ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Special Issue: *The Neurosciences and Music VI*
ORIGINAL ARTICLE

Cross-classification of musical and vocal emotions in the auditory cortex

Sébastien Paquette,[1,2] Sylvain Takerkart,[3] Shinji Saget,[3] Isabelle Peretz,[1] and Pascal Belin[1,3,4]

[1]Department of Psychology, International Laboratory for Brain Music and Sound Research, Université de Montréal, Montreal, Canada. [2]Department of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts. [3]Institut de Neurosciences de La Timone, CNRS & Aix-Marseille University, Marseille, France. [4]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, United Kingdom

Address for correspondence: Sébastien Paquette, Department of Neurology, Beth Israel Deaconess Medical Center, 330 Brookline Avenue - Palmer 127, Boston, MA 02215. spaquet1@bidmc.harvard.edu; Pascal Belin, Institut de Neurosciences de La Timone, CNRS & Aix-Marseille University, Marseille, France. pascal.belin@univ-amu.fr.

Whether emotions carried by voice and music are processed by the brain using similar mechanisms has long been investigated. Yet neuroimaging studies do not provide a clear picture, mainly due to lack of control over stimuli. Here, we report a functional magnetic resonance imaging (fMRI) study using comparable stimulus material in the voice and music domains—the Montreal Affective Voices and the Musical Emotional Bursts—which include nonverbal short bursts of happiness, fear, sadness, and neutral expressions. We use a multivariate emotion-classification fMRI analysis involving cross-timbre classification as a means of comparing the neural mechanisms involved in processing emotional information in the two domains. We find, for affective stimuli in the violin, clarinet, or voice timbres, that local fMRI patterns in the bilateral auditory cortex and upper premotor regions support above-chance emotion classification when training and testing sets are performed within the same timbre category. More importantly, classifier performance generalized well across timbre in cross-classifying schemes, albeit with a slight accuracy drop when crossing the voice–music boundary, providing evidence for a shared neural code for processing musical and vocal emotions, with possibly a cost for the voice due to its evolutionary significance.

Keywords: music; voice; emotion; multivoxel pattern analysis; cross-classification; auditory cortex

Introduction

Voice and music are arguably the most effective sound categories for conveying emotions. These emotions can be recognized easily from short affect bursts without any verbal content, such as giggles or screams.[1,2] Similarly, musical emotions can be recognized quickly, in as little as a quarter of a second—the equivalent of a chord or a few melodic notes.[3,4] Moreover, the expression of emotion in voice and music is based on the modulation of comparable acoustical cues (e.g., tempo, intensity, pitch range).[5,6] For example, the span of major musical intervals has been found to be more similar to that of excited speech, whereas the range of minor intervals was more similar to that of subdued speech.[5,7] It has thus been proposed that music could tap into neuronal circuits that have evolved primarily for the processing of biologically important emotional vocalizations;[8,9] the common acoustical characteristics shared by both domains would promote neural recycling/co-opting of circuits.[10]

Accordingly, several studies have compared emotional processing of vocal and musical stimuli, using behavioral tests in patients or neuroimaging techniques in healthy volunteers. Evidence indicates that brain-damaged patients with lesions involving the amygdala show emotion processing deficits for both vocal[11–13] and musical[14] stimuli, consistent with the pivotal role of this structure in processing emotional information in both domains.

A few studies have used neuroimaging to compare brain activity patterns in response to affective stimuli in the vocal and musical domains.[15,16] Results suggest largely overlapping activation patterns with important differences: voice processing is associated

doi: 10.1111/nyas.13666

with higher activity in the middle/superior temporal cortex and in the superior temporal sulcus (STS) and music processing with higher activity in the superior temporal gyrus (STG), including the planum temporale and Heschl's gyrus.

Those differences highlight a first challenge in comparing the neural substrate of vocal and musical emotions. Some differences in activation patterns could reflect genuine differences in timbre-specific emotional processing and the selective recruitment of category-specific cortex: the voice-selective temporal voice areas (TVAs),[17–19] particularly its emotional subdivision,[20] or music-selective areas.[21,22] But the differences could also reflect across-domain differences in the presence of linguistic information. Escoffier *et al.* used long (>10 s) musical and vocal stimuli and contrasted brain activation elicited by complex familiar music with that elicited by meaningless vowels;[16] Aubé *et al.* contrasted short excerpts of music written in the Western tonal system (some on discrete pitch instrument) with vocalization stimuli without controlling for the segmental structure (phonological units; vowels, consonants) of the auditory stimuli.[15] Although highly informative, these studies highlight the need for properly matched stimulus sets in order to minimize activation differences that may reflect low-level acoustics rather than higher level emotional value.

Here, we use two stimulus sets consisting of short, nonverbal affective stimuli. The Montreal Affective Voices (MAV) are a set of affect bursts produced by several actors and portraying different emotions on the vowel "a";[1] the Musical Emotional Bursts (MEB) consist of short musical bouts produced by professional musicians and expressing four emotions (happiness, sadness, fear, and neutral) on a violin or a clarinet.[23] We used two different musical timbres to separate music-specific processing from timbre-specific processing. Both MAV and MEB stimuli are nonverbal and of similar short duration and thus constitute appropriate material for a cross-domain comparison.

A challenge in using conventional functional magnetic resonance imaging (fMRI) analysis to test the hypothesis of similar neural networks engaged by the musical and vocal domain is that it predicts no difference. The absence of differences may be due to several factors, the most likely being lack of sensitivity to small differences. Here, we circumvent this problem by using multivoxel pattern analysis (MVPA)[24] instead of the traditional subtraction-based massive univariate analysis: inference is based on the ability of machine learning classifiers to classify emotions based on local fMRI patterns measured in response to the affective stimuli. Specifically, we use cross-classification as a powerful test of shared neural networks: if cerebral areas represent the emotional information content of the vocal and musical stimuli similarly, then an MVPA classifier trained on fMRI patterns elicited by musical stimuli should yield above-chance performance in classifying fMRI patterns measured for vocal stimuli, even though it was not trained with them; conversely, a classifier trained with fMRI patterns measured for vocal stimuli should also perform well for those measured in response to the musical stimuli.

Methods

Participants

Twenty participants (10 female, 19 right-handed; average Edinburgh Handedness Inventory score: 79.4) were recruited through the departmental subject mailing list of the University of Glasgow. They reported normal hearing and no neurological disorders. Participants were between 18 and 29 years (mean = 22.9) and had on average 5.6 years (0–18 years) of musical education. Participants provided written informed consent before participation, in accordance with the Declaration of Helsinki. The local ethics committee at the University of Glasgow approved the protocol.

Stimuli

One hundred and twenty previously validated stimuli were used in the study: 40 (20 male, 20 female) short nonverbal emotional interjections selected from the MAV[1] and 80 (40 clarinet, 40 violin) short musical bursts selected from the MEB.[23] The 40 stimuli in each timbre were 10 different samples of each of four emotions (fear, sadness, happiness, neutral). Stimuli were recorded from 10 different actors (five male), 10 violinists, and 10 clarinetists. Stimulus duration was not significantly different across the three timbres (MAV: 1.31 $\pm$ 0.12 s; MEB: 1.57 $\pm$ 0.07 s; $F(2,117) = 2.05$, $P = 0.13$). All stimuli are available online (sebastienpaquette.com/downloads).

Procedures

Stimuli were presented in blocks, each composed of 40 stimuli (10 stimuli per emotion) of a given

timbre (three timbres) presented in a pseudorandom order with a 2–2.5 s jittered interstimulus interval. Four blocks of each of the three timbres (voice, clarinet, violin) were presented in a pseudorandom order separated by 20 s of silence, for a total of 12 blocks of auditory stimulation. Stimuli were played to the participant at a comfortable listening level over MRI-compatible in-ear headphones (S14, Sensimetrics Corporation) using an M-Audio Audiophile 2496 soundcard. Participants were instructed to keep their eyes closed. To make sure participants were paying attention to the stimuli, they were also instructed to perform a 1-back task by pressing a button using their index finger on an MRI-compatible response pad (Lumitouch, PhotonControl) every time they heard the same sound twice in a row (four repetitions per block). Participants did not miss more than two consecutive button presses.

After scanning, participants performed a four-alternative (fear, sadness, happiness, neutral) forced-choice emotion categorization task on the 120 stimuli.

MRI acquisition

Structural and blood oxygenation level–dependent functional images were acquired on a 3T Tim Trio Scanner (Siemens) and a 32-channel head coil at the Centre for Cognitive Neuroimaging (CCNi) of the Institute of Neuroscience and Psychology at the University of Glasgow. Functional images during the task were acquired using continuing fMRI scanning and a parallel-accelerated multi-echo echo planar imaging sequence (TE: 9.4/21.2/33.45/57). Volumes consisted of 32 axial slices (voxel size: $3 \times 3 \times 3$ mm^3; gap: 25%) and were acquired at a TR of 2.47 seconds. The acquisition ended with a T1-weighted anatomical scan (voxel size: $1 \times 1 \times 1$ mm^3).

MRI preprocessing

Data preprocessing was performed using statistical parametric mapping (SPM8; Wellcome Trust Centre for Neuroimaging), and analyses were performed using SPM12. All images were realigned to correct for head motion with the first volume as a reference. Images corresponding to the different echo times were then combined by weighted summation using the parallel-acquired inhomogeneity-desensitized method. T1-weighted structural images were coregistered to the mean image created by the realignment procedure and were used for normalization of functional images onto the Montreal Neurological Institute (MNI) atlas using normalization parameters derived from segmentation of the anatomical image. Finally, each image was smoothed with an isotropic 8-mm full-width-at-half-maximum Gaussian kernel.

MVPA analyses

In order to conduct our decoding analyses, a univariate general linear model was first fitted for each subject for estimating beta maps associated with each emotion, separately for each of the 12 blocks—48 beta maps in total. All MVPA analyses involved decoding which emotion was present in the stimuli that contributed to the corresponding beta map. We used a supervised learning approach that comprised training a classifier (support vector machine, linear kernel, $C = 1$) on a subset of the data (training data) and computing its generalization performance on the remaining data (test data). The process was repeated throughout several splits of the data set into training and test sets in order to obtain a robust estimate of the decoding accuracy. More specifically, two different analyses were performed.

First, a sliding-window strategy (searchlight, radius 8 mm) was used to map the within-timbre decoding power throughout the whole brain. For this, we defined a fourfold cross-validation within the four blocks of data that corresponded to a given timbre (voice, clarinet, violin), run independently using the data from each timbre. The results were averaged across folds, resulting in three classification accuracy maps per subject (one per timbre), showing for all gray matter voxels the emotion-classification accuracy of the classifier trained on fMRI data from the local sphere centered on that voxel in response to auditory stimulation with stimuli from that timbre.

A nonparametric, second-level analysis of the individual timbre-specific classification maps (after subtraction of the theoretical chance level 0.25) was performed using SnPM (http://warwick.ac.uk/snpm) with 10,000 permutations of conditions. This yielded a map of all voxels in which the distribution of subject-specific accuracy values combined across the three timbres was significantly different from 0 at $P = 0.05$ (family-wise error (FWE) correction for multiple comparisons). The results of this whole-brain searchlight analysis identified four regions where the within-timbre

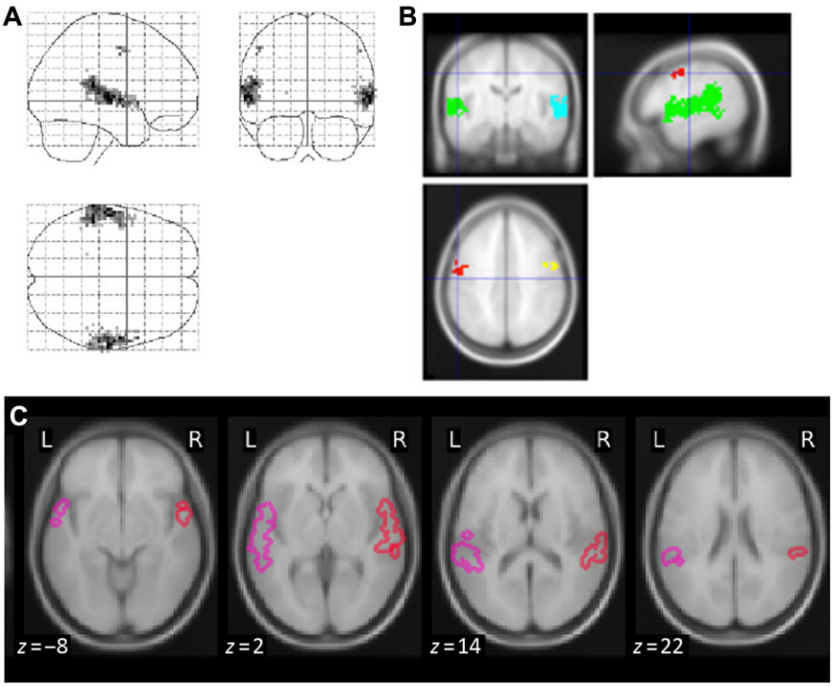

Figure 1. Cerebral regions supporting within-timbre emotion classification. (A) The four cerebral regions supporting above-chance within-timbre emotion classification shown in glass-brain representation: two large regions of the left and right auditory cortex encompassing most of the superior temporal gyrus and two small regions of the left and right premotor cortex. (B) The four ROIs shown on selected slices in the three orientations. Green: left temporal; blue: right temporal; red: left premotor; yellow: right premotor. (C) Outline of the two temporal ROIs on axial sections.

emotion decoding accuracy was significantly greater than chance (Fig. 1).

Second, we examined the consistency of the multivariate patterns across timbres via cross-timbre emotion decoding. Within each region identified in the first analysis, we trained a classifier based on the fMRI pattern measured in the whole region while the subjects listened to stimuli of a given timbre and then measured its generalization power on data recorded when stimuli in other timbres were presented to the subject. This included six different analyses corresponding to the off-diagonal terms in a 3 × 3 training timbre × testing timbre matrix (cf. 3 (1) train on violin, test on clarinet; (2) train on violin, test on voice; (3) train on clarinet, test on violin; (4) train on clarinet, test on voice; (5) train on voice, test on violin; (6) train on voice, test on clarinet), for each of which a fourfold cross-validation was defined, similar to the first analysis. The cross-timbre decoding accuracy levels obtained for each of the four regions were averaged across subjects, and we tested whether this mean power was significantly above chance level (0.25) using a Wilcoxon signed-rank test to account for the nonnormal distribution of the accuracy scores. Of particular interest here are accuracy values obtained by classification schemes involving only music timbres (four cells of the 3 × 3 matrix) compared with those obtained for classification schemes involving voice (the remaining five cells of the matrix) in order to determine the cost of crossing the music–voice timbre boundary.

In order to locate where the cross-timbre emotion decoding was occurring, we used searchlight mapping, the same technique used for the within-timbre classification, but restricted it to voxels within the two auditory cortex regions for increased statistical power.

Results

Behavioral accuracy

To compare the emotion-recognition accuracy between music and voice, participants performed a four-alternative forced-choice emotion (fear, sadness, happiness, neutral) categorization task after

the fMRI experiment. Emotion classification accuracy was high for each of the three timbres (average across emotions: voice: 97.9%; violin: 82.8%; clarinet: 70.9%). A 3 (timbre) × 4 (emotions) repeated-measures analysis of variance of categorization accuracy scores revealed main effects of timbre ($F(2,38) = 71.23$, $P < 0.001$) and emotion ($F(3,57) = 18.11$, $P < 0.001$) modulated by a timbre × emotion interaction (Greenhouse-Geisser corrected; $F(2.97,56.36) = 20.59$, $P < 0.001$). The main effect of timbre reflects higher overall accuracies for vocal stimuli; the interaction with emotion reflects the lower accuracies obtained for clarinet stimuli, particularly for fear and neutral expressions, which were the only stimulus categories with a recognition accuracy score below 75%.

Within-timbre emotion decoding

We first used a searchlight MVPA approach to identify potential areas supporting within-timbre emotion classification to investigate whether the fMRI signal in local spherical clusters of voxels would allow a support vector machine classifier to yield above-chance classification accuracy at classifying emotions in (test) stimuli of a given timbre, after having been trained with fMRI patterns measured in response to other (training) stimuli of the same timbre (see Methods section above).

The map of group-level, above-chance machine classification accuracy scores for within-timbre classification is shown in Figure 1. It highlights two pairs of symmetrical regions: two large regions of the left and right auditory cortex encompassing most of the STG and anterior STS bilaterally (MNI peak coordinates: left: –63, –19, 4; right: 57, –7, 1) and two small clusters of voxels in the premotor cortex anterior to the upper part of the central sulcus (left: –48, –7, 49; right: 54, –1, 46). Each voxel within these four regions is the center of an 8-mm radius sphere that yielded a distribution of emotion classification scores across subjects and timbres significantly above the 25% chance level.

Cross-timbre emotion decoding

Next, we probed the cross-timbre classification power of the fMRI-based emotion classifiers. For each of the four regions identified by the searchlight within-timbre classification (Fig. 1), new classifiers were trained (this time based on all voxels within the region) with emotional labels from stimuli in one of the three timbres and then tested on stimuli from another timbre. The approach can be thought of as exploring the different cells of a 3 × 3 matrix of trained versus tested timbres: cells on the diagonal represent within-timbre classification, while off-diagonal cells represent across-timbre classification. For each subject, we computed the emotion classification accuracy for each of the nine cells of the training–testing matrix and used nonparametric statistics to determine whether the distribution of accuracy scores was significantly above chance in each off-diagonal cell (see "Methods" section above). The results are shown in Figure 2. All off-diagonal cells showed significantly above-chance accuracy for each of the four cerebral regions (all P-values < 0.001 except train: voice, test: clarinet $P < 0.007$ in temporal and left premotor regions of interest (ROIs)).

As a crucial test of the cross-timbre generalization of emotion classification, we compared accuracy values for diagonal (within-timbre) versus off-diagonal (across-timbre) cells of the 3 × 3 training–testing timbre matrix. None of the four ROIs showed a significant difference between accuracy values for within- versus across-timbre classification in a two-sample t-test (right temporal: within-timbre = 12.1 ± 1.0% above chance level, across-timbre = 10.7 ± 0.7%, $P = 0.23$; left temporal: within-timbre = 11.3 ± 0.9%, across-timbre = 10.5 ± 0.7%, $P = 0.54$; right premotor: within-timbre = 8.7 ± 1.0%, across-timbre = 7.8 ± 0.6%, $P = 0.46$; left premotor: within-timbre = 10.3 ± 1.1%, across-timbre = 8.7 ± 0.7%, $P = 0.25$).

The comparison of the accuracy values yielded by classification schemes involving only music timbres versus those obtained for classification schemes involving voice revealed lower accuracy values for classification schemes involving voice, a difference that was strongly significant in both auditory cortices (right temporal: music = 14.0 ± 0.8%, voice = 8.9 ± 0.7%, $P = 7.6 \times 10^{-6}$; left temporal: music = 14.0 ± 0.8%, voice = 8.2 ± 0.7%, $P = 2.6 \times 10^{-7}$; right premotor: music = 9.8 ± 0.8%, voice = 6.8 ± 0.6%, $P = 0.006$; left premotor: music = 10.2 ± 0.9%, voice = 8.5 ± 0.9%, $P = 0.17$).

Finally, the searchlight mapping testing of the six off-diagonal training–testing schemes for the cross-timbre classification yielded significant results. Figure 3 shows the results in the form of maps of auditory cortex voxels with above-chance accuracy at the group level (determined using nonparametric

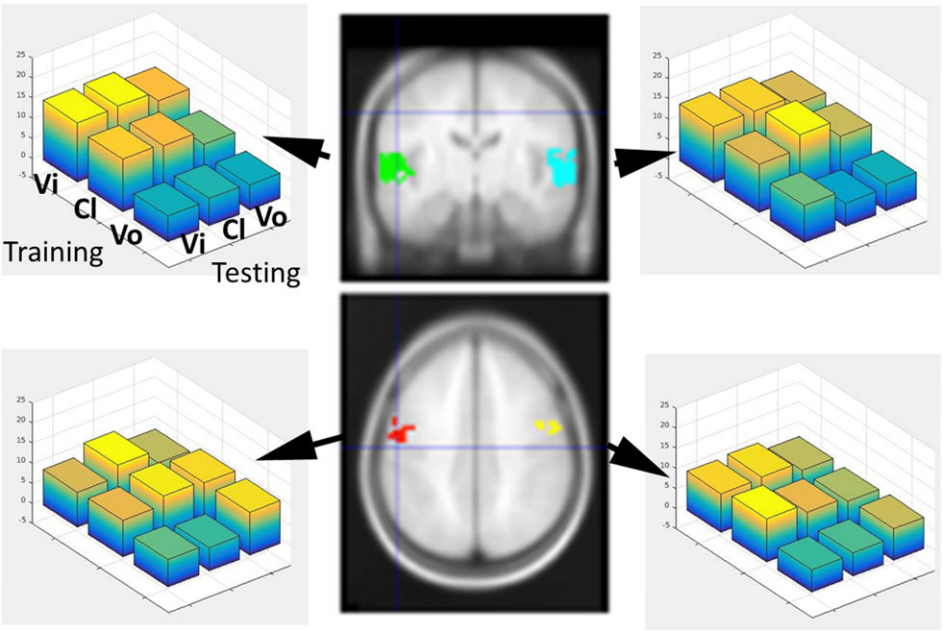

Figure 2. Within- and across-timbre classification accuracy in auditory and premotor regions. For each of the four regions identified in the within-timbre classification analysis (green: left auditory cortex; blue: right auditory cortex; red: left premotor cortex; yellow: right premotor cortex), the figure shows group-average emotion classification accuracies for each cell of the 3 × 3 training–testing timbre matrix (3D plots, in % above-chance level). Within-timbre classification performance (upper left–lower right diagonals) is indistinguishable from across-timbre classification performance (off-diagonal cells), between 5% and 15% above-chance in all cases. VI, violin timbre stimuli; Cl, clarinet timbre stimuli; Vo, voice timbre stimuli.

statistics) for each of the six off-diagonal schemes. All maps show significant ($P < 0.05$ FWE) voxels, although in different numbers and spatial distribution (Fig. 3).

Discussion

We show that fMRI patterns of cerebral activity measured during auditory stimulation with vocal or musical (violin, clarinet) emotional sounds can be used to decode emotions in novel stimuli—even from a different category. In auditory cortex and premotor regions in both hemispheres, classifiers trained at categorizing emotions based on local fMRI patterns measured for voice stimuli were also successful at categorizing emotions in fMRI patterns measured for violin/clarinet stimuli, and vice versa. These results constitute compelling evidence for a shared neural code for auditory emotion processing across different timbres (violin, clarinet, voice).

Within-timbre emotion classification

Large regions of the primary and secondary auditory cortex extending to para-belt STS regions enabled within-timbre emotion classification: in these regions, classifiers trained with local fMRI signal measured in response to emotional stimuli in one timbre were able to successfully generalize performance to fMRI patterns measured in response to other stimuli from the same timbre. This likely reflects the processing of acoustical differences underlying different perceived emotions, relatively consistent across stimuli within a given timbre category. For vocal sounds, the findings are in line with prior studies that used MVPA to decode emotion category from fMRI patterns in the voice-selective TVAs[25] or at the whole-brain level.[26] Ethofer *et al.* obtained above-chance classification of emotional category based on fMRI measures of cerebral activity in the TVAs in areas included in the present areas of auditory cortex.[25] Kotz *et al.* used a whole-brain searchlight mapping strategy similar to the one used here and also found that vocal emotional category could be decoded above chance from fMRI patterns in regions of the anterior and posterior STS bilaterally.[26] They also found that emotion category could be decoded from a number of extratemporal

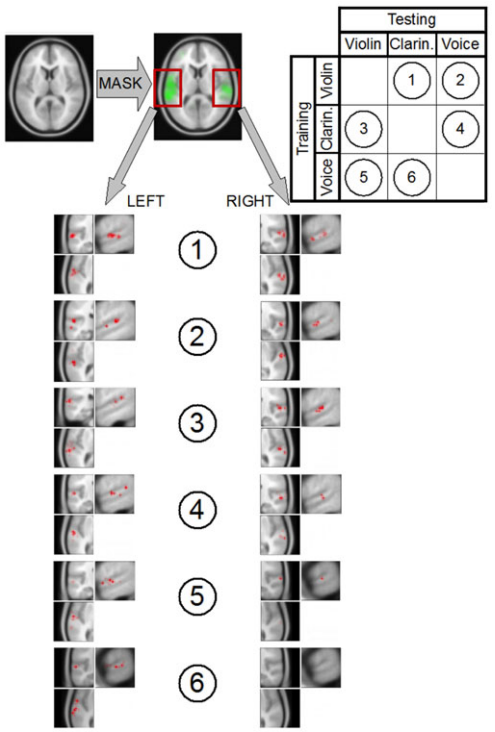

Figure 3. Across-timber searchlight mapping in auditory areas. Number panels show slices through the auditory cortex in the three orientations (left panels: left hemisphere; right panels: right hemisphere), overlaid on an anatomical T1-weighted image, with voxels corresponding to significantly ($P < 0.05$ FWE) above-chance emotion classification accuracy for the cell of the 3 × 3 training–testing timbre matrix corresponding to the circled number. Significant voxels (searchlight sphere centers) are present for all cross-classification schemes except for training voice–testing clarinet (6), which only yielded significant voxels in the left hemisphere.

regions, particularly in the right inferior prefrontal cortex, which was not observed here. This difference may be ascribed to task demands. The participants in Kotz *et al.*'s study were engaged in an active emotion-discrimination task, while here they performed a 1-back task that did not involve any explicit emotional component. More unexpected was the involvement of the bilateral superior premotor areas, which were not found in prior MVPA studies. Premotor involvement might be related to the involvement of vocal movements in the perception of emotional vocalizations.[27]

Across-timbre emotion classification

The most striking result of the present study is that multivariate emotion decoders based on fMRI activity in the auditory cortex were effective across timbres: not only did classifiers trained on fMRI responses to stimuli from one timbre generalize performance to other stimuli of the same timbre, they also generalized to stimuli from different timbres. In fact, it did so to the extent that there was no significant performance difference between within- and across-timbre emotion classification accuracies in any of the four ROIs considered. This is clearly illustrated in Figure 2, where diagonal and off-diagonal cells have comparable levels of above-chance percent accuracy in all four ROIs. Thus, the information code for emotion contained in fMRI patterns in these regions operates essentially independently of timbre, since there is no perceptible accuracy loss in shifting timbres between training and testing schemes when all three timbres are considered.

However, classifying schemes involving voice either as training or testing timbre and an instrument (violin or clarinet) yielded significantly lower accuracies than instrument-only classifying schemes in bilateral auditory cortices, revealing the cost (≈5%) of crossing the music–voice timbre boundary. Perhaps the activity observed in the auditory cortices for violin and clarinet stimuli were more similar than the one observed for vocal stimuli, leading to higher cross-classification accuracies between these two musical timbres than with voice.

This loss of accuracy could be unintentionally linked to our stimuli choices. There is perhaps greater variability in emotion production across 10 actors (10 vocal tracts) compared with the sounds produced by different violins or clarinets. Alternatively, this slight cost of crossing boundaries between domains may reflect a faster or more widespread neural response to the voice than to musical instrument sounds due to the evolutionary significance of the voice.

Although the musical bursts were created to be highly similar to the vocal bursts, some acoustical differences may remain.[5–7,23,28,29] The next logical step lies in using brain imaging and artificial stimuli (or a larger sample of stimuli[28]) in a protocol where emotional acoustical cues can be manipulated (or can vary naturally) in vocal and nonvocal emotions expressions.

Overall, these results constitute strong evidence that the cerebral code for emotion in sound operates similarly across voice and other musical timbres (violin, clarinet) in the auditory cortex. This

finding is in excellent agreement with a recent study from the field of computational voice paralinguistics that showed good cross-domain generalization performance for deep-learning algorithms trained on emotion categorization on multiple audio sources from one domain and then tested on audio stimuli from the other domain.[28] Together, the results provide compelling support for the notion of a universal acoustic code for auditory emotion across different timbres (violin, clarinet, voice).

Acknowledgment

We thank Frances Crabbe and Marc Becirspahic for their help with data collection. S.P. was supported by a European Union Erasmus Mundus mobility fellowship in Auditory Cognitive Neuroscience and a fellowship from the Canadian Institutes of Health Research. P.B. was supported by Grants BBJ003654/1 and BB/1006494/1 from the British Biotechnology and Biological Sciences Research Council, Grant AJE201214 from the French Fondation pour la Recherche Médicale, and Grants ANR-16-CONV-0002 (Institute for Language, Communication and the Brain) and ANR-11-LABX-0036 (Brain and Language Research Institute) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). S.P., P.B., and I.P. developed the project. S.P. collected the data. S.P., S.T., and S.S. analyzed the data. S.T. and S.S. prepared the figures. All authors were involved in writing, editing, and reviewing the manuscript.

Competing interests

The authors declare no competing financial interests.

References

1. Belin, P., S. Fillion-Bilodeau & F. Gosselin. 2008. The Montreal Affective Voices: a validated set of nonverbal affect bursts for research on auditory affective processing. *Behav. Res. Methods* **40:** 531–539.
2. Scherer, K. 1994. Affect bursts: essays on emotion theory. In *Emotions: Essays on Emotion Theory*. S.H.M. van Goozen, N.E. van de Poll & J.A. Sergeant, Eds.: 161–193. Hillsdale, NJ: Lawrence Erlbaum.
3. Bigand, E., S. Filipic & P. Lalitte. 2005. The time course of emotional responses to music. *Ann. N.Y. Acad. Sci.* **1060:** 429–437.
4. Peretz, I., L. Gagnon & B. Bouchard. 1998. Music and emotion: perceptual determinants, immediacy, and isolation after brain damage. *Cognition* **68:** 111–141.
5. Curtis, M.E. & J.J. Bharucha. 2010. The minor third communicates sadness in speech, mirroring its use in music. *Emotion* **10:** 335–348.
6. Juslin, P. & P. Laukka. 2003. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol. Bull.* **129:** 770–814.
7. Bowling, D.D.L., J. Sundararajan, S. Han, *et al.* 2012. Expression of emotion in eastern and western music mirrors vocalization. *PLoS One* **7:** e31942.
8. Peretz, I., D. Vuvan, M.-É. Lagrois, *et al.* 2015. Neural overlap in processing music and speech. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370**. https://doi.org/10.1098/rstb.2014.0090.
9. Spencer, H. 1857. The origin and function of music. *Fraser's Mag.* **56:** 396–408.
10. Dehaene, S. & L. Cohen. 2007. Cultural recycling of cortical maps. *Neuron* **56:** 384–398.
11. Dellacherie, D., D. Hasboun, M. Baulac, *et al.* 2011. Impaired recognition of fear in voices and reduced anxiety after unilateral temporal lobe resection. *Neuropsychologia* **49**: 618–629.
12. Scott, S., A.W. Young, A.J. Calder, *et al.* 1997. Auditory recognition of emotion after amygdalotomy: impairment of fear and anger. *Nature* **385:** 254–257.
13. Sprengelmeyer, R., A.W. Young, U. Schroeder, *et al.* 1999. *Knowing no fear*. **266:** 2451–2456.
14. Gosselin, N., I. Peretz, E. Johnsen, *et al.* 2007. Amygdala damage impairs emotion recognition from music. *Neuropsychologia* **45:** 236–244.
15. Aubé, W., A. Angulo-Perkins, I. Peretz, *et al.* 2015. Fear across the senses: brain responses to music, vocalizations and facial expressions. *Soc. Cogn. Affect. Neurosci.* **10:** 399–407.
16. Escoffier, N., J. Zhong & A. Schirmer. 2013. Emotional expressions in voice and music: same code, same effect? *Hum. Brain Mapp.* **34:** 1796–1810.
17. Belin, P., R.J. Zatorre, P. Lafaille, *et al.* 2000. Voice-selective areas in human auditory cortex. *Nature* **403:** 309–312.
18. Pernet, C.R., P. McAleer, M. Latinus, *et al.* 2015. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage* **119:** 164–174.
19. Agus, T.R., S. Paquette, C. Suied, *et al.* 2017. Voice selectivity in the temporal voice area despite matched low- level acoustic cues. *Sci. Rep.* **7:** 1–7.
20. Ethofer, T., J. Bretscher, M. Geschwind, *et al.* 2012. Emotional voice areas: anatomic location, functional properties, and structural connections revealed by combined fMRI/DTI. *Cereb. Cortex* **22:** 191–200.
21. Norman-Haignere, S., N.G. Kanwisher & J.H. McDermott. 2015. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron* **88:** 1281–1296.
22. Armony, J.L., W. Aubé, A. Angulo-Perkins, *et al.* 2015. The specificity of neural responses to music and their relation to voice processing: an fMRI-adaptation study. *Neurosci. Lett.* **593:** 35–39.
23. Paquette, S., I. Peretz & P. Belin. 2013. The musical emotional bursts: a validated set of musical affect bursts to investigate auditory affective processing. *Front. Psychol.* **4:** 509.

24. Haxby, J. V., A.C. Connolly & J.S. Guntupalli. 2014. Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci.* **37:** 435–456.
25. Ethofer, T., D. Van De Ville, K. Scherer, *et al.* 2009. Decoding of emotional information in voice-sensitive cortices. *Curr. Biol.* **19:** 1028–1033.
26. Kotz, S., C. Kalberlah & J. Bahlmann. 2013. Predicting vocal emotion expressions from the human brain. *Hum. Brain* **34:** 1971–1981.
27. Warren, J.E., D.A. Sauter, F. Eisner, *et al.* 2006. Positive emotions preferentially engage an auditory-motor "mirror" system. *J. Neurosci.* **26:** 13067–75.
28. Coutinho, E. & B. Schuller. 2017. Shared acoustic codes underlie emotional communication in music and speech—evidence from deep transfer learning. *PLoS One* **12:** e0179289.
29. Ma, W. & W.F. Thompson. 2015. Human emotions track changes in the acoustic environment. *Proc. Natl. Acad. Sci. USA* **112:** 14563–14568.