reduced the effects that we observed.

Conclusions

We found that the music-selective component of fMRI responses in auditory cortex was sensitive to both pitch structure and temporal structure of music, suggesting that the processing of these two structural dimensions is intertwined. However, much remains to be learned about cortical music selectivity. Future studies could employ the same methods used here to test

specific hypotheses about the importance of particular music features (e.g. tonality), or the timescale of musical structure required to drive responses. Answering these questions will bring us closer to a scientific understanding of the quintessentially human capacity for music.

ACKNOWLEDGEMENTS

This work was supported by NSF grant BCS-1634050 and NIH grant R01DC017970 to J.H.M. and NIH grant DP1HD091947 to N.K. S.V.N.H was supported by a K99/R00 award from the NIH (DC018051). The Athinoula A. Martinos Imaging Center at MIT is supported by the NIH Shared instrumentation grant #S10OD021569.

REFERENCES

Albouy P, Benjamin L, Morillon B, Zatorre RJ (2020) Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. Science 367:16.

Alluri V, Toiviainen P, Jääskeläinen IP, Glerean E, Sams M, Brattico E (2012) Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. Neuroimage 59:3677–3689.

Asokan MM, Williamson RS, Hancock KE, Polley DB (2021) Inverted central auditory hierarchies for encoding local intervals and global temporal patterns. Curr Biol 31:1–9.

Ayotte J, Peretz I, Rousseau I, Bard C, Bojanowski M (2000) Patterns of music agnosia associated with middle cerebral artery infarcts. Brain 123:1926–1938.

Bigand E, Madurell F, Tillmann B, Pineau M (1999) Effect of global structure and temporal organization on chord processing. J Exp Psychol Hum Percept Perform 25:184–197.

Boebinger D, Norman-Haignere S V., McDermott JH, Kanwisher N (2021) Music-selective neural populations arise without musical training. J Neurophysiol 125:2237–2263.

Boltz MG (1999) The processing of melodic and temporal information: Independent or unified dimensions? Int J Phytoremediation 21:67–79.

Brett M, Grahn JA (2007) Rhythm and beat perception in motor areas of the brain. J Cogn Neurosci 19:893–906.

Brochard R, Abecasis D, Potter D, Ragot R, Drake C (2003) The “Ticktock” of our Internal Clock: Direct Brain Evidence of Subjective Accents in Isochronous Sequences. Psychol Sci 14:362–366.

Brust JCM (1980) Music and Language: Musical alexia and agraphia. Brain 103:367–392.

Crowder RG, Neath I (1995) The Influence of Pitch on Time Perception in Short Melodies. Music Percept 12:379–386.

Cuddy LL, Badertscher B (1987) Recovery of the tonal hierarchy : 41.

Da Costa S, Van Der Zwaag W, Marques JP, Frackowiak RSJ, Clarke S, Saenz M (2011) Human Primary Auditory Cortex Follows the Shape of Heschl’s Gyrus. J Neurosci 31:14067–14075.

Dale AM (1999) Optimal experimental design for event-related fMRI. Hum Brain Mapp 8:109–114.

Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis: I. Segmentation and surface reconstruction. Neuroimage 9:179–194.
Dalla Bella S, Peretz I (1999) Music agnosias: Selective impairments of music recognition after brain damage. Int J Phytoremediation 21:209–216.
Desain P, Honing H (2003) The formation of rhythmic categories and metric priming. Perception 32:341–365.
Di Liberto GM, Pelofi C, Bianco R, Patel P, Mehta AD, Herrero JL, de Cheveigné A, Shamma S, Mesgarani N (2020) Cortical encoding of melodic expectations in human temporal cortex. Elife 148:26–74.
Di Pietro M, Laganaro M, Leemann B, Schnider A (2004) Receptive amusia: Temporal auditory processing deficit in a professional musician following a left temporo-parietal lesion. Neuropsychologia 42:868–877.
Dice LR (1945) Measures of the Amount of Ecologic Association Between Species. Ecology 26:297–302.
Dowling WJ (1978) Scale and contour: Two components of a theory of memory for melodies. Psychol Rev 85:341–354.
Ellis D (2014) midi2aiff.scpt. Available at: https://github.com/dpwe/alignmidi/blob/master/midi2aiff.scpt.
Eustache F, Lechevalier B, Viader F, Lambert J (1990) Identification and discrimination disorders in auditory perception: A report on two cases. Neuropsychologia 28:257–270.
Fedorenko E, Hsieh PJ, Nieto-Castañón A, Whitfield-Gabrieli S, Kanwisher N (2010) New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. J Neurophysiol 104:1177–1194.
Fedorenko E, McDermott JH, Norman-Haignere S V., Kanwisher NG (2012) Sensitivity to musical structure in the human brain. J Neurophysiol 108:3289–3300.
Fischer J, Mikhael JG, Tenenbaum JB, Kanwisher N, Hespos SJ, Tsao D (2016) Functional neuroanatomy of intuitive physical inference. Proc Natl Acad Sci.
Fries W, Swihart AA (1990) Disturbance of rhythm sense following right hemisphere damage. Neuropsychologia 28:1317–1323.
Grant-Allen C (1878) Note-Deafness. Mind 3:157–167.
Greve DN, Fischl B (2009) Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48:63–72.
Griffiths TD, Johnsrude I, Dean JL, Green GGR (1999) A common neural substrate for the analysis of pitch and duration pattern in segmented sound? Neuroreport 10:3825–3830.
Hamilton LS, Edwards E, Chang EF (2018) A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. Curr Biol 28:1860–1871.
Hattiangadi N, Pillion JP, Slomine B, Christensen J, Trovato MK, Speedie LJ (2005) Characteristics of auditory agnosia in a child with severe traumatic brain injury: A case report. Brain Lang 92:12–25.
Herdener M, Esposito F, Scheffler K, Schneider P, Logothetis NK, Uludag K, Kayser C (2013) Spatial representations of temporal and spectral sound cues in human auditory cortex. Cortex 49:2822–2833.
Hofman S, Klein C, Arlazoroff A (1993) Common hemisphericity of language and music in a musician a case report. J Commun Disord 26:73–82.

Hullett PW, Hamilton LS, Mesgarani N, Schreiner CE, Chang EF (2016) Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. J Neurosci 36:2014–2026.
Humphries C, Liebenthal E, Binder JR (2010) Tonotopic organization of human auditory cortex. Neuroimage 50:1202–1211.
Hyde KL, Peretz I (2004) Brains that are out of tune but in time. Psychol Sci 15:356–360.
Jackendoff R, Lerdahl F (2006) The capacity for music: What is it, and what's special about it? Cognition 100:33–72.
Janata P, Birk JL, Van Horn JD, Leman M, Tillmann B, Bharucha JJ (2002) The cortical topography of tonal structures underlying western music. Science 298:2167–2170.
Jenkinson M, Smith S (2001) A global optimisation method for robust affine registration of brain images. Med Image Anal 5:143–156.
Johannes S, Jöbges ME, Dengler R, Münte TF (1998) Cortical auditory disorders: A case of non-verbal disturbances assessed with event-related brain potentials. Behav Neurol 11:55–73.
Johkura K, Matsumoto S, Hasegawa O, Kuroiwa Y (1998) Defective auditory recognition after small hemorrhage in the inferior colliculi. J Neurol Sci 161:91–96.
Johnsrude IS, Penhune VB, Zatorre RJ (2000) Functional specificity in the right human auditory cortex for perceiving pitch direction. Brain 123:155–163.
Jones MR, Boltz M (1989) Dynamic Attending and Responses to Time. Psychol Rev 96:459–491.
Jones MR, Boltz M, Kidd G (1982) Controlled attending as a function of melodic and temporal context. Percept Psychophys 32:211–218.
Jones MR, Summerell L, Marshburn E (1987) Recognizing Melodies: A Dynamic Interpretation. Q J Exp Psychol Sect A 39:89–121.
Julian JB, Fedorenko E, Webster J, Kanwisher N (2012) An algorithmic method for functionally defining regions of interest in the ventral visual pathway. Neuroimage 60:2357–2364.
Kay KN, Rokem A, Winawer J, Dougherty RF, Wandell BA (2013) GLMdenoise: A fast, automated technique for denoising task-based fMRI data. Front Neurosci 7:1–15.
Kidd G, Boltz M, Jones MR (1984) Some effects of rhythmic context on melody recognition. Am J Psychol 97:153–173.
Koelsch S, Vuust P, Friston K (2018) Predictive Processes and the Peculiar Case of Music. Trends Cogn Sci 23:63–77.
Krumhansl CL, Zatorre RJ (2003a) Tonal processing in music: The role of right frontal cortex. Meet Cogn Neurosci Soc New York City.
Krumhansl CL, Zatorre RJ (2003b) Rhythmic and tonal violations in music: An fMRI study. Meet Organ Hum Brain Mapping, New York City.
LaCroix AN, Diaz AF, Rogalsky C (2015) The relationship between the neural computations for speech and music perception is context-dependent: an activation likelihood estimate study. Front Psychol 6:1–19.
Leaver AM, Rauschecker JP (2010) Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J Neurosci 30:7604–7612.
Lee YS, Janata P, Frost C, Hanke M, Granger R (2011) Investigation of melodic contour processing in the brain using multivariate pattern-based fMRI. Neuroimage 57:293–300.

Liégeois-Chauvel C, Peretz I, Babaï M, Laguitton V, Chauvel P (1998) Contribution of different cortical areas in the temporal lobes to music processing. Brain 121:1853–1867.
Liu F, Patel AD, Fourcin A, Stewart L (2010) Intonation processing in congenital amusia: Discrimination, identification and imitation. Brain 133:1682–1693.
Loftus R, Masson J (1994) Using confidence intervals in within-subject designs. Psychon Bull Rev 1:476–490.
Makris I, Mullet É (2003) Judging the Pleasantness of Contour-Rhythm-Pitch-Timbre Musical Combinations. Am J Psychol 116:581–611.
Matthews TE, Witek MAG, Lund T, Vuust P, Penhune VB (2020) The sensation of groove engages motor and reward networks. Neuroimage 214:116768.
Mavlov L (1980) Amusia Due to Rhythm Agnosia in a Musician with Left Hemisphere Damage: A Non-Auditory Supramodal Defect. Cortex1 16:331–338.
Mazzucchi A, Marchini C, Budai R, Parma M (1982) A case of receptive amusia with prominent timbre perception defect. J Neurol Neurosurg Psychiatry 45:644–647.
McPherson MJ, Dolan SE, Durango A, Ossandon T, Valdés J, Undurraga EA, Jacoby N, Godoy RA, McDermott JH (2020) Perceptual fusion of musical notes by native Amazonians suggests universal representations of musical intervals. Nat Commun 11:1–14.
Mendez MF, Geehan GR (1988) Cortical auditory disorders: Clinical and psychoacoustic features. J Neurol Neurosurg Psychiatry 51:1–9.
Monahan CB, Carterette EC (1985) Pitch and Duration as Determinants of Musical Space. Music Percept 3:1–32.
Nieto-Castañón A, Fedorenko EG (2012) Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. Neuroimage 63:1646–1669.
Norman-Haignere S V., Feather J, Boebinger D, Brunner P, Ritaccio A, McDermott JH, Schalk G, Kanwisher N (2021) Intracranial recordings from human auditory cortex reveal a neural population selective for musical song. bioRxiv.
Norman-Haignere S V., Kanwisher NG, McDermott JH (2013) Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. J Neurosci 33:19451–19469.
Norman-Haignere S V., Kanwisher NG, McDermott JH (2015) Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. Neuron 88:1281–1296.
Norman-Haignere S V., McDermott JH (2018) Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. PLoS Biol 16.
Oram N, Cuddy LL, Oram N (1995) Responsiveness of Western adults to pitch-distributional information in melodic sequences. Psychol Res 57:103–118.
Palmer C, Krumhansl CL (1987) Independent Temporal and Pitch Structures in Determination of Musical Phrases. J Exp Psychol Hum Percept Perform 13:116–126.
Palmer C, Krumhansl CL (1990) Mental Representations for Musical Meter. J Exp Psychol Hum Percept Perform 16:728–741.
Patel AD, Iversen JR (2014) The evolutionary neuroscience of musical beat perception: The Action Simulation for Auditory Prediction (ASAP) hypothesis. Front Syst Neurosci 8:1–14.
Patel AD, Peretz I, Tramo M, Labreque R (1998) Processing prosodic and musical patterns: A

neuropsychological investigation. Brain Lang 61:123–144.
Peretz I (1990) Processing of local and global musical information by unilateral brain-damaged patients. Brain 113:1185–1205.
Peretz I (1993) Auditory atonalia for melodies. Cogn Neuropsychol 10:21–56.
Peretz I (1996) Can We Lose Memory for Music? A Case of Music Agnosia in a Nonmusician. J Cogn Neurosci 8:481–496.
Peretz I, Ayotte J, Zatorre RJ, Mehler J, Ahad P, Penhune VB, Jutras B (2002) Congenital Amusia: A disorder of fine-grained pitch discrimination. Neuron 33:185–191.
Peretz I, Coltheart M (2003) Modularity of music processing. Nat Neurosci 6:688–691.
Peretz I, Hyde KL (2003) What is specific to music processing? Insights from congenital amusia. Trends Cogn Sci 7:362–367.
Peretz I, Kolinsky R (1993) Boundaries of Separability between Melody and Rhythm in Music Discrimination: A Neuropsychological Perspective. Q J Exp Psychol Sect A 46:301–325.
Peretz I, Kolinsky R, Tramo M, Labrecque R, Hublet C, Demeurisse G, Belleville S (1994) Functional dissociations following bilateral lesions of auditory cortex. Brain 117:1283–1301.
Peretz I, Vuvan D, Laine Lagrois M-É, Armony JL (2015) Neural overlap in processing music and speech. Phil Trans R Soc B 370.
Phillips-Silver J, Toiviainen P, Gosselin N, Piché O, Nozaradan S, Palmer C, Peretz I (2011) Born to dance but beat deaf: A new form of congenital amusia. Neuropsychologia 49:961–969.
Piccirilli M, Sciarma T, Luzzi S (2000) Modularity of music: Evidence from a case of pure amusia. J Neurol Neurosurg Psychiatry 69:541–545.
Pitt MA, Monahan CB (1987) The perceived similarity of auditory polyrhythms. Percept Psychophys 41:534–546.
Povel DJ, Okkerman H (1981) Accents in equitone sequences. Percept Psychophys 30:565–572.
Prince JB (2011) The integration of stimulus dimensions in the perception of music. Q J Exp Psychol 64:2125–2152.
Prince JB, Schmuckler MA (2014) The tonal-metric hierarchy: A corpus analysis. Music Percept 31:254–270.
Prince JB, Tan SEJ, Schmuckler MA (2020) Listeners perceive complex pitch-temporal structure in melodies. Mem Cogn 48:526–540.
Prince JB, Thompson WF, Schmuckler MA (2009) Pitch and Time, Tonality and Meter: How Do Musical Dimensions Combine? J Exp Psychol Hum Percept Perform 35:1598–1617.
Rogalsky C, Rong F, Saberi K, Hickok G (2011) Functional Anatomy of Language and Music Perception: Temporal and Structural Factors Investigated Using Functional Magnetic Resonance Imaging. J Neurosci 31:3843–3852.
Santoro R, Moerel M, De Martino F, Goebel R, Ugurbil K, Yacoub E, Formisano E (2014) Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. PLoS Comput Biol 10.
Schellenberg EG, Stalinski SM, Marks BM (2014) Memory for surface features of unfamiliar melodies: Independent effects of changes in pitch and tempo. Psychol Res 78:84–95.
Schmuckler MA, Boltz MG (1994) Harmonic and rhythmic influences on musical expectancy. Percept Psychophys 56:313–325.

Schönwiesner M, Zatorre RJ (2009) Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc Natl Acad Sci 106:14611-.
Sihvonen AJ, Ripollés P, Leo V, Rodríguez-Fornells A, Soinila S, Särkämö T (2016) Neural basis of acquired amusia and its recovery after stroke. J Neurosci 36:8872–8881.
Smith NA, Schmuckler MA (2004) The Perception of Tonal Structure Through the Differentiation and Organization of Pitches. J Exp Psychol Hum Percept Perform 30:268–286.
Stewart L, Von Kriegstein K, Warren JD, Griffiths TD (2006) Music and the brain: Disorders of musical listening. Brain 129:2533–2553.
Tanaka Y, Yamadori A, Mori E (1987) Pure word deafness following bilateral lesions: A psychophysical analysis. Brain 110:381–403.
Thompson WF (1994) Sensitivity to combinations of musical parameters: Pitch with duration, and pitch pattern with durational pattern. Percept Psychophys 56:363–374.
Thompson WF, Hall MD, Pressing JL (2001) Illusory Conjunctions of Pitch and Duration in Unfamiliar Tone Sequences. J Exp Psychol Hum Percept Perform 27:128–140.
Tierney A, Dick F, Deutsch D, Sereno M (2013) Speech versus Song: Multiple Pitch-Sensitive Areas Revealed by a Naturally Occurring Musical Illusion. Cereb Cortex 23:249–254.
Tillmann B, Lebrun-Guillaud G (2006) Influence of tonal and temporal expectations on chord processing and on completion judgments of chord sequences. Psychol Res 70:345–358.
Traer J, Norman-Haignere S V., McDermott JH (2021) Causal inference in environmental sound recognition. Cognition 214:104627.
Tramo MJ, Bharucha JJ, Musiek FE (1990) Music Perception and Cognition Following Bilateral Lesions of Auditory Cortex. J Cogn Neurosci 2.
Tramo MJ, Shah GD, Braida LD (2002) Functional role of auditory cortex in frequency processing and pitch perception. J Neurophysiol 87:122–139.
Warrier CM, Zatorre RJ (2004) Right temporal cortex is critical for utilization of melodic contextual cues in a pitch constancy task. Brain 127:1616–1625.
Williams JA, Margulis EH, Nastase SA, Chen J, Hasson U, Norman KA, Baldassano C (2021) High-order areas and auditory cortex both represent the high-level event structure of music. bioRxiv:2021.01.26.428291.
Wilson SJ, Pressing JL, Wales RJ (2002) Modelling rhythmic function in a musician post-stroke. Neuropsychologia 40:1494–1505.
Woods KJP, McDermott JH (2018) Schema learning for the cocktail party problem. Proc Natl Acad Sci U S A 115:E3313–E3322.
Woods KJP, Siegel MH, Traer J, McDermott JH (2017) Headphone screening to facilitate web-based auditory experiments. Attention, Perception, Psychophys 79:2064–2072.
Zatorre RJ (1985) Discrimination and recognition of tonal melodies after unilateral cerebral excisions. Neuropsychologia 23:31–41.
Zatorre RJ, Belin P (2001) Spectral and temporal processing in human auditory cortex. Cereb Cortex 11:946–953.
Zatorre RJ, Belin P, Penhune VB (2002) Structure and function of auditory cortex: music and speech. Trends Cogn Sci 6.