NeuroImage 166 (2018) 60–70

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Encoding of natural timbre dimensions in human auditory cortex

Emily J. Allen [a,*], Michelle Moerel [b,c], Agustín Lage-Castellanos [b,d], Federico De Martino [b,e], Elia Formisano [b,c], Andrew J. Oxenham [a]

[a] Department of Psychology, University of Minnesota, Minneapolis, MN 55455, United States
[b] Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, 6200 MD, Maastricht, The Netherlands
[c] Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, 6200 MD, Maastricht, The Netherlands
[d] Department of NeuroInformatics, Cuban Center for Neuroscience, Street 190 e/25 and 27 Cubanacán Playa Havana, CP 11600, Cuba
[e] Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Minneapolis, MN 55455, United States

ARTICLE INFO

Keywords:
Auditory cortex
Encoding models
Music
Perception
Timbre

ABSTRACT

Timbre, or sound quality, is a crucial but poorly understood dimension of auditory perception that is important in describing speech, music, and environmental sounds. The present study investigates the cortical representation of different timbral dimensions. Encoding models have typically incorporated the physical characteristics of sounds as features when attempting to understand their neural representation with functional MRI. Here we test an encoding model that is based on five subjectively derived dimensions of timbre to predict cortical responses to natural orchestral sounds. Results show that this timbre model can outperform other models based on spectral characteristics, and can perform as well as a complex joint spectrotemporal modulation model. In cortical regions at the medial border of Heschl's gyrus, bilaterally, and regions at its posterior adjacency in the right hemisphere, the timbre model outperforms even the complex joint spectrotemporal modulation model. These findings suggest that the responses of cortical neuronal populations in auditory cortex may reflect the encoding of perceptual timbre dimensions.

Introduction

Timbre, the perceptual quality or color of a sound, is defined as everything by which a listener can distinguish between two sounds with the same loudness, pitch, spatial location, and duration (ANSI, 2013). For instance, it is differences in timbre that allow us to distinguish a violin from a guitar, or one vowel sound from another. Among the typical adjectives that fall under the category of timbre are “brightness”, “clarity”, “harshness”, “fullness”, and “noisiness” (Stepanek, 2006). Efforts have been made to identify and quantify the most salient aspects of timbre through the use of multidimensional scaling (MDS) techniques (e.g., Grey, 1977; Elliott et al., 2013). MDS utilizes subjective measures to determine how perceptually similar a selection of sounds are to one another, thereby creating a geometric representation that derives the subjective distances between a diverse set of stimuli using as few dimensions as possible (Grey, 1977). After collecting similarity ratings for musical instrument sounds with unique timbres, Grey (1977) used MDS to identify three dimensions that best represented the distribution of timbres. The first dimension was related to the spectral energy distribution of the sounds (ranging from a low to high spectral centroid, corresponding to timbral descriptors ranging from dull to bright), and the other two related to temporal patterns, such as whether the onset was rapid (like a struck piano note or a plucked guitar string) or slow (as is characteristic of many woodwind instruments) and the synchronicity of higher harmonic transients.

Grey's influential study contained only sixteen instrumental sounds from three instrument families, placing some limits on the generalizability of the outcomes, and used sounds that may not have all had exactly the same fundamental frequency (F0), which itself may have affected some aspects of timbre judgments (e.g., Moore and Glasberg, 1990; Warrier and Zatorre, 2002; Allen and Oxenham, 2014). Elliott et al. (2013) extended Grey's approach by using 42 natural orchestral instruments from five instrument families, all with the same F0 (311 Hz, the E♭ above middle C). After collecting similarity and semantic ratings, they performed multiple analyses, including MDS. They consistently found five dimensions to be both necessary and sufficient for describing the timbre space of these orchestral sounds.

The aim of the current study was to determine whether similar dimensions can be identified in the cortical representations of timbral differences. Although the literature on the neural representations of

* Corresponding author.
E-mail address: prac0010@umn.edu (E.J. Allen).

https://doi.org/10.1016/j.neuroimage.2017.10.050
Received 31 March 2017; Received in revised form 19 October 2017; Accepted 24 October 2017
Available online 4 November 2017

timbre is limited, there is some evidence to suggest it is processed in both primary and secondary auditory cortical regions including superior temporal sulcus (STS), posterior Heschl's gyrus (HG), and planum temporale (PT), bilaterally, with possible hemispheric asymmetries (Casey et al., 2012; Halpern et al., 2004; Menon et al., 2002; Staeren et al., 2009; Warren et al., 2005). However, previous studies have not attempted to differentiate the neural representations of different timbral dimensions, and have not explored the possibility that a subjectively based model of timbre could predict patterns of cortical activation in response to sound. In the present study, we use fMRI encoding (Kay et al., 2008; Moerel et al., 2012; Santoro et al., 2014) to determine whether neural populations in the cortex can represent the timbre dimensions identified by Elliott et al. (2013), and compare this model's performance with that of models based on the spectral and temporal characteristics of the sounds.

Materials and methods

Ethics statement

The experimental procedures were approved by the Institutional Review Board (IRB) for human subject research at the University of Minnesota. Written informed consent was obtained from each participant before starting the measurements.

Participants

Ten right-handed subjects (mean age of 28.6 years, standard deviation [STD] = 8.6 years; five females, five males) participated in this study. All subjects had normal hearing, defined as audiometric pure-tone thresholds of 20 dB hearing level (HL) or better, at octave frequencies between 250 Hz and 8 kHz, and were recruited from the University of Minnesota community. Musical experience of subjects ranged from zero to 18 years, with eight of the 10 subjects having at least 10 years of musical experience.

Stimuli and procedure

The stimulus set consisted of 42 professionally recorded natural Western orchestral instrument sounds, taken from the study of Elliott et al. (2013). The sounds were originally obtained from the McGill University Master Samples collection (Opolko and Wapnick, 2006) and were manipulated to all have the same F0 of 311 Hz (E♭), and a subjective duration of 1 s, as described in Elliott et al. (2013). Spectrograms for a subset of these sounds are shown in Fig. 1. Instrument families included strings, flutes, brass, single reeds, and double reeds. When the rms of the stimuli was normalized, the perceptual loudness of the sounds at the level of 75 dB SPL varied noticeably. In order to equalize the perceived loudness of the stimuli, we processed them using a loudness model (Chen et al., 2011; Moore, 2014), and scaled the sounds to produce roughly equal predicted loudness for each sound. This resulted in perceptually equal loudness for 41 of the 42 sounds. One of the sounds, a muted C trumpet, required manual adjustment to subjectively match the perceptual loudness of the other sounds, presumably because certain aspects of the sound (e.g., sharp attack and broad spectrum) were not adequately captured by the loudness model. The adjusted level was selected by four raters (inter-rater differences were no more than 2 dB).

After the loudness adjustments, the average level of the sounds was 74 dB SPL and the range was 62–81 dB SPL (STD = 3.2 dB). Sounds were presented via MRI-compatible Sensimetrics (Malden, MA) S14 earphones with custom filters.

Magnetic resonance imaging

Images were acquired in a 3T MR scanner (Siemens Prisma) at the

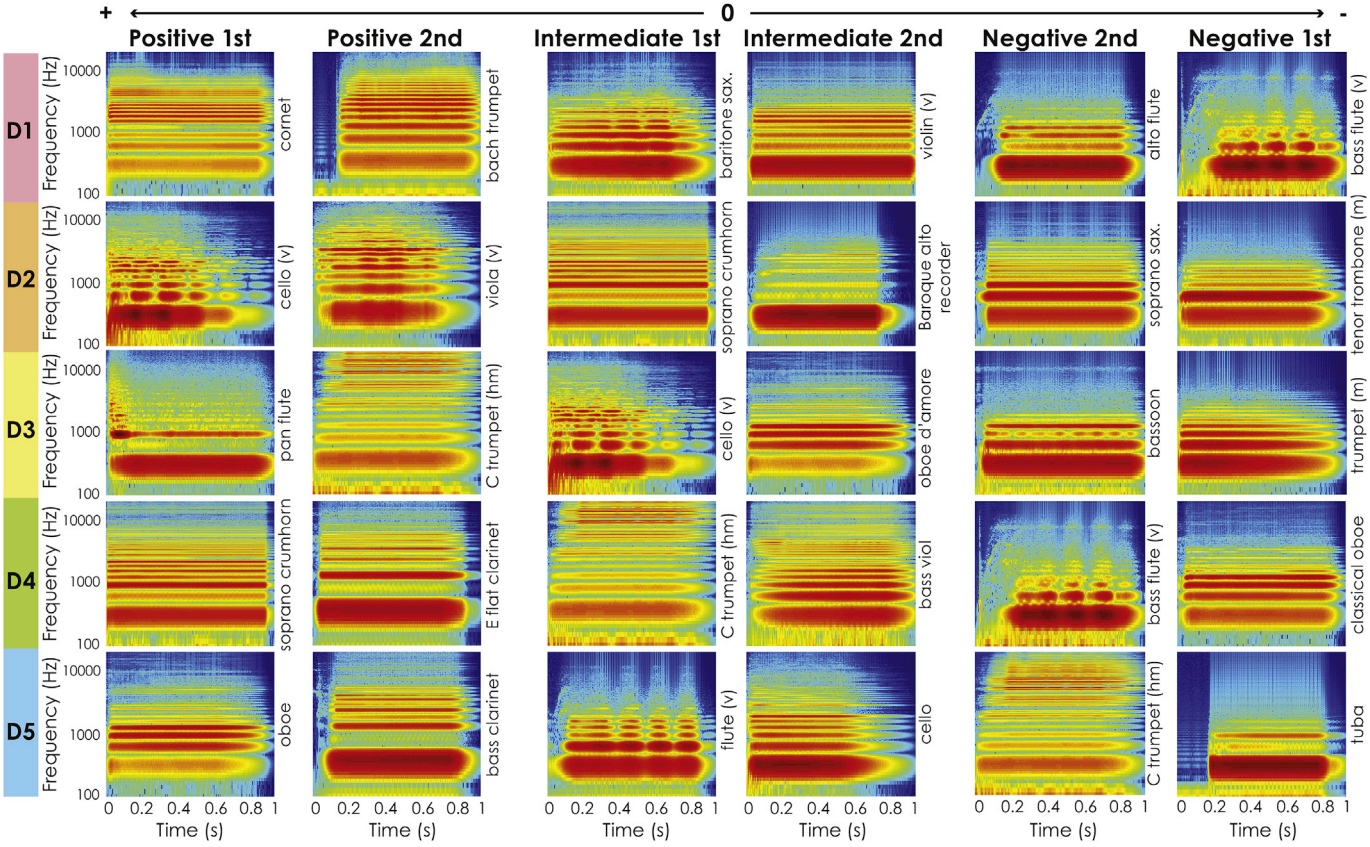

Fig. 1. Spectrograms of the sounds with (columns from left to right) the two most positive, two intermediate, and the two most negative values on each of the five timbre dimensions (rows). Abbreviations: v = vibrato, m = muted, h = harmonic.

Center for Magnetic Resonance Research (CMRR, University of Minnesota) using a 32-channel head coil. For each subject, we collected anatomical images and a functional dataset. The MPRAGE T1-weighted anatomical image parameters were: repetition time (TR) = 2600 ms; echo time (TE) = 3.02 ms; matrix size = 256 × 256; 1 mm isotropic voxels. The acquisition parameters for the functional scans were: TR = 2400 ms; time of acquisition (TA) = 1000 ms; silent gap = TR – TA = 1400 ms; TE = 30 ms; multiband factor = 4; number of slices = 44; matrix size = 672 × 672; 2 mm isotropic voxels. Slices were angled to align with the Sylvian Fissure, and covered the majority of the brain. However, for most subjects the top of the parietal and frontal lobes were excluded, along with the bottom of the occipital lobe.

The functional dataset followed an event-related design, where the sounds were presented in the silent gaps between acquisitions. Six functional runs were collected per subject. In each run, a unique subset of seven of the 42 sounds was repeated four times in pseudo-random order. The division of sounds into separate sets of seven was important for maintaining independence between training and testing datasets in the fMRI encoding analysis (see below). The stimuli within each sound set were manually selected to include a variety of instruments across multiple instrument families. These sound sets remained consistent across subjects, but the presentation order of the stimuli within each set was randomized, and the order of the sets throughout the scanning session was counterbalanced across subjects in a Latin-square design. The presentation times of the sound trials were pseudo-randomly jittered with an interstimulus interval of 2, 3, 4, or 5 TRs. Three silent trials (with no stimuli present) and three catch trials were also included in each run. For the catch trials, intended to keep subjects alert, they were instructed to perform a one-back task in which they pressed a button any time a successive repeat of the same sound was presented. This one-back task never occurred for the same sound more than once in a given run. For the one-back task repeats, the maximum jitter was set to 4 TRs (9.6 s). The one-back task catch trials were excluded from analysis. With the 28 test sounds (four repetitions of seven sounds from the collection) and 3 catch-trial sounds, a total of 31 sounds were presented per run, along with 3 silent trials. Including about 10 s of silence preceding each run and about 5 s following each run, the total duration of one run was approximately 5 min.

The data were preprocessed in BrainVoyager QX (Brain Innovation, Maastricht, The Netherlands). Preprocessing included slice scan time correction (using cubic spline), 3D motion correction (using trilinear/sinc interpolation) aligned to the first volume of the first run, and a high-pass filter (GLM-Fourier) cutoff of 3 cycles per run. Distortion correction was performed using the Correction based on Opposite Phase Encoding (COPE) plugin in BrainVoyager QX, which estimated distortions based on volumes from a posterior-anterior (PA) phase-encoding (PE) direction and volumes from an AP PE direction (Fritz et al., 2014), and applied corrections to the functional data. Functional slices were coregistered to the anatomical data, and then normalized to Talairach space (Talairach and Tournoux, 1988). Automatic segmentation with manual corrections of the grey matter (GM) - white matter (WM) boundary was performed using the anatomical data. Using this boundary, each hemisphere for each subject was then inflated and brought to Cortex Based Aligned (CBA) space (Goebel et al., 2006). CBA-averaged group-level GM-WM meshes were also generated in BrainVoyager QX.

Sound representation by the encoding models

We used fMRI encoding to test several hypotheses for how the brain represents the timbre of natural orchestral instruments. Under the fMRI encoding approach, each hypothesis is defined as an encoding model. We can distinguish between hypotheses by comparing the accuracy with which each of the trained models is able to predict the fMRI response patterns to novel testing sounds. We tested the performance of four encoding models, described below.

First, the subjective *timbre* model represents the hypothesis that responses to the sounds are well described by the five dimensions of timbre identified by Elliott et al. (2013) (see Fig. 1). The first dimension, D1, was semantically described as 'hard, sharp, high-frequency energy balance'. The second dimension, D2, was described as 'varying level, dynamic, vibrato, ringing release'. D3 was characterized as 'noisy, small instrument, unpleasant'. Sounds scoring high on D4 were described as 'compact, steady pitch, pure'. Finally, D5 had no significant correlates among semantic descriptor pairs. Fig. 2A shows the sounds' representation in the space of the *timbre* model. The values of each sound on each of the five dimensions were taken from Elliott et al. (2013). As they were obtained using MDS, the five timbral dimensions were not correlated (Fig. 2B).

Second, the *joint spectrotemporal modulation* (*STM*) model represents the hypothesis that cortical sound processing is well represented by the frequency-specific spectrotemporal modulation tuning of neuronal populations. Sounds are expressed by their frequency-specific spectrotemporal modulation content, obtained as the output of a two-stage biologically inspired model of auditory processing (Chi et al., 2005; Santoro et al., 2014; NSL Tools package, available at http://www.isr.umn.edu/Labs/NSL/Software.htm). This model is similar to the *timbre* model in that it takes into account both spectral and temporal properties of sound, but relies solely on the physical description of sound (transformed via simulated auditory processing), and not on any human subjective judgments. The first stage of this model mimics 'early' auditory processing, and consists of 128 overlapping bandpass filters equally spaced along a logarithmic frequency axis (180–7040 Hz; range of 5.3 octaves). The output of this 'early' stage is a spectrogram, which serves as input to the second 'cortical' stage of the model. This stage uses a set of modulation filters (temporal modulation center frequencies, ω) and spectral modulation center frequencies (cycles/octave, Ω) to extract the spectrotemporal modulation content from the spectrograms. The modulation filters are applied at each time-frequency bin, and the absolute value of the complex-valued model output is then averaged over time. The full STM model contained ω = 30 features, and Ω = 15 features. We divided the frequency axis into 128 bins with equal bandwidth in octaves, and averaged the modulation energy within each frequency bin, resulting in 57,600 features (128×30×15). The sounds' frequency-specific spectrotemporal modulation characteristics as represented by this full model are shown in Fig. 2D–F. This full model was then reduced to 36 features in order to fit it to the fMRI data. The 36 features were: ω = [3, 9, 27] Hz × Ω = [0.5, 1, 2] cycles/octave, with the frequency axis divided into 4 bins with equal bandwidth in octaves. The spectral and temporal modulation filters had Q_{3dB} values of 1.2 and 1.8, respectively. The 36-feature limit was chosen on account of having 42 unique sounds in our stimulus set and wanting to ensure that the number of features in the model was less than the number of unique sounds in our stimulus set. Correlations between the model's 36 features are shown in Fig. 2C.

Third, the *cochlear filter mean* model represents the hypothesis that responses to the sounds are well described by the spectral content of the sounds and the frequency tuning exhibited in the cochlea. This model therefore postulates that the cortical responses reflect primarily the long-term spectral profile of sounds, as filtered by the cochlea, without regard to their temporal properties. The representation of the sounds in the space of this model was obtained based on the output of the first stage of the model underlying the *STM* model. The resulting "cochleograms" were averaged over time, and the frequency axis was divided equally into 36 logarithmic frequency bins (resulting in 36 model features).

Finally, the *spectral centroid* model represents the hypothesis that cortical coding of timbre is dominated by the spectral centroid of a sound, corresponding to the perception of "brightness" or "sharpness" (e.g., von Bismarck, 1974), as represented by Grey's (1977) first dimension, and reflected by cortical tuning to the sounds' spectral centroids. This is essentially a simplified version of the cochlear filter mean model, in that it postulates that the spectral centroid of the sound dominates the representation over other spectral features. The spectral center of gravity c, for each sound was identified by taking the sum of the frequencies f_i,

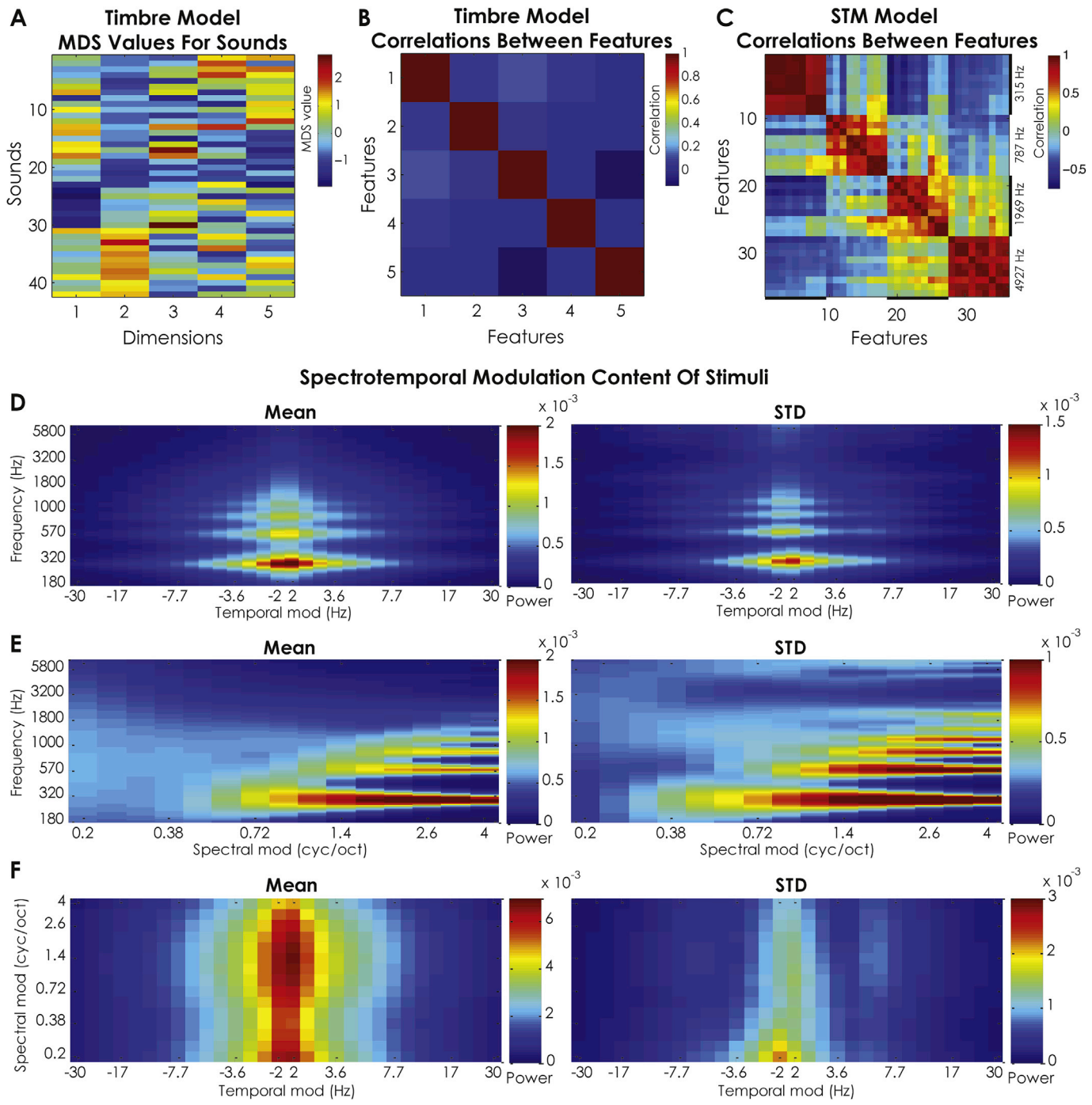

Fig. 2. Sound representation by the *timbre* and *STM* models and frequency-specific spectrotemporal modulation content of the sounds. (A) MDS values for all 42 sounds across the five dimensions (i.e., features) of the *timbre* model, taken from Elliott et al. (2013). (B) Correlation between each of the five *timbre* model features. (C) Correlation between each of the 36 *STM* model features reflecting a high correlation between spectrotemporal modulation features within the same frequency bin. Frequency bins are labeled on the right y-axis. (D) The distribution of temporal modulations across frequency, (E) the distribution of spectral modulation across frequency, and (F) spectral modulations as a function of temporal modulations. The mean and standard deviation (STD) across sounds are shown in the left and right column, respectively.

weighted by their normalized amplitudes a_i:

$$c = \frac{\Sigma(f_i a_i)}{\Sigma a_i}$$

The sounds' representation in the model space was then obtained by creating a [1 x f] vector of zeros for each sound (where f represents the center frequencies of the frequency bins of the *cochlear filter mean* model), and assigning the frequency bin that contained that sound's spectral centroid with a value of one. Frequency bins that did not contain the centroid for any of the 42 sounds were removed. A total of 17 frequency bins remained, resulting in 17 features for this model.

Model training and testing

Model training and testing was done using MATLAB (Mathworks, Natick, MA). We performed the analysis independently for the training and testing runs, which contained completely distinct sets of sounds. That is, model training and testing were performed with 6-fold cross-validation. For each cross-validation, 5 runs (i.e., 35 sounds) served for

model training and one run (i.e., 7 sounds) was left out for model testing.

The fMRI responses to the 42 natural orchestral instrument stimuli were estimated as follows. For each cross-validation, the training data were used to compute noise regressors using the GLMdenoise technique (Kay et al., 2013; GLMdenoise available at: http://kendrickkay.net/GLMdenoise/), and to estimate the hemodynamic response function (HRF) of each voxel across all sounds. This HRF was fixed, and was used in a regression analysis that included the regressors as estimated by GLMdenoise, to estimate the amplitude of the voxel's response (i.e., the beta weight) to each of the training and testing sounds. Next, we identified the voxels that responded significantly to the sounds ($T > 3.5$, $p < 0.001$, uncorrected). For these voxels, regularized linear regression (ridge regression; see Santoro et al., 2014; for details) was used to compute the relationship between the measured fMRI responses and the stimulus features of each model. This relationship (i.e., the trained model) represented how much each feature contributed to a given voxel's response, referred to as the voxel's population response function.

The trained model was evaluated by its ability to predict the fMRI responses to the set of testing sounds that were not used for model training. First, to gain insight into overall model performance (across all regions with a significant response to the sounds) we computed a sound identification *prediction accuracy score*. Activity patterns for each of the test sounds were used to predict the sound identity based on its correlations with the predicted patterns of activity for each of the seven test sounds. These correlations were then sorted and assigned a rank score between one and seven (seven being the lowest rank). In the case of perfect performance, the correlation between the predicted and actual patterns would always be ranked higher for comparisons within the same sound than across different sounds, so the correlation rank, r_i, would always be 1. In the case of chance performance, the expected correlation rank would be in the middle, i.e., 4. Prediction accuracy P_i was then computed for each sound i using the following formula:

$$P_i = 1 - \left(\frac{r_i - 1}{N_{test} - 1}\right)$$

where r_i is the rank across the $N_{test} = 7$ sounds in the test set. The overall prediction accuracy was then computed as the mean of P across all sounds (i.e., averaging across the 6 cross-validation folds), yielding a value between zero and one (perfect prediction score = 1; chance = 0.5). This method for calculating prediction accuracy, while less common than forced-choice accuracy measures that look exclusively at stimuli that are accurately classified (i.e., those that ranked first), has the advantage of taking into account the whole distribution of ranks (beyond those ranked first) to assess the model performance (see e.g., Kay et al., 2008; Moerel et al., 2012; Santoro et al., 2014).

Second, in order to gain insight into the variations in model performance throughout brain areas, we evaluated model accuracy per voxel. For each voxel, we computed the correlation between predicted and measured responses to the testing sounds. Resulting correlations were Fisher's z transformed, and averaged across cross-validations to obtain a map of prediction accuracy per subject for each encoding model.

Group map generation and analysis

Group maps of model prediction accuracy were computed by smoothing single subject prediction accuracy maps, with local averaging up to a distance of four vertices (repeat value = 4) that were then brought into CBA space. For each vertex that was included in at least eight individual subject maps, a one sample *t*-test was performed to test if the observed prediction accuracy (i.e., the correlation between predicted and observed responses to testing sounds) was significantly greater than 0. Following the correction for multiple comparisons using False Discovery Rate (FDR), resulting maps were thresholded at q (FDR) < 0.05.

In order to compare the prediction accuracy of two encoding models, single subject prediction accuracy maps were smoothed (repeat value = 4) and brought into CBA space. For each vertex that was included in at least eight individual maps, a paired samples *t*-test was performed to test if there was a significant difference between the prediction accuracies of the two encoding models. If more than eight subjects were available for a given vertex, paired t-tests were run on a random selection of eight subjects out of all available subjects (this step was taken to ensure equal degrees of freedom and equal number of possible permutations across vertices, see below). To correct for multiple comparisons we used a cluster size thresholding method based on nonparametric permutations. That is, for each vertex we applied the paired *t*-test to all possible permutations of the eight subjects across the two models ($2^8 = 256$ permutations), resulting in 256 permuted maps. We then generated a null distribution of cluster size, considering a single-voxel threshold of $t > 1.8$. Cluster sizes that occurred less frequently than in 5% in the null distribution were considered significant.

Finally, we created group maps for each dimension of the trained *timbre* model. This was an exploratory analysis, with the aim of gaining insight into the cortical representation of the timbre dimensions. For each timbre dimension, we obtained the single subject map as the voxels' weights under the trained *timbre* model and smoothed the maps with a Gaussian kernel of 2 mm full-width at half-maximum (FWHM). We converted the individual subject maps to binary maps, by setting the voxel to −1 or 1 if the weight was smaller or greater than zero, respectively. Next, the individual subject binary maps were brought to CBA space. Probability maps were created by assigning each voxel with the proportion of subjects that showed the same sign in their weight map (chance = 0.5; perfect congruency among subjects = 1; map threshold set to 0.75).

Results

We observed significant responses to the sounds throughout the superior temporal cortex bilaterally (see Fig. 3). The temporal auditory responsive regions included Heschl's gyrus (HG), and adjacent regions on Heschl's sulcus (HS), planum polare (PP), planum temporale (PT), superior temporal gyrus (STG), and superior temporal sulcus (STS). Beyond the auditory cortices, we observed responses to the sounds in the inferior frontal gyrus, the inferior frontal sulcus, the postcentral gyrus, and the intraparietal sulcus.

Prediction accuracies for the four encoding models are shown in Fig. 4. All models except for the *spectral centroid* model performed significantly above chance (0.5) in a one-tailed *t*-test (mean [SE]; *timbre*: 63% [0.02], $t_9 = 5.97$, $P = 0.0001$, $d = 1.89$; *STM*: 60% [0.01], $t_9 = 6.72$, $P < 0.0001$, $d = 2.12$; *cochlear filter mean*: 56% [0.02], $t_9 = 3.39$, $P = 0.004$, $d = 1.07$). The *timbre* model performed significantly better than the *cochlear filter mean* model ($t_9 = 2.93$, $P = 0.02$, $d = 0.93$), and the *spectral centroid* model ($t_9 = 3.89$, $P = 0.004$, $d = 1.21$). The *STM* model also performed significantly better than the *spectral centroid* model ($t_9 = 3.70$, $P = 0.005$, $d = 1.13$). There was no significant difference between the *timbre* model and the *STM* model ($t_9 = 1.26$, $P = 0.24$, $d = 0.40$), nor between the *cochlear filter mean* model and the *spectral centroid* model ($T_{(9)} = 0.41$, $P = 0.69$, $d = 0.49$).

To test whether the *STM* model's prediction accuracy might improve with the inclusion of more features, we also ran a version of the model that contained 576 features (36 frequency bins X 4 spectral modulations [0.5 1 2 4] X 4 temporal modulations [1 3 9 27]. The average prediction accuracy [SE] in this case was: 59% [0.02], which was not significantly different from the 36-feature version ($t_9 = 0.48$, $P = 0.64$, $d = 0.15$), nor did it outperform the *timbre* model ($t_9 = 1.59$, $P = 0.15$, $d = 0.50$).

Cortical variation in encoding model prediction accuracy

Fig. 5A shows variations in model performance throughout the cortex. These maps indicate how well the measured responses from individual voxels to sounds were represented by the different models. Given that the *spectral centroid* model did not perform significantly above chance, we excluded it from further analysis. Although all models

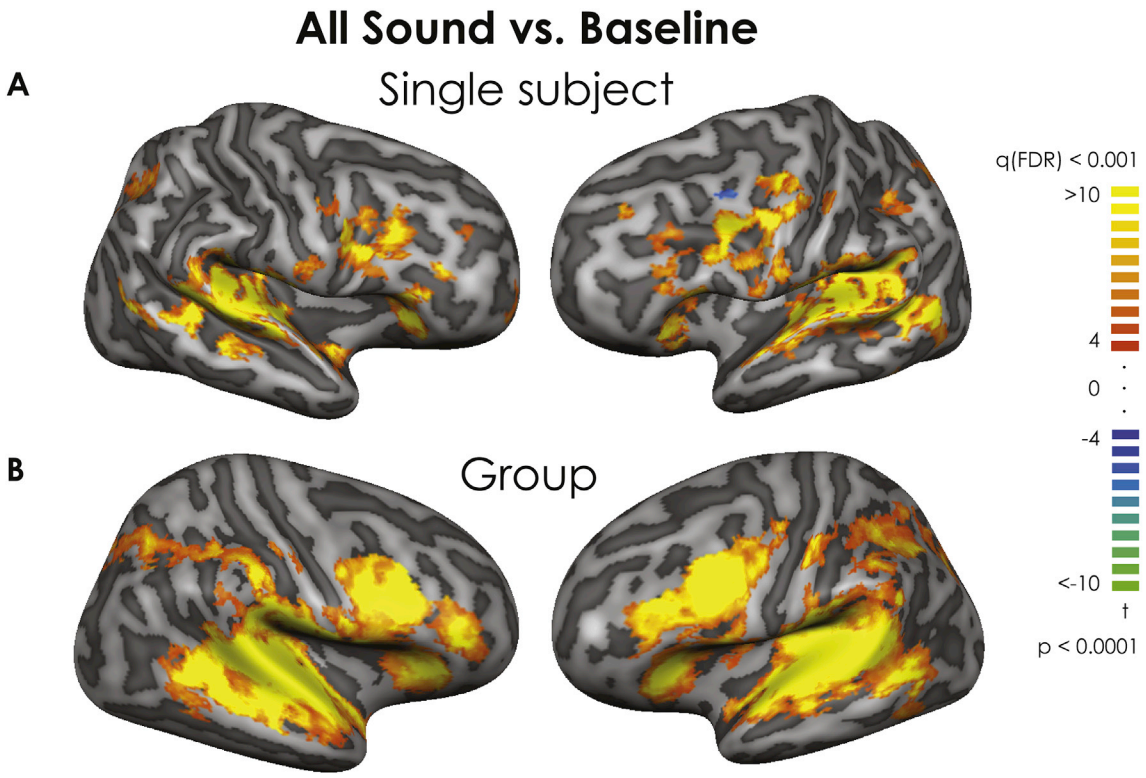

Fig. 3. Brain maps showing average activation across all runs and across all sounds compared to baseline. (A) fMRI response of a single subject to sound stimuli. (B) Group-level fixed-effects GLM maps. Both the single subject and group maps are thresholded at $P < 10^{-4}$ (corresponding to q (FDR) < 0.001), cluster thresholded (cluster size = 25), with nearest-neighbor interpolation.

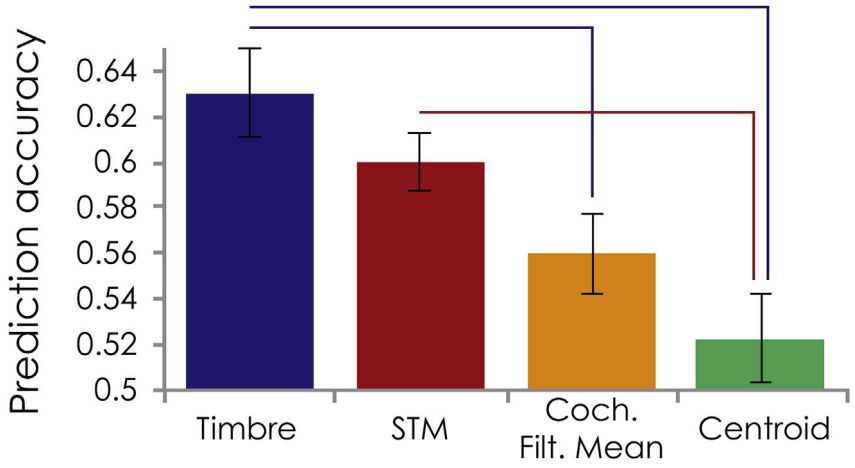

Fig. 4. Mean prediction accuracy across the encoding models. Average model performance across ten subjects for the *timbre*, *STM*, *cochlear filter mean*, and *spectral centroid* models. Error bars represent ± 1 standard error of the mean. Blue lines indicate which models performed significantly worse than the *timbre* model, and red lines indicate which models performed significantly worse than the *STM* model. No other significant differences were found across models.

displayed the highest prediction accuracies around the superior temporal plane (STP) and STG, significantly above-chance accuracy was also observed in frontal regions. Note that differences in the performance of a single model across the brain could result from location-specific differences in noise level (for a review, see Schoppe et al., 2016), and therefore the differences within each panel of Fig. 5A should be interpreted with caution.

Contrast maps

In order to compare the models in terms of the significant regional differences in their performance, we contrasted each model with the *timbre* model (see Fig. 5B). Warmer colors indicate regions in which the *timbre* model has significantly better performance compared to the other models, and cooler colors indicate regions where the other models have significantly better performance than the *timbre* model. Overall, the maps show more warm colors than cool colors, reflecting the overall higher performance of the *timbre* model (i.e., higher sound identification score). The timbre model outperformed all other models in representing processing in right hemispheric regions posterior to HG (covering HS and anterior PT). A comparison of the two best-performing models, the *STM* and *timbre* models, revealed considerable overlap, but also some regional differences. Specifically, the *timbre* model's representation is superior to that of the *STM* model in regions at the medial end of HG bilaterally, and

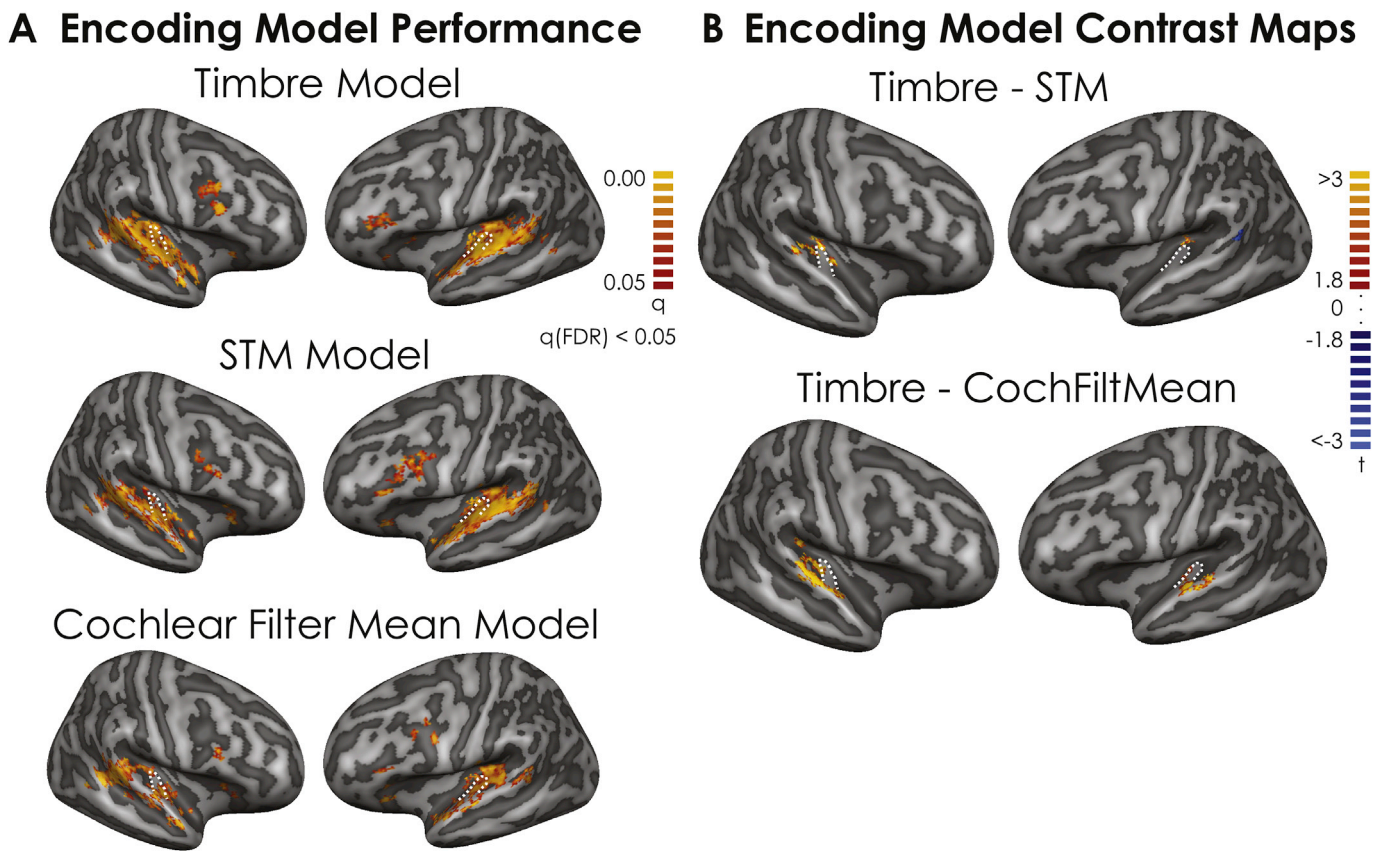

Fig. 5. Group-level model performance. (A) The maps show the cortical regions with a significant (q [FDR] < 0.05) correlation between measured and predicted responses to sounds. From top to bottom, performance of the *timbre*, *STM*, *cochlear filter mean*, and *spectral centroid* model are shown. (B) Group-level differences between models. Positive values (warmer colors) indicate voxels for which the *timbre* model performed significantly better, and negative values (cooler colors) indicate voxels for which the *STM* or *cochlear filter mean* (in the top and bottom panel, respectively) performed significantly better. White dotted lines indicate HG.

at the posterior and anterior adjacency of HG (i.e., HS and first transverse temporal sulcus (FTS), respectively) in the right hemisphere. These areas may reflect either primary or belt regions of auditory cortex (Moerel et al., 2014). The *timbre* model also outperforms the *STM* model in a small region on the STG of the right hemisphere, likely reflecting a belt region of auditory cortex. Conversely, the *STM* model outperforms the *timbre* model in a small region at the posterior end of the STG in the left hemisphere, potentially corresponding to the parabelt region of the auditory cortex (Moerel et al., 2014). Furthermore, compared to the *cochlear filter mean* model, the timbre model performs better in regions along the HG and STG bilaterally, and HS in the right hemisphere. The superior performance seen in lateral HG may correspond to a difference in core auditory regions, while the differences observed in HS of the right hemisphere and the STG bilaterally may correspond to belt and parabelt regions, respectively (Moerel et al., 2014).

Analysis of the timbre dimensions

According to Elliott et al. (2013), around 90% of the perceptual variance in the acoustic stimuli is explained by these five dimensions, and the dimensions are ordered by the amount of variance explained, with D1-D3 explaining the most variance. In order to explore a possible correspondence between this perceptual variance and the neural variance, we tested each dimension of the timbre model separately. The mean prediction accuracy results were: D1: 56%, D2: 60%, D3: 58%, D4: 49%, D5: 52%. In a one-tailed *t*-test, the first three dimensions were significantly above chance ($t_9 = 4.00$, $P = 0.003$, $d = 1.26$; $t_9 = 5.28$, $P = 0.001$, $d = 1.67$; and $t_9 = 4.21$, $P = 0.002$, $d = 1.33$, respectively), suggesting the first three dimensions best predict responses to novel test sounds.

We explored the overlap in the sound representations captured by the *timbre* and *STM* models by using canonical correlation analysis (CCA) (Hotelling, 1936) and linear regression. CCA was used to identify two new sets of features that share the largest amount of information (i.e., the maximum correlation), and linear regression was used to compute the transformation that best describes the features of one model in terms of the features of the other. We describe each approach and report the results below.

CCA and linear regression procedures

CCA was performed in a four-fold cross-validation loop (where a random 75% of the sounds and their representation in the models' space were used for training, and the remaining 25% for testing on an independent data set), repeated 1000 times, to evaluate the canonical correlation using an independent data set. Overfitting of the *STM* model (36 features, 42 sounds) was prevented by using the first 14 principal components (PCs) of the model. These 14 PCs explained 99.8% of the variance in the training data and 98.2% of the variance in the test data. The PC decomposition was performed on the training data and the test data was projected on the PC space. Since the *timbre* model contains only five features, dimensionality reduction on the *timbre* model was not needed. For each cross-validation, the CCA was run on the training data. Next, we computed the proportion of variance in the original *STM* model that could be explained by the canonical covariates of the *timbre* model, and likewise, the proportion of variance in the original *timbre* model that could be explained by the canonical covariates of the *STM* model. For each cross-validation, this was computed by projecting the test data of each model to the canonical covariate space obtained on the training data. On the test data sets, a linear regression between the full set of canonical covariates of one model to the set of original features of the other model was performed. Performing this analysis on test data

independent from the (training) data used to compute the canonical covariates avoids overfitting.

The linear regression between the two models was also performed in a four-fold cross validation loop, repeated 1000 times, and the average values of the explained variance on the test data were reported. Each feature of one model was described as a linear function of all of the features in the other model. The total variance in one model that could be explained by the other was computed as the sum of the explained variances of each feature. When the *STM* model was used as the independent dataset, overfitting was prevented by means of principal components regularization. For consistency with CCA analysis, the linear regression was performed on the subspace spanned by the first 14PCs of the *STM* model. When the *timbre* model was used as the independent variable, no regularization was required and ordinary least squares (OLS) regression was used.

CCA and linear regression results

For the CCA we found, on average, across cross-validations and 1000 repetitions, that the canonical covariates of the *timbre* model explained 34.4% of the variance of the original *STM* model, while the canonical covariates of the *STM* model explained 41.6% of the variance of the *timbre* model. For the linear regression we found, on average, across cross-validations and 1000 repetitions, that a linear combination of the features of the *timbre* model explained 37.1% of the variance of the original *STM* model, while a linear combination of the features of the *STM* model explained 38.2% of the variance of the *timbre* model. The CCA results are in overall accordance with the linear regression results and suggest that while there is a clear overlap between the two models, offering the possibility of (partially) understanding the *timbre* model in terms of basic acoustic features, there remains a substantial amount of variance in the *timbre* model that cannot be explained by the *STM* model and vice versa.

Linking the timbre dimensions to acoustic features

To further explore the acoustic basis of each of the timbre dimensions, we display 3D correlation heat maps between the *STM* model features and each of the five *timbre* model dimensions (Fig. 6A). Additionally, to explore the neurobiological correlate of each of the five timbre dimensions and quantify the consistency across subjects, we conduct an exploratory analysis of the trained *timbre* model, displaying those voxels for which the sign of the voxel's weight in the trained timbre model is consistent across the majority of subjects (Fig. 6B).

The first timbre dimension, D1, is semantically associated with “hard, sharp, high-frequency energy balance” (Elliott et al., 2013), and correlates most strongly with a combination of high frequencies and slow temporal modulations (Fig. 6A). The positive weights on medial HG suggest that these regions respond more strongly to sounds that score high on D1. In contrast, negative weights are distributed along STG, indicating that these cortical locations respond more strongly to sounds that score low on D1 (Fig. 6B). This may reflect the tonotopic organization of the auditory cortex, with a high frequency preference at the medial border of HG, and a low frequency preference along the STG (Langers et al., 2007; Moerel et al., 2012), suggesting this dimension, at least in part, reflects the frequency content of sounds.

D2 is semantically associated with “varying level, dynamic, vibrato, and ringing release”, and is positively correlated with fast temporal modulations, especially in combination with intermediate frequency features (Fig. 6A). These characteristics seem appropriate for the semantic descriptor “ringing release”. In contrast, negative correlations with low temporal modulations are seen at low-to mid-range frequencies and low spectral modulations. D2 weights were consistently positive across a large number of voxels on the supratemporal plane (STP), indicating that these regions respond more strongly to faster temporal modulations. This is in accordance with previous studies that showed a strong bilateral activation of the auditory cortices for sounds with fast temporal modulations (e.g., Zatorre and Belin, 2001; Joanisse and DeSouza, 2014).

D3, which is semantically associated with “noisy, small instrument, and unpleasant”, correlates positively with high frequency features of the *STM* model especially when combined with fast temporal modulations (possibly corresponding to greater spectral irregularity and roughness), and negatively to low frequency features (Fig. 6A). In contrast, the strongest negative correlations were found for slow temporal modulations at low-to mid-range frequencies. This suggests that high frequency sounds with fast modulations may be perceived as more noisy and unpleasant. Like D2 weights, D3 weights were consistently positive across the STP. This is in accordance with previous work, which found unpleasant sounds to be associated with increased bilateral activation throughout auditory cortex (Plichta et al., 2011).

D4 corresponds to “compact, steady pitch, pure”, and correlates positively with the lowest *STM* frequency features, and negatively with mid-range frequency features. Positive D4 weights appear on primary auditory cortical regions centered on HG, suggesting that these regions respond more strongly to more compact and pure sounds. In contrast, negative weights are situated along the STG, which may respond more strongly to broader, more complex sounds. This organization is consistent with hierarchical auditory processing, with simple tones being processed in early auditory cortical areas and more complex sounds undergoing greater processing in secondary or tertiary auditory regions (Patterson et al., 2002; Tian and Rauschecker, 2004).

D5 is difficult to interpret, as the previous work by Elliott et al. (2013) did not reveal a semantic association with this dimension. D5 has strong positive correlations with features that combine mid-range frequencies, slow temporal modulations (~3 Hz), and middle spectral modulations (~1 cycle/octave; Fig. 6A). Furthermore, the anterolateral portion of HG displays positive D5 weights, bilaterally. This may point toward a lower-level dimension in the processing hierarchy, potentially associated with pitch strength (Penagos et al., 2004).

Discussion

In this study, we used fMRI encoding to compare a timbre model derived from listeners' ratings of the sounds with acoustic models based on physical sound characteristics. We observed that the *timbre* model was able to predict a significant portion of the variance in the sound-evoked cortical activation. Furthermore, it performed significantly better than the other models tested, with the exception of a complex joint *spectrotemporal modulation* model. This finding, along with the observation that the two models shared a large part of the variation in the stimulus domain and the inferior performance of the uniquely spectral encoding models, supports the idea that joint spectrotemporal features are critical for capturing timbre perception (Patil et al., 2012).

However, we observed that the *timbre* model outperformed the joint *STM* model in a subset of the auditory cortical locations. Specifically, the *timbre* model performed significantly better in regions medial and posterior to HG, particularly in the right hemisphere. This suggests that while the *timbre* model only contains five features, it may be capturing some semantic or perceptual tuning properties of the auditory cortex that extend beyond those captured by the *spectrotemporal* model. Specifically, the differences observed in terms of the amount of shared variance between the *timbre* and *STM* models identified via CCA and linear regression may be a result of the *timbre* model capturing some nonlinear combination of physical features not represented in the *STM* model. This may be a distinguishing component of higher-level semantic processing (Kay and Yeatman, 2017). In light of this, it would be tempting to combine these two models in hopes of achieving better model performance. However, concatenation of these models is suboptimal as the *timbre* model is made of features that are orthogonal to each other and the *STM* model has many collinear features. As a result, the regularization to be applied to each model separately differs substantially and concatenation would result in over-penalizing the *timbre* model. Therefore, an area that warrants future research is the development of methods to

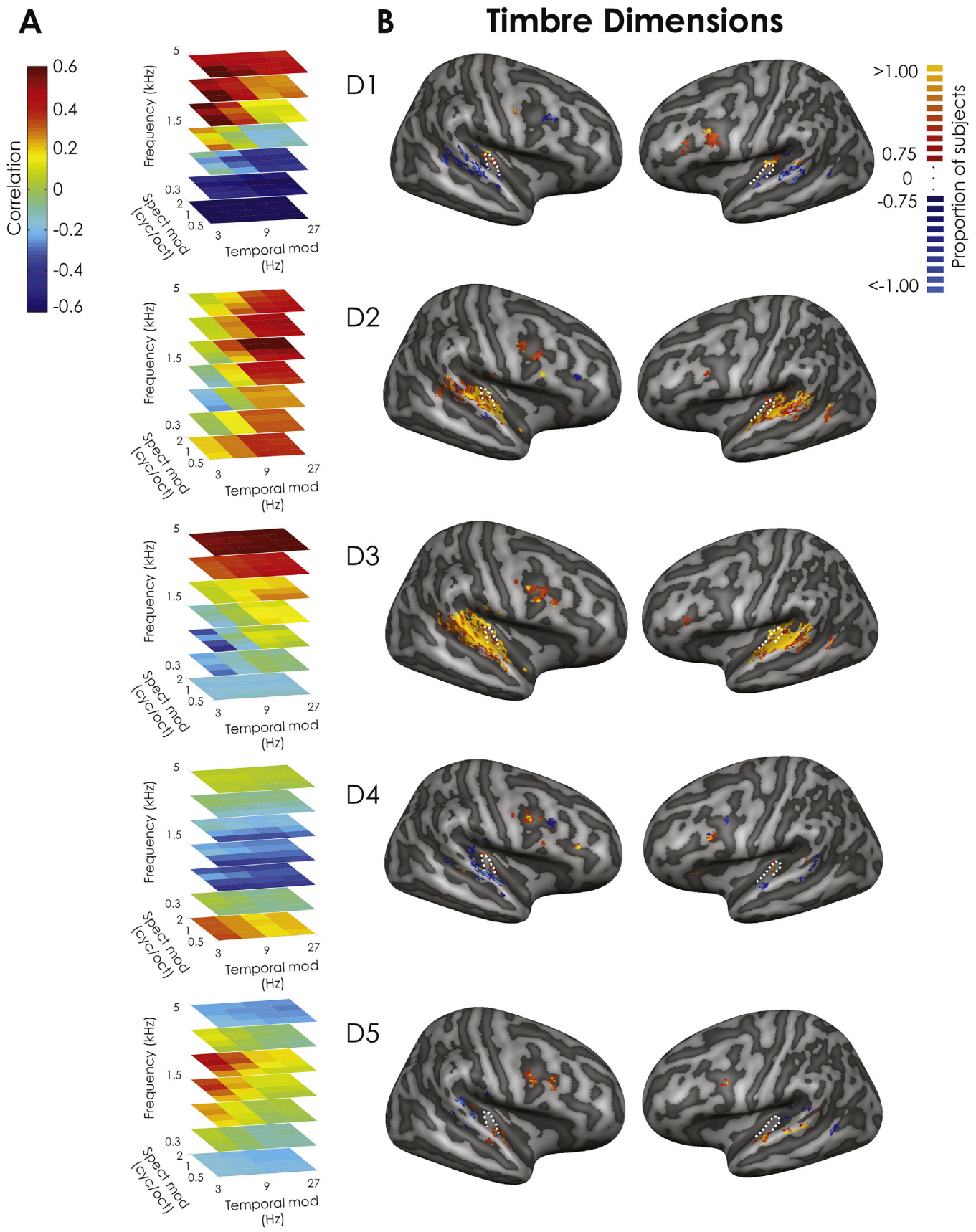

Fig. 6. Exploratory analyses of the timbre dimensions. (A) Slice plots showing marginal correlations between each of the five *timbre* dimensions and the features in the *STM* Model at several different frequencies. (B) Group-level maps of the five dimensions of the *timbre* model. For each timbre dimension, warm and cool colors reflect across-subject consistently positive and consistently negative scores, respectively. A positive or negative weight reflects that as sounds scored higher or lower on that dimension, respectively, the BOLD response in the voxel increased. White dotted lines indicate HG.

optimally combine models that explain different parts of the variance (see e.g., de Heer et al., 2017).

In addition to auditory regions, responses to sounds in frontal regions, such as the inferior frontal gyrus (IFG), were consistently predicted above chance across models. This may indicate that timbre features are also represented in frontal regions, but could also reflect higher-level auditory processing that is correlated with the features of the employed encoding models. One possible explanation is that model accuracy in frontal regions could be driven by sound recognition, since our stimuli were common musical instruments. Maeder et al. (2001) found certain regions to be more active for sound recognition compared to sound localization, including the left posterior IFG. Further, Broca's area may be included in the well-predicted cortical regions. While Broca's region is typically thought to be a higher-level language processing area, it has been suggested to also play a role in music processing (for a review, see Fadiga et al., 2009).

Timbre is a notoriously elusive acoustic feature to define and to investigate experimentally. In this study, the use of fMRI encoding (Naselaris and Kay, 2015) allowed us to explicitly test the representation of timbre-varying sounds throughout cortical neuronal populations. Employing natural sounds, this approach furthermore ensured that timbre varied across sounds in an ecologically valid manner. While many earlier studies have used encoding models that represented the physical characteristics of natural images (Kay et al., 2008; Naselaris et al., 2009) or sounds (Santoro et al., 2014), our work along with more recent studies (Huth et al., 2016; Kay and Yeatman, 2017) demonstrates the utility of incorporating higher-level perceptual features into the encoding models. This represents a next evolution in fMRI encoding, where the method can be used to tackle those aspects of perception and cognition that are extremely challenging to capture using classical approaches.

The *timbre* model provides an efficient representation of processing in human auditory cortex via a compact model whose features are based on subjective ratings of timbre. Our results suggest that the distributed neural representation of timbre in the cortex may align with perceptual categorizations of timbre. Consequently, it may be possible to assign semantic labels to the multidimensional tuning of neuronal populations. Since the employed *timbre* model was customized for this particular set of orchestral instruments, studies that test a broader range of stimuli (i.e., more musical instruments, speech, and other natural sounds) are recommended in order to determine the extent of this model's generalizability.

Acknowledgements

This work was supported by the National Institute of Deafness and other Communication Disorders at the National Institutes of Health (grant number R01 DC005216), the Brain Imaging Initiative of the College Liberal Arts, University of Minnesota, the Erasmus Mundus Student Exchange Network in Auditory Cognitive Neuroscience (ACN), the Netherlands Organisation for Scientific Research (NWO; VENI grant 451-15-012, and VICI grant 453-12-002), and the Dutch Province of Limburg. Juraj Mesik, Philip Burton, Cheryl Olman, Jordan Beim, and Taffeta Elliott provided helpful advice and assistance. The authors declare no competing financial interests.

References

Allen, E.J., Oxenham, A.J., 2014. Symmetric interactions and interference between pitch and timbre. J. Acoust. Soc. Am. 135, 1371–1379.

ANSI, 2013. S1.1-2013, American National Standard Acoustical Terminology (American National Standards Institute, New York, 1960).

Casey, M., Thompson, J., Kang, O., Raizada, R., Wheatley, T., 2012. Population Codes Representing Musical Timbre for High-level FMRI Categorization of Music Genres, in: Machine Learning and Interpretation in Neuroimaging. Springer, Berlin Heidelberg, pp. 34–41.

Chen, Z., Hu, G., Glasberg, B.R., Moore, B.C.J., 2011. A new method of calculating auditory excitation patterns and loudness for steady sounds. Hear. Res. 282, 204–215. https://doi.org/10.1016/j.heares.2011.08.001.

Chi, T., Ru, P., Shamma, S. a, 2005. Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. 118, 887. https://doi.org/10.1121/1.1945807.

de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E., 2017. The hierarchical cortical organization of human speech processing. J. Neurosci. 37, 6539–6557. https://doi.org/10.1523/JNEUROSCI.3267-16.2017.

Elliott, T.M., Hamilton, L.S., Theunissen, F.E., 2013. Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. J. Acoust. Soc. Am. 133, 389–404. https://doi.org/10.1121/1.4770244.

Fadiga, L., Craighero, L., D'Ausilio, A., 2009. Broca's area in language, action, and music. Ann. N. Y. Acad. Sci. 1169, 448–458. https://doi.org/10.1111/j.1749-6632.2009.04582.x.

Fritz, L., Mulders, J., Breman, H., Peters, J., Bastiani, M., Roebroeck, A., Andersson, J., Ashburner, J., Weiskopf, N., Goebel, R., 2014. Comparison of EPI distortion correction methods at 3T and 7T!. In: OHBM Annual Meeting.

Goebel, R., Esposito, F., Formisano, E., 2006. Analysis of functional image analysis contest (FIAC) data with BrainVoyager QX: from single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. Hum. Brain Mapp. 27, 392–401. https://doi.org/10.1002/hbm.20249.

Grey, J.M., 1977. Multidimensional perceptual scaling of musical timbres. J. Acoust. Soc. Am. 61, 1270–1277.

Halpern, A.R., Zatorre, R.J., Bouffard, M., Johnson, J.A., 2004. Behavioral and neural correlates of perceived and imagined musical timbre. Neuropsychologia 42, 1281–1292. https://doi.org/10.1016/j.neuropsychologia.2003.12.017.

Hotelling, H., 1936. Relations between two sets of variables. Biometrica 28, 321–377.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458. https://doi.org/10.1038/nature17637.

Joanisse, M.F., DeSouza, D.D., 2014. Sensitivity of human auditory cortex to rapid frequency modulation revealed by multivariate representational similarity analysis. Fr. ont.Neurosci 8, 306. https://doi.org/10.3389/fnins.2014.00306.

Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008. Identifying natural images from human brain activity. Nature 452, 352–355. https://doi.org/10.1038/nature06713.Identifying.

Kay, K.N., Rokem, A., Winawer, J., Dougherty, R.F., Wandell, B.A., 2013. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. Front. Neurosci. https://doi.org/10.3389/fnins.2013.00247.

Kay, K.N., Yeatman, J.D., 2017. Bottom-up and top-down computations in word- and face-selective cortex. Elife 6, 191–195. https://doi.org/10.7554/eLife.22341.

Langers, D.R.M., Backes, W.H., van Dijk, P., 2007. Representation of lateralization and tonotopy in primary versus secondary human auditory cortex. Neuroimage 34, 264–273. https://doi.org/10.1016/j.neuroimage.2006.09.002.

Maeder, P., Meuli, R., Adriani, M., Bellmann, A., Fornari, E., Thiran, J.-P., Pittet, A., Clarke, S., 2001. Distinct pathways involved in sound recognition and localization: a human fMRI study. Neuroimage 14, 802–816. https://doi.org/10.1006/nimg.2001.0888.

Menon, V., Levitin, D.J., Smith, B.K., Lembke, A., Krasnow, B.D., Glazer, D., Glover, G.H., McAdams, S., 2002. Neural correlates of timbre change in harmonic sounds. Neuroimage 17, 1742–1754. https://doi.org/10.1006/nimg.2002.1295.

Moerel, M., De Martino, F., Formisano, E., 2014. An anatomical and functional topography of human auditory cortical areas. Front. Neurosci. 8, 1–14. https://doi.org/10.3389/fnins.2014.00225.

Moerel, M., De Martino, F., Formisano, E., 2012. Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. J. Neurosci. 32, 14205–14216. https://doi.org/10.1523/JNEUROSCI.1388-12.2012.

Moore, B.C.J., 2014. Development and current status of the "cambridge" loudness models. Trends hear. 18 https://doi.org/10.1177/2331216514550620, 233121651455062.

Moore, B.C.J., Glasberg, B.R., 1990. Frequency discrimination of complex tones with overlapping and non-overlapping harmonics. J. Acoust. Soc. Am. 87, 2163–2177.

Naselaris, T., Kay, K.N., 2015. Resolving ambiguities of MVPA using explicit models of representation. Trends Cogn. Sci. 19, 551–554. https://doi.org/10.1016/j.tics.2015.07.005.

Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., Gallant, J.L., 2009. Bayesian reconstruction of natural images from human brain activity. Neuron 63, 902–915. https://doi.org/10.1016/j.neuron.2009.09.006.

Opolko, F., Wapnick, J., 2006. The McGill University Master Samples Collection on DVD.

Patil, K., Pressnitzer, D., Shamma, S., Elhilali, M., 2012. Music in our ears: the biological bases of musical timbre perception. PLoS Comput. Biol. 8, e1002759. https://doi.org/10.1371/journal.pcbi.1002759.

Patterson, R.D., Uppenkamp, S., Johnsrude, I.S., Griffiths, T.D., 2002. The processing of temporal pitch and melody information in auditory cortex. Neuron 36, 767–776. https://doi.org/10.1016/S0896-6273(02)01060-7.

Penagos, H., Melcher, J.R., Oxenham, A.J., 2004. A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. J. Neurosci. 24, 6810–6815. https://doi.org/10.1523/JNEUROSCI.0383-04.2004.

Plichta, M.M., Gerdes, A.B.M., Alpers, G.W., Harnisch, W., Brill, S., Wieser, M.J., Fallgatter, A.J., 2011. Auditory cortex activation is modulated by emotion: a functional near-infrared spectroscopy (fNIRS) study. Neuroimage 55, 1200–1207. https://doi.org/10.1016/j.neuroimage.2011.01.011.

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, e1003412. https://doi.org/10.1371/journal.pcbi.1003412.

Schoppe, O., Harper, N.S., Willmore, B.D.B., King, A.J., Schnupp, J.W.H., 2016. Measuring the performance of neural models. Front. Comput. Neurosci. 10, 10. https://doi.org/10.3389/fncom.2016.00010.

Staeren, N., Renvall, H., De Martino, F., Goebel, R., Formisano, E., 2009. Sound categories are represented as distributed patterns in the human auditory cortex. Curr. Biol. 19, 498–502. https://doi.org/10.1016/j.cub.2009.01.066.

Stepanek, J., 2006. Musical sound timbre: verbal descriptions and dimensions. In: Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06), pp. 121–126. Montreal.

Talairach, J., Tournoux, P., 1988. Co-planar Stereotaxic Atlas of the Human Brain. Thieme Medical, New York.

Tian, B., Rauschecker, J.P., 2004. Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. J. Neurophysiol. 92, 2993–3013. https://doi.org/10.1152/jn.00472.2003.

von Bismarck, G., 1974. Timbre of steady sounds: a factorial investigation of its verbal attributes. Acustica 30, 146–159.

Warren, J.D., Jennings, A.R., Griffiths, T.D., 2005. Analysis of the spectral envelope of sounds by the human brain. Neuroimage 24, 1052–1057. https://doi.org/10.1016/j.neuroimage.2004.10.031.

Warrier, C.M., Zatorre, R.J., 2002. Influence of tonal context and timbral variation on perception of pitch. Percept. Psychophys. 64, 198–207.

Zatorre, R.J., Belin, P., 2001. Spectral and temporal processing in human auditory cortex. Cereb. Cortex 11, 946–953.