Cortical Correlates of Attention to Auditory Features

Emily J. Allen,¹ Philip C. Burton,¹ Juraj Mesik,¹ Cheryl A. Olman,¹ and Andrew J. Oxenham¹

Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455

Pitch and timbre are two primary features of auditory perception that are generally considered independent. However, an increase in pitch (produced by a change in fundamental frequency) can be confused with an increase in brightness (an attribute of timbre related to spectral centroid) and vice versa. Previous work indicates that pitch and timbre are processed in overlapping regions of the auditory cortex, but are separable to some extent via multivoxel pattern analysis. Here, we tested whether attention to one or other feature increases the spatial separation of their cortical representations and if attention can enhance the cortical representation of these features in the absence of any physical change in the stimulus. Ten human subjects (four female, six male) listened to pairs of tone triplets varying in pitch, timbre, or both and judged which tone triplet had the higher pitch or brighter timbre. Variations in each feature engaged common auditory regions with no clear distinctions at a univariate level. Attending to one did not improve the separability of the neural representations of pitch and timbre at the univariate level. At the multivariate level, the classifier performed above chance in distinguishing between conditions in which pitch or timbre was discriminated. The results confirm that the computations underlying pitch and timbre perception are subserved by strongly overlapping cortical regions, but reveal that attention to one or other feature leads to distinguishable activation patterns even in the absence of physical differences in the stimuli.

Key words: attention; auditory cortex; fMRI; pitch; timbre

Significance Statement

Although pitch and timbre are generally thought of as independent auditory features of a sound, pitch height and timbral brightness can be confused for one another. This study shows that pitch and timbre variations are represented in overlapping regions of auditory cortex, but that they produce distinguishable patterns of activation. Most importantly, the patterns of activation can be distinguished based on whether subjects attended to pitch or timbre even when the stimuli remained physically identical. The results therefore show that variations in pitch and timbre are represented by overlapping neural networks, but that attention to different features of the same sound can lead to distinguishable patterns of activation.

Introduction

Pitch and timbre are two fundamental perceptual dimensions of sound. Variations in pitch carry information about intonation and melody, whereas timbre is closely related to sound quality and identity. Despite the importance of pitch and timbre in auditory and speech perception, it remains unclear how they are represented in the cortex. Although these dimensions are often studied independently, several studies have shown that they can

interact with one another perceptually (Beal, 1985; Melara and Marks, 1990; Moore and Glasberg, 1990; Marozeau and de Ch-veigné, 2007; Borchert et al., 2011; Allen and Oxenham, 2014). Given their perceptual interactions, it is plausible that they are represented in overlapping cortical regions. A recent fMRI study found that pitch and timbre variations were in fact represented in largely overlapping regions of the auditory cortex, although the patterns of activation could be distinguished using multivoxel pattern analysis (MVPA) (Allen et al., 2017). However, this conclusion was based on a passive listening task. In an earlier fMRI study, Warren et al. (2005) also found overlapping bilateral regions in the temporal lobes that were active for sounds varying in either fundamental frequency (F0) or spectral envelope shape. They found additional activation when alternating between harmonic and noise stimuli and inferred that the midportion of the right STS may process information specifically related to spectral envelopes (i.e., “brightness”). However, although subjects were instructed to attend to the stimuli, there was no explicit auditory task, so no manipulation of attention, and the stimuli were not equated for perceptual salience.

Received March 2, 2018; revised Feb. 12, 2019; accepted Feb. 13, 2019.

Author contributions: E.J.A., J.M., C.A.O., and A.J.O. designed research; E.J.A. and J.M. performed research; E.J.A., P.C.B., J.M., C.A.O., and A.J.O. analyzed data; E.J.A. wrote the first draft of the paper; E.J.A., P.C.B., J.M., C.A.O., and A.J.O. edited the paper; E.J.A., J.M., and A.J.O. wrote the paper.

This work was supported by the National Institutes of Health (Grant R01 DC005216), the High Performance Connectome Upgrade for Human 3T MR Scanner (Grant 1S10OD017974-01), and the Brain Imaging Initiative of the College Liberal Arts, University of Minnesota. We thank Andrea Grant, Anahita Mehta, and Jordan Beim for helpful assistance.

The authors declare no competing financial interests.

Correspondence should be addressed to Emily J. Allen at prac0010@umn.edu.

<https://doi.org/10.1523/JNEUROSCI.0588-18.2019>

Copyright © 2019 the authors

Auditory attention has been found to modulate activity in wide regions of the superior temporal gyrus (STG) (Jäncke et al., 1999; Degerman et al., 2006). A recent meta-analysis by Alho et al. (2014) compared neural representations of several sound dimensions and categories (pitch, spatial location, speech, and voice processing) during active and passive fMRI measurements. Although speech or voice processing loci were not found to change with attention, pitch was found to activate more posterior and lateral areas in STG during active tasks, whereas the passive listening loci were shifted more anteriorly, toward the lateral end of Heschl's gyrus, the macroanatomical landmark that corresponds most closely to primary auditory cortex. Although some studies have examined cortical representations of timbral dimensions (Menon et al., 2002; Reiterer et al., 2008; Allen et al., 2018), none has yet examined the effects of modulating attention to timbre. It thus remains possible that an attentionally demanding task may enhance the spatial separability of the cortical representations of pitch and timbre.

In addition to possible differences between active and passive listening conditions, subjects' attention can be directed to a particular sound feature. Recent studies have shown that attention can enhance the representation of a specific sound within a mixture (Mesgarani and Chang, 2012; Ding and Simon, 2012a,b) and there is some evidence for neuronal modulation in visual cortex as a function of attention to different features within a visual stimulus (Saenz et al., 2002). However, it is unknown whether attention to a specific auditory feature selectively enhances the representation of that feature over others rather than just enhancing the representation of the entire object.

This study investigated whether task-based attention enhances the separation of the neural correlates of pitch and timbre relative to that found in a passive-listening task (Allen et al., 2017) and whether neural correlates of attention to either pitch or timbre emerge even when the physical stimulus remains identical. The results suggest that the representations of pitch and timbre variation are subserved by strongly overlapping cortical regions even in the active task conditions, but reveal that attention to one or other dimension can lead to distinguishable activation patterns using MVPA even in the absence of physical differences between the stimuli.

Materials and Methods

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

Participants. Ten right-handed subjects (mean age: 27.5, SD: 3.2; 4 females, 6 males) from the University of Minnesota community participated in the experiment. All subjects had normal hearing, defined as audiometric pure tone thresholds of 20 dB hearing level or better at octave frequencies between 250 Hz and 8 kHz. The musical experience of the subjects ranged from 0 to 20 years, with the majority of subjects ($n = 7$) reporting 8 or more years of formal musical training, 1 reporting between 3 and 7 years, and the remaining 2 reporting 2 years or less.

Prescan experimental design. Before being scanned, each subject's difference limens (DLs) were measured for pitch and timbre discrimination. For pitch discrimination, we measured the DL for F0, the periodicity of a sound, a physical variable most closely associated with pitch. For timbre, we measured the DL for the spectral centroid or center frequency (CF), a physical manipulation that leads to reported changes in the timbral dimension of "brightness" or "sharpness" (Fastl and Zwicker, 2007). The paradigm was similar to that used by Allen and Oxenham (2014) and Allen et al. (2017). Band-pass-filtered, resolved harmonic complex tones (Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994) eliciting robust pitch and timbre percepts were generated in MATLAB (The

Figure 1. Spectrum of a harmonic complex tone with an F0 of 200 Hz and 24 dB/octave slopes around a 900 Hz spectral centroid or CF. The blue dot indicates the location of F0. The green dot indicates the location of the CF. Increasing or decreasing the F0 changes the spacing between the harmonics, which increases or decreases the perceived pitch of the tone, respectively. Increasing or decreasing the CF shifts the envelope of the complex to the right or left, respectively, which changes the relative amplitudes of the harmonics without changing the frequencies, leading to increases or decreases in the perceived brightness of the tone.

Figure 2. Chart showing all combinations of attended dimensions (pitch or timbre) and number of varying dimensions (one or two) in the study, totaling four different experimental conditions: PA, PwT, TA, and TwP. Note that the stimuli in the PwT and TwP conditions are identical; the only difference is the dimension to which subjects were instructed to attend.

MathWorks) and presented using the AFC toolbox for auditory psychophysics (Ewert, 2013). Pairs of successive tone triplets, three identical tones presented consecutively, were played diotically through HD 650 headphones (Sennheiser) at a sampling rate of 44,100 Hz. Each tone was 200 ms in duration for a total triplet duration of 600 ms, with a 300 ms interstimulus interval between the two triplets. The tones had 20 ms raised-cosine onset and offset ramps. Harmonics up to 10,000 Hz were added in sine phase and scaled to produce 24 dB/octave slopes around the center frequency (i.e., the spectral centroid). The 3 dB bandwidth of the stimulus around the center frequency was one-quarter octave, with the tone complexes having no flat band-pass region. Sounds were presented at an overall level of 70 dB sound pressure level (SPL).

Subjects listened to pairs of tone triplets presented sequentially and on each trial selected the triplet with the higher pitch or brighter timbre (i.e., a standard two-alternative forced-choice procedure). Triplets were cho-

sen with the aim of increasing the perceptual salience of the stimuli and obtaining a more robust cortical response (Thomas et al., 2015). Stimuli were paired with boxes on the screen that would light up with each triplet, along with the question, “Which pitch was higher?” or “Which timbre was brighter?” depending on the task. Feedback was provided after each response, indicating whether the subject was correct or incorrect. For the pitch condition, the CF of the stimulus remained unchanged at 900 Hz and the nominal F0 was roved $\pm 10\%$ uniformly around 200 Hz across trials. For the timbre task, the F0 remained unchanged at 200 Hz and the nominal CF was roved $\pm 10\%$ uniformly around 900 Hz across trials (Fig. 1). A two-down, one-up adaptive tracking rule was used to vary the difference in F0 (ΔF_0) or the difference in CF (ΔCF) between the two triplets to track the DL corresponding to performance of 70.7% correct (Levitt, 1971). In each trial, the two values of F0 or CF were geometrically centered around the nominal value and the order of presentation (low-high or high-low) was randomized with equal *a priori* probability. The starting value of ΔF_0 or ΔCF was 20%, defined relative to the lower of the two F0s or CFs, and was initially increased or decreased by a factor of 2. Following the second staircase reversal (i.e., the first direction change in the tracking variable from “up” to “down”), this factor was decreased to 1.26 and was then further decreased to the final factor of 1.12 after two more reversals. After six reversals at the smallest step size, the run was terminated and the DL in each run was calculated as the geometric mean of the ΔF_0 or ΔCF values at the last six reversals points. Each subject’s final DL for each dimension was based on the geometric mean DL across six task blocks. All blocks of one dimension were completed before beginning measurements in the next dimension and this ordering was counterbalanced across subjects.

After DLs were calculated, discrimination performance was measured using a method of constant stimuli with the attended dimension (F0 or spectral centroid) difference set to five times the DL measured for each individual subject (5DL) and the unattended dimension either being held constant or also varying by 5DL. The reason for multiplying the DL by 5 was 3-fold: (1) to allow subjects to perform near ceiling, confirming that they are attending to the correct dimension on each task; (2) to allow for the fact that the task was presented in the acoustically noisy MRI scanner environment (because the DLs were originally measured in silence); and (3) to ensure that the changes in pitch and timbre were approximately equally salient (Allen and Oxenham, 2014). Performance was based on 100 trials of each: pitch alone (PA) comparisons (when only F0 is varying), timbre alone (TA) comparisons (when only spectral centroid is varying), both pitch and timbre varying, but with subjects attending to only pitch (PwT), and both pitch and timbre varying, but with subjects attending only to timbre (TwP). In all cases, the subjects were instructed to select the tone with either the higher pitch or brighter timbre in the tone pair (Fig. 2).

Experimental design during scan. The stimuli were presented at an overall level of approximately 75 dB SPL via MRI-compatible Sensimetrics S14 earphones with custom filters designed to compensate for the frequency response of the hardware. Other stimulus parameters were the same as those for prescanner behavioral testing. During each task scan, subjects completed 10 trials of each of the four discrimination tasks using differences that were set to be five times the DL for each subject. The direction of change for each dimension (up or down) was selected ran-

Figure 3. Schematic diagram of an experimental session. **A**, Magnified view of a single block containing five trials. **B**, Counterbalancing of condition blocks within and across runs for a given subject. **C**, Magnified view of a single trial. Arrows represent jitter of the stimulus onset. b, Block; r, run; t, trial; acq, image acquisition. Note that the stimuli used in the PwT and TwP conditions were identical; only the task varied.

Figure 4. Mask of the auditory cortex and frontal lobe regions. Left to right, Left hemisphere, front-facing view, and right hemisphere.

domly and independently on each trial with equal *a priori* probability. These trials were evenly divided into eight blocks (with five trials in each block), one for each condition (Fig. 3). Stimulus onset during the silent period was randomly jittered to begin between 250 and 1000 ms after previous image acquisition. Blocks were separated by rest periods of between 10.75 and 12.25 s (due to stimulus presentation jitter) to measure the baseline signal. Trials were presented during inter-acquisition intervals to reduce acoustic contamination from the scanner. Following each trial, subjects had 2 s to respond. Subjects’ responses were collected via a button box. Stimuli were paired with boxes projected onto a screen and viewed through a mirror mounted on top of the head coil. The words “pitch,” “timbre,” “rest,” and “end” appeared on the screen as task cues. The order of the conditions in a run was randomized and there was a total of 60 trials per condition. Feedback for correct and incorrect responses appeared in the form of happy and sad emoticons, respectively. Missed responses were followed by a presentation of an asterisk symbol on the screen. Due to some technical difficulties with the button box, occasional subject responses were missed despite subjects having pressed the button during the allotted response window (mean number of missed responses across subjects: 1.9 of 240, SD: 2.1). These missed responses were not included in the calculation of task performance.

MRI. The data were acquired at a 3 T Siemens Prisma MRI scanner. To minimize the contamination of the functional data with scanner noise, we used a pulse sequence with sparse temporal acquisition (Hall et al., 1999). The pulse sequence used slice accelerated multiband (factor 2) echo planar imaging (EPI) (Xu et al., 2013) with a repetition time (TR) of 7 s (acquisition time of 2 s and an inter-acquisition silent interval of 5 s), providing a voxel resolution of 2 mm isotropic. Each functional volume

Figure 5. Cross-validation procedure for MVPA. **A**, Six functional runs represented as rows, each parsed into eight blocks of trials, totaling 48 blocks per subject. **B**, In each fold of the cross-validation, 40 of the blocks were used for training and eight were used for testing, with two blocks of each condition type represented in the testing data. b, Block; r, run.

had 48 slices angled upward to avoid the eyes in an effort to reduce eye movement artifacts while covering most of the brain. However, in many subjects, the posterior portion of the parietal cortex was not included. A total of 6 functional scans were acquired for each subject, each of which

Table 1. Prescan behavioral task performance

Condition(s)	Mean % (SD)	Rank order	Z-statistic	p-value (adjusted alpha = 0.008)
All pitch	96.8 (4.3)	All pitch < all timbre	−0.17	0.87
All timbre	97.3 (3.0)	TA < PA	−0.35	0.72
PA	98.3 (2.6)	PwT < TwP	−0.54	0.59
TA	98.1 (3.0)	BV < BA	−3.01	0.003*
BA	98.2 (2.7)	PwT < PA	−2.54	0.011
BV	95.9 (4.2)	TwP < TA	−1.49	0.14
PwT	95.2 (5.2)			
TwP	96.5 (3.0)			

Shown are the mean and SD of all conditions and Wilcoxon signed-rank tests Z-statistics and p-values comparing task performance between conditions. Asterisks indicate significant differences.

took approximately 6 min to complete and consisted of 48 volumes. To correct the spatial distortions from inhomogeneity in the B₀ magnetic field, reversed phase encode EPI volumes were also acquired. To localize functional activations, we additionally collected anatomical T1-weighted images that were coregistered with the EPI data.

Statistical analysis. Data were preprocessed using the Analysis of Functional NeuroImages (AFNI) software package (Cox, 1996) and FSL 5.0.9 (<http://fsl.fmrib.ox.ac.uk/>). Statistical analyses and visualization were performed with AFNI. Preprocessing included skull stripping, slice time correction, distortion correction, six-parameter motion correction, spatial smoothing (4 mm FWHM Gaussian blur), and prewhitening.

For each subject, an event-related general linear model (GLM) analysis was performed that included regressors for each of the four experimental conditions, six motion parameters, and Legendre polynomials up to the fourth order to account for baseline drift (modeled separately for each run). Each subject's brain was transformed into Montreal Neurological Institute (MNI) space (Mazziotta et al., 1995). Beta weights (regression coefficients) for individual voxels were estimated by the GLM for each condition for each subject, as were contrasts comparing conditions within pitch, within timbre, between pitch and timbre, and a contrast comparing all sounds to baseline. Cortical surface-based visualization was done in AFNI's SUMA (SURface MAPPING) <https://afni.nimh.nih.gov/Suma>. Group-level statistics were mapped onto the FreeSurfer brain surface MNI N27.

Group-level analyses with subject as a random effect included paired-sample *t* tests performed on the unmasked, unthresholded β weights for each contrast using the AFNI program 3dttest++. Voxels were thresholded at $p < 0.01$, uncorrected. Correction for multiple comparisons was achieved by determining the minimum significant cluster size. Taking into account increasing concerns over the risk of inflated false-positives with this method, as reported by Eklund et al. (2016), a nonparametric permutation test was used. This permutation

test randomized the signs of the residuals of the model among subjects and then performed a *t* test, with these steps iterated 10,000 times to determine nearest neighbor, faces touching, two-sided cluster thresholds. This method, implemented within 3dttest++ using the “clustsim” option, determined the probability, with each voxel having a 1% chance of displaying a false-positive, of clusters of a given size occurring by

Table 2. Behavioral task performance in the scanner

Condition(s)	Mean % (SD)
All pitch	99.3 (2.3)
All timbre	99.2 (1.2)
PA	98.6 (3.2)
TA	98.8 (1.4)
BA	98.7 (2.4)
BV	99.8 (0.5)
PwT	100 (0.0)
TwP	99.7 (0.7)

Figure 6. Group-level statistical map on an inflated brain showing the mean of all sound conditions relative to baseline (masked, thresholded at the single voxel level, $p < 0.01$; cluster thresholded, $p < 0.05$). Color scale values range from -1 to 1 units of percentage change. For reference, the grayed out areas, which are the regions not colored blue in Figure 4, have been added to denote regions not included in analysis.

chance. Based on these probabilities, clusters smaller than those that would occur by chance $>5\%$ of the time were filtered out of the results to achieve a cluster-level $\alpha = 0.05$. The t tests were conducted within a gray matter mask containing anatomically defined auditory cortices and frontal lobe regions (Fig. 4). The mask, which was created on the cortical surface, was made up of the following gyri and sulci in the left and right hemispheres: superior temporal (including banks), Heschl's, supramarginal, precentral, superior frontal, middle frontal (caudate and rostral), inferior frontal (opercularis, triangularis, orbitalis), orbito-frontal (lateral and medial), and anterior cingulate (caudal and rostral), as well as the insulae, temporal poles, and frontal poles. These regions were defined by the Desikan–Killiany atlas (Desikan et al., 2006). Frontal regions were included in the mask because of the incorporation of an attentional task, which made activation in frontal regions seem plausible. In addition, in our previous study (Allen et al., 2018), which included an active task with subjects listening to different musical instrument sounds (i.e., natural variations in timbre), we found robust responses in several frontal regions.

MVPA. In addition to univariate analyses, we used MVPA, which has the advantage of being more sensitive to differences between conditions than the univariate approach because it examines differential patterns of activity across voxels as opposed to averaging across them (Norman et al., 2006) and may reveal differences at the voxel level that are not apparent via standard univariate analyses. MVPA was performed using Princeton's MVPA toolbox for MATLAB with a backpropagation classifier algorithm (<http://code.google.com/p/princeton-mvpa-toolbox/>). MATLAB's neural network toolbox with no hidden layers was used for classifier training/testing. To restrict the number of voxels in our analyses, we added a functionally defined mask based on our univariate analysis results. This mask contained voxels that were most active during the auditory tasks (all sound conditions contrasted with the silence baseline) thresholded to the 2000 most active (positive) voxels across both hemispheres for each subject. This cutoff was chosen so that the number of voxels in each mask was the same across subjects and to limit the number of voxels used for classification (De Martino et al., 2008; Schindler et al., 2013).

The preprocessed data were detrended using a third-order polynomial via AFNI's processing pipeline. Spike-contaminated trials in the time

Figure 7. MVPA average classifier performance for each of the 10 subjects (indicated by bar color) on three different classifications. The horizontal red lines denote chance performance and the horizontal black lines denote average performance across subjects.

course of the preprocessed data exceeding 4 SDs from the mean of the run (excluding silence) were excluded from analysis. The remaining trials within each block were then averaged together, resulting in one value per block per voxel. The runs were z -scored and subjected to sixfold cross-validated MVPA analysis, where each training set consisted of five runs, totaling 40 patterns, whereas the remaining run (eight patterns) was used as the testing set (Fig. 5). Voxels that did not vary significantly in activation across the conditions based on an ANOVA done on the training set with a threshold of $p < 0.05$ were considered uninformative. In each cross-validation fold, these uninformative voxels were removed from the initial 2000-voxel mask. Classifier performance was computed for each fold separately and then averaged together for a final prediction accuracy score.

Results

Prescan behavioral task performance

The geometric mean F0 DL across subjects was 0.50% [95% confidence interval (CI) = 0.3–0.9] and the average spectral-

Figure 8. MVPA average classifier performance for each of the 10 subjects (indicated by bar color). Left to right, TA versus PA, PwT versus TwP, training on alone conditions (PA vs TA) and testing on the varying conditions (PwT vs TwP), and training on the varying conditions and testing on the alone conditions. The horizontal red lines denote chance performance and the horizontal black lines denote average performance across subjects.

Table 3. Exploratory MVPA

	Mean % (SD)	$t_{(9)}$	p -value
Subset of conditions			
PA vs TA	80.8 (8.4)	8.4	0.0001*
PwT vs TwP	82.9 (7.7)	7.7	0.0001*
PA vs PwT	77.9 (10.2)	8.6	0.0001*
TA vs TwP	81.3 (8.6)	11.5	0.0001*
Train on BA, test on BV	55.0 (14.0)	1.13	0.14
Train on BV, test on BA	57.9 (16.5)	1.52	0.082
Randomized labels			
Four-way (chance: 25%)	24.4 (7.4)	0.27	0.60
Pitch vs timbre	51.0 (9.7)	0.34	0.37
BA vs BV	52.1 (10.9)	0.60	0.28

Shown are classifier performance on subsets of the conditions and with randomized labels (should be at chance).

centroid DL was 2.7% (95% CI = 2.1–3.6), in good agreement with musicians's DLs from an earlier study using similar stimuli (Allen and Oxenham, 2014). This agreement was expected because most of the current subjects would be classified as musicians, having had >8 years of musical training. As anticipated, performance on the constant-stimuli task using the scaled 5DL variation was high, with a mean proportion of correct responses of 97.0% (SD: 3.7%) across all conditions. Average performance in each condition is reported in Table 1. Due to the near-ceiling performance in these tasks, a nonparametric Friedman test was used on the four conditions, which indicated a significant effect of condition ($\chi^2 = 11.7$, $p < 0.008$). We then ran Wilcoxon signed-rank tests to compare conditions. As reported in Table 1, after a Bonferroni correction for multiple comparisons with an adjusted α of 0.008, no significant difference in performance between the pitch and timbre conditions was found. This was true when comparing PA and TA conditions as well as the PwT and TwP conditions. There was, however, a significant difference between the BA and BV conditions, suggesting that the conditions in which both dimensions were varying were more challenging than the conditions in which only one dimension was varying. However, this significance was marginal between PA and PwT after correction and was not significant for TA and TwP.

Behavioral task performance during scanning

Similar to performance in the prescanner session, performance in the scanner was near ceiling (mean: 99.3%, SD: 1.8). Average performance within conditions is reported in Table 2. This time, a nonparametric Friedman test did not reveal a significant effect of condition ($\chi^2 = 7.77$, $p = 0.051$).

Group-level analysis

For all conditions, we found robust activation of the auditory cortices when comparing sound stimulation with baseline silence (Fig. 6). However, when comparing activation between pairs of conditions that involved sound stimulation (P-T, BA-BV, PA-TA, and PwT-TwP), no significant voxels remained after thresholding.

MVPA results

As shown in Figure 7, average 4-way classifier performance for distinguishing between pitch alone, timbre alone, pitch varying, and timbre varying was 74.4% (SD = 9.7 percentage points), which was significantly above chance (25%) based on a one-tailed t test ($t_{(9)} = 16.1$, $p < 0.0001$). Average classifier performance for pitch conditions versus timbre conditions was 82.1% (SD = 6.5), which was also significantly above chance (50%) ($t_{(9)} = 15.5$, $p < 0.0001$). Average classifier performance for both alone (BA) conditions versus both varying (BV) conditions was 80.0% (SD = 7.0), which, again, was significantly above chance (50%) ($t_{(9)} = 13.5$, $p < 0.0001$).

Exploratory MVPA and univariate analyses

We also performed MVPA on subsets of the conditions PA versus TA and PwT versus TwP conditions (Fig. 8, left), as well as comparing PA versus PwT and TA versus TwP. In all cases, classifier performance was significantly above chance. Additionally, we tested how well the MVPA classifier performed if it was trained on the BA conditions (PA vs TA) but tested on the BV conditions (PwT vs TwP) and vice versa (Fig. 8, right). In neither case was average performance across subjects significantly above chance. All results are reported in Table 3.

We analyzed the errors in classifier performance for the four-way classifier to determine whether, despite its high performance, there were any conditions that were consistently confused with one another (e.g., when the correct condition was PA, was it classified more often as PwT than TA or TwP?). Results are shown in Figure 9. It is difficult to draw any strong conclusions because there are no obvious indications of the classifier consistently confusing one condition for another.

Next, as a sanity check, we randomized the condition labels to determine whether classifier performance would drop to chance, as would be expected. Indeed, with random labels, no classifier performed significantly above chance (Table 3).

Finally, we reran the MVPA using two additional masks. The first mask was an intersection of our original 2000-voxel mask for each subject and anatomically defined auditory regions; the second mask was an intersection between the same 2000-voxel mask and the anatomically defined frontal regions. This separation of the 2000-voxel masks into the frontal versus auditory regions was done to test whether the high MVPA classifier performance was

primarily driven by voxels in a particular region. The number of voxels included in each mask for each subject is shown in Table 4. MVPA performance within these masks is shown in Figure 10 and Table 5.

The four-way classifier when using the 2000-voxel mask intersected with auditory regions resulted in an average performance of 70.4% (SD = 8.3, which, again, was significantly above chance ($t_{(9)} = 17.3$, $p < 0.0001$). The 4-way classifier when using the 2000-voxel mask intersected with frontal regions resulted in an average performance of 61.9% (SD = 15.4), which was also significantly above chance ($t_{(9)} = 7.6$, $p < 0.0001$). This difference was significant based on a paired-samples t test ($t_{(9)} = 2.28$, $p < 0.048$), suggesting that activity in auditory regions was somewhat more informative than activity in frontal regions in distinguishing between stimuli and tasks.

Discussion

The present study aimed to determine whether task-related attention to one dimension when listening to sounds varying in pitch and/or brightness would lead to more spatially distinct representations of pitch and timbre. Univariate analyses suggested that this was not the case either at the group level or for the majority of subjects at the individual level. No significant differences in activation were observed between the pitch-varying and timbre-varying conditions regardless of whether variations in the other dimension were present. Therefore, it seems that the spatial overlap between representations of pitch and brightness is observed under active and passive (Allen et al., 2017) listening conditions. This similarity in the spatial distribution of cortical processing of the two dimensions may not be entirely surprising given the perceptual interactions and dependencies that have been identified between these dimensions (Melara and Marks, 1990; Borchert et al., 2011; Allen and Oxenham, 2014).

Although the univariate analyses could not differentiate between condition pairs, high MVPA classification performance was obtained with the four-way classifier (PA vs PwT vs TA vs TwP), as well as the two-way classifiers tested (alone vs BV, pitch vs timbre, PA vs TA, and PwT vs TwP). Despite there being no significant overall spatial differences between pitch and timbre representation, it is clear that the patterns of activation within these regions are distinct. This outcome is consistent with earlier findings obtained under passive listening conditions (Allen et al., 2017) and extends them by showing that pattern classification remains possible both in the presence and absence of an attentional task. The only classifiers that did not exceed chance performance on average were training on PA and TA while testing on PwT and TwP and vice versa. It is possible that performance suffered with these two classifiers due to representations of pitch and timbre interacting in a nonlinear fashion (e.g., via divisive normalization) and that these nonlinearities are stronger for the conditions in which both dimensions are varying. As a result, this may lead to features that are helping classifiers when one dimension is varying that are not helpful when both dimensions are varying. Therefore, when both types of stimuli are not present in the training dataset, the classifier may struggle.

Figure 9. Classifier confusion patterns for the four-way classifier. Bar graphs show average classifier guesses for each condition (PA, TA, PwT, and TwP). Individual subject classifier guesses are superimposed on the corresponding bars. x-axis labels indicate the correct condition.

Table 4. Number of voxels included in each mask for each subject

Subject	2000-Voxel mask intersected with auditory regions (voxels)	2000-Voxel mask intersected with frontal regions (voxels)
1	1111	889
2	1722	278
3	1443	557
4	1535	465
5	1730	270
6	1434	566
7	1328	672
8	1312	688
9	1331	208
10	1649	351
Mean (SD)	1459.5 (200.5)	494.4 (219.7)

The left panel shows the number of voxels included when the original 2000-voxel mask is intersected with the anatomically defined auditory regions. The right panel shows the number of voxels included when the same 2000-voxel mask is intersected with the anatomically defined frontal regions.

Particularly striking is the fact that, in the absence of physical differences in the stimuli and without significant differences in the behavioral performance of the subjects, the classifier performed well above chance for identifying the pitch and timbre conditions, successfully distinguishing between the PwT and TwP conditions. These results suggest that representations within auditory cortex may reflect the attention to particular sound features within a single auditory object. In this way, the outcomes extend earlier work using EEG, MEG, and ECoG, which has shown that attention to entire auditory objects or streams (e.g., one voice in the presence of another) can profoundly alter cortical activity (Hillyard et al., 1973; Ding and Simon, 2012a; Zion Golumbic et al., 2013). There is evidence to suggest that such feature-based attentional modulation also exists in visual cortex for visual objects (Saenz et al., 2002).

It is important to note that timbre is well established as a multidimensional dimension (Grey, 1977; Elliott et al., 2013; Allen et al., 2018). The present study manipulated timbre along just one dimension by shifting the spectral centroid, which affects a sound's perceived "brightness." Therefore, we are unable to draw

Figure 10. MVPA average classifier performance for each of the 10 subjects comparing the intersection of the 2000-voxel mask with the anatomically defined auditory regions with the intersection of the 2000-voxel mask with the anatomically defined frontal regions. The horizontal red lines denote chance performance and the horizontal black lines denote average performance across subjects.

Table 5. Average classifier performance within two masks: the intersection of the 2000-voxel mask with the anatomically defined auditory regions (left) and with the intersection of the 2000-voxel mask with the anatomically defined frontal regions (right)

Classifier	Auditory mean % (SD)	Frontal mean % (SD)
Four-way	70.4 (8.3)	61.9 (15.4)
Pitch vs timbre	79.6 (7.7)	73.8 (16.1)
BA vs BV	80.0 (9.7)	77.1 (8.7)
PA vs TA	82.5 (8.3)	75.0 (21.2)
PwT vs TwP	77.1 (13.8)	82.1 (8.8)
Train on PA vs TA, test on PwT vs TwP	52.9 (16.4)	54.6 (22.8)
Train on PwT vs TwP, test on PA vs TA	53.8 (13.4)	58.8 (22.0)

conclusions regarding other manipulations of timbre such as temporal modulations. It may be that cortical representations of the spectral properties of timbre, such as spectral centroid, would be more similar to the cortical representations of F0 than, for example, variations in attack and decay.

It is also worth noting that average performance in the behavioral tasks was very high, exceeding 90% correct in all conditions. Although this provides a strong indication that subjects attended to the correct dimension, it may also imply that the attentional demands of the task were not high. A possible follow-up study might involve increasing the task difficulty to determine whether spatial differences would emerge that were not found in the present study. However, it is important in such a design that the task difficulty is equated across dimensions so any differences that emerge cannot simply be explained by the task in one dimension being more challenging than the other.

Another future direction would be to use encoding models to explicitly characterize pitch and timbre selectivity throughout the auditory cortex and thus explore in more depth how these populations are modulated by attention. These approaches have been successfully used to characterize suppressive stimulus interactions in the visual system (Brouwer and Heeger, 2011) and the somatosensory system (Brouwer et al., 2015) and could be similarly useful in understanding the interactions between representations of pitch and timbre.

Overall, these results show that attending to either pitch or brightness does not result in a spatial separation in their representations that is detectable via conventional univariate analyses. Nevertheless, the patterns of activation produced when attending to pitch or timbre features within a sound appear to be distinct even when the stimuli remain physically identical. This finding reveals that activation in both auditory and frontal cortical regions can reflect attention to specific sound features within single auditory objects.

References

Alho K, Rinne T, Herron TJ, Woods DL (2014) Stimulus-dependent activations and attention-related modulations in the auditory cortex: a meta-analysis of fMRI studies. *Hear Res* 307:29–41.

Allen EJ, Oxenham AJ (2014) Symmetric interactions and interference between pitch and timbre. *J Acoust Soc Am* 135:1371–1379.

Allen EJ, Burton PC, Olman CA, Oxenham AJ (2017) Representations of pitch and timbre variation in human auditory cortex. *J Neurosci* 37:1284–1293.

Allen EJ, Moerel M, Lage-Castellanos A, De Martino F, Formisano E, Oxenham AJ (2018) Encoding of natural timbre dimensions in human auditory cortex. *Neuroimage* 166:60–70.

- Beal AL (1985) The skill of recognizing musical structures. *Mem Cognit* 13:405–412.
- Borchert EM, Micheyl C, Oxenham AJ (2011) Perceptual grouping affects pitch judgments across time and frequency. *J Exp Psychol Hum Percept Perform* 37:257–269.
- Brouwer GJ, Heeger DJ (2011) Cross-orientation suppression in human visual cortex. *J Neurophysiol* 106:2108–2119.
- Brouwer GJ, Arnedo V, Offen S, Heeger DJ, Grant AC (2015) Normalization in human somatosensory cortex. *J Neurophysiol* 114:2588–2599.
- Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res* 29:162–173.
- Degerman A, Rinne T, Salmi J, Salonen O, Alho K (2006) Selective attention to sound location or pitch studied with fMRI. *Brain Res* 1077:123–134.
- De Martino F, Valente G, Staeren N, Ashburner J, Goebel R, Formisano E (2008) Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage* 43:44–58.
- Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31:968–980.
- Ding N, Simon JZ (2012a) Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc Natl Acad Sci U S A* 109:11854–11859.
- Ding N, Simon JZ (2012b) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J Neurophysiol* 107:78–89.
- Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci U S A* 113:7900–7905.
- Elliott TM, Hamilton LS, Theunissen FE (2013) Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. *J Acoust Soc Am* 133:389–404.
- Ewert SD (2013) AFC - A modular framework for running psychoacoustic experiments and computational perception models. In: *Proceedings of the International Conference on Acoustics AIA-DAGA*, pp 1326–1329. Merano, Italy. Available at http://pub.dega-akustik.de/AIA_DAGA_2013/data/articles/000400.pdf.
- Fastl H, Zwicker E (2007) *Psychoacoustics: facts and models*. New York: Springer.
- Grey JM (1977) Multidimensional perceptual scaling of musical timbres. *J Acoust Soc Am* 61:1270–1277.
- Hall DA, Haggard MP, Akeroyd MA, Palmer AR, Summerfield AQ, Elliott MR, Gurney EM, Bowtell RW (1999) “Sparse” temporal sampling in auditory fMRI. *Hum Brain Mapp* 7:213–223.
- Hillyard SA, Hink RF, Schwent VL, Picton TW (1973) Electrical signs of selective attention in the human brain. *Science* 182:177–180.
- Houtsma AJM, Smurzynski J (1990) Pitch identification and discrimination for complex tones with many harmonics. *J Acoust Soc Am* 87:304–310.
- Jäncke L, Mirzazade S, Shah NJ (1999) Attention modulates activity in the primary and the secondary auditory cortex: a functional magnetic resonance imaging study in human subjects. *Neurosci Lett* 266:125–128.
- Levitt H (1971) Transformed up-down methods in psychoacoustics. *J Acoust Soc Am* 49:467–477.
- Marozeau J, de Cheveigné A (2007) The effect of fundamental frequency on the brightness dimension of timbre. *J Acoust Soc Am* 121:383–387.
- Mazziotta JC, Toga AW, Evans A, Fox P, Lancaster J (1995) A probabilistic atlas of the human brain: theory and rationale for its development. *Neuroimage* 2:89–101.
- Melara RD, Marks LE (1990) Interaction among auditory dimensions: timbre, pitch, and loudness. *Percept Psychophys* 48:169–178.
- Menon V, Levitin DJ, Smith BK, Lembke A, Krasnow BD, Glazer D, Glover GH, McAdams S (2002) Neural correlates of timbre change in harmonic sounds. *Neuroimage* 17:1742–1754.
- Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485:233–236.
- Moore BC, Glasberg BR (1990) Frequency discrimination of complex tones with overlapping and non-overlapping harmonics. *J Acoust Soc Am* 87:2163–2177.
- Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* 10:424–430.
- Reiterer S, Erb M, Grodd W, Wildgruber D (2008) Cerebral processing of timbre and loudness: fMRI evidence for a contribution of Broca’s area to basic auditory discrimination. *Brain Imaging Behav* 2:1–10.
- Saenz M, Buracas GT, Boynton GM (2002) Global effects of feature-based attention in human visual cortex. *Nat Neurosci* 5:631–632.
- Schindler A, Herdener M, Bartels A (2013) Coding of melodic gestalt in human auditory cortex. *Cereb Cortex* 23:2987–2993.
- Shackleton TM, Carlyon RP (1994) The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. *J Acoust Soc Am* 95:3529–3540.
- Thomas JM, Huber E, Stecker GC, Boynton GM, Saenz M, Fine I (2015) Population receptive field estimates of human auditory cortex. *Neuroimage* 105:428–439.
- Warren JD, Jennings AR, Griffiths TD (2005) Analysis of the spectral envelope of sounds by the human brain. *Neuroimage* 24:1052–1057.
- Xu J, Moeller S, Auerbach EJ, Strupp J, Smith SM, Feinberg DA, Yacoub E, Ugurbil K (2013) Evaluation of slice accelerations using multiband echo planar imaging at 3 T. *Neuroimage* 83:991–1001.
- Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013) Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” *Neuron* 77:980–991.