Population codes representing musical timbre for high-level fMRI categorization of music genres

Michael Casey[1], Jessica Thompson[1], Olivia Kang[2], Rajeev Raizada[3], and Thalia Wheatley[2]

[1] Bregman Music and Auditory Research Studio, Department of Music
[2] Wheatley Lab, Department of Psychological and Brain Sciences
Dartmouth College, Hanover, NH 03755, USA
[3] Raizada Lab, Department of Human Development
Cornell University, Ithaca, NY 14853, USA
Michael.A.Casey@dartmouth.edu, Jessica.Thompson.GR@dartmouth.edu,
Olivia.E.Kang@dartmouth.edu, raizada@cornell.edu,
Thalia.P.Wheatley@dartmouth.edu

Abstract. We present experimental evidence in support of distributed neural codes for timbre that are implicated in discrimination of musical styles. We used functional magnetic resonance imaging (fMRI) in humans and multivariate pattern analysis (MVPA) to identify activation patterns that encode the perception of rich music audio stimuli from five different musical styles. We show that musical styles can be automatically classified from population codes in bilateral superior temporal sulcus (STS). To investigate the possible link between the acoustic features of the auditory stimuli and neural population codes in STS, we conducted a representational similarity analysis and a multivariate regression-retrieval task. We found that the similarity structure of timbral features of our stimuli resembled the similarity structure of the STS more than any other type of acoustic feature. We also found that a regression model trained on timbral features outperformed models trained on other types of audio features. Our results show that human brain responses to complex, natural music can be differentiated by timbral audio features, emphasizing the importance of timbre in auditory perception.

Keywords: music, timbre code, STS, multivariate analysis, cepstrum

1 Introduction

Multivariate statistical methods are becoming increasingly popular in neuroimaging analysis. It has been shown that multivariate pattern analysis (MVPA) can reveal information that is undetectable by conventional univariate methods [1]. Much of the work using this approach has focused on the encoding of visual perceptual experiences. Only very recently have researchers begun to apply these methods to the auditory domain, then generally employing only simple stimuli such as isolated tones and monophonic melodic phrases. By contrast, we investigate the neural codes of rich auditory stimuli: real-world commercial music recordings, which contain multiple parallel and complex streams of acoustic information that are distributed in frequency and time.

Recent studies have used MVPA to discriminate neurological responses to several different categories of sound. In one fMRI study, subjects were presented with sounds of cats, female singers, and acoustic guitars. Using MVPA, the authors found that this sound category information could be attributed to spatially distributed areas over the superior temporal cortices [2]. The activation patterns that encode the perceptual interpretation of physically identical but ambiguous phonemes were investigated using MVPA. It was shown that these subjective perceptual interpretations were retrievable from fMRI measurements of brain activity in the superior temporal cortex [3]. Whole-brain MVPA methods were used to identify regions in which the local pattern of activity accurately discriminated between ascending and descending melodies. Three distinct areas of interest were revealed: the right superior temporal sulcus, the left inferior parietal lobule, and the anterior cingulate cortex. These results are in-line with previous studies that found the right superior temporal sulcus to be implicated in melodic processing [4]. Overall, these studies show that MVPA can be used to determine how mental representations of sound categories can be mapped to patterns of neural activity.

Timbre is how sound is described independent of its loudness and pitch, corresponding to the identifiable properties of a sound that remain invariant under those transformations. Timbre is one of the primary cues by which humans discriminate sounds. However, the neural correlates of timbre perception have been severely under studied compared to other aspects of sound like pitch and location. Much of the limited previous work has focused on the lateralization of timbre perception. In an early study on this topic, patients with right- but not left-sided temporal lesions were impaired on a timbre discrimination task [5][6]. Subsequent studies have further described this asymmetry in terms of the types of cues involved in the timbre discrimination task. In a series of studies by Sampson and colleagues, only the right temporal cortex was implicated in tasks that involved onset dynamics and spectral timbre [7], but both temporal cortices were implicated when tones were presented in the context of a melody [8]. Menon and colleagues investigated the neural correlates of timbre using melodies that differed in attack time, spectral centroid, and spectral flux. They found left temporal cortex activations were significantly more posterior than right temporal cortex activations, suggesting a functional asymmetry in their respective contributions to timbre processing [9]. Although these results clearly demonstrate the importance of both temporal cortices in timbre discrimination, the precise neural organization of timbre perception is largely unknown.

2 Materials and Methods

To further investigate the neural encoding of sound category information, we designed an experiment using twenty five natural music stimuli equally divided into five different musical styles: (1) Ambient, (2) 50s RocknRoll, (3) Heavy Metal, (4) Symphonic, and (5) Roots Country. Audio was procured as 44.1kHz, stereo, high-quality AAC 192kbps files. We extracted six-second excerpts from the center of each file, edited to start synchronously with the metrical grid—i.e. on a down beat, if one existed. Excerpts were normalized so that their RMS values were equal, and a 50ms quarter-sine ramp was applied at the start and end of each excerpt to suppress transients.

Participants consisted of 6 females, and 9 males, ages 18-25, who had varying levels of musical expertise. We used a Philips 3T scanner with 32-channel head coil and Lumina button box with one fiber-optic response pad and four colored push buttons. The field of view was 240×240mm with 3mm voxels corresponding to an 80×80 matrix ($240/3 = 80$) for 35 axial slices, thus yielding 224,000 voxels per volume. The scanner repetition rate (TR) was 2000ms. We collected data in 8 runs, each presenting all 25 stimuli in exhaustive category pairings. Category ordering was balanced using maximum length sequences (MLS) to optimally mitigate order effects [10]. Stimuli presentations were interleaved with fixation tasks that ranged from 4-8 seconds. At four randomized intervals per run, an attention probe question appeared on the screen that asked whether the preceding audio clip contained a particular musical feature (e.g., electric guitar). Subjects responded “yes” or “no” to these questions via the response pad. These trials helped to ensure that subjects attended to the music across trials. Data from these trials were discarded from the analyses.

Functional and anatomical images were preprocessed using the AFNI tool chain [11]. As the voxels are not collected concurrently, a timing correction procedure was used to align voxel response functions in time. Volumes were motion corrected to align to the anatomical image. Transient spikes in the signal were suppressed with the AFNI program *3dDespike*. Head motion was included as a regressor to account for signal changes due to motion artifact and linear trends were removed. Data were then smoothed with a 4 mm full width at half maximum (FWHM) smoothing kernel. The image data was further processed by applying per-subject anatomical masks of the STS, which has previously been implicated in sound category discrimination [2, 4, 12]. STS masks were defined manually based on individual subject-specific anatomical landmarks. The data were converted to event related data sets by mapping the volumes to high-dimensional vectors, detrending and zscoring using the rest conditions, then extracting only the data corresponding to stimulus presentations. This yielded 25 stimuli × 8 runs = 200 feature vectors per subject. Singular value decomposition (SVD) was performed on the data to further reduce the dimensionality.

3 Multivariate Analysis

3.1 Musical Category Classification in Bilateral STS

Bilateral STS-masked fMRI data were classified into the five different musical categories using a linear support vector machine (SVM) classifier using within-subject, leave-one-run-out cross validation to evaluate the classification results.

Data were SVD-reduced using the training data to compute a basis for each trial. The subject-mean classifier confusion and standard-error matrix is shown in Table 1. The mean classification accuracy was 0.60, with ±0.03 standard error, which was significantly above the baseline (0.20). Percussive categories (Rock and Roll, Country, and Heavy Metal) were more likely to be confused with one another whereas Ambient was most likely to be confused with Classical and vice versa. The non-percussive categories (Ambient and Classical) were more accurately classified (0.76 mean accuracy, ±0.04 standard error) than the percussive categories (0.5 mean accuracy, ±0.05 standard error). This difference between percussive and non-percussive accuracies was not explained by sampling bias or event density. Percussive and non-percussive confusions are shown in **bold** column-wise.

Table 1. Bilateral STS classifier confusion and standard error

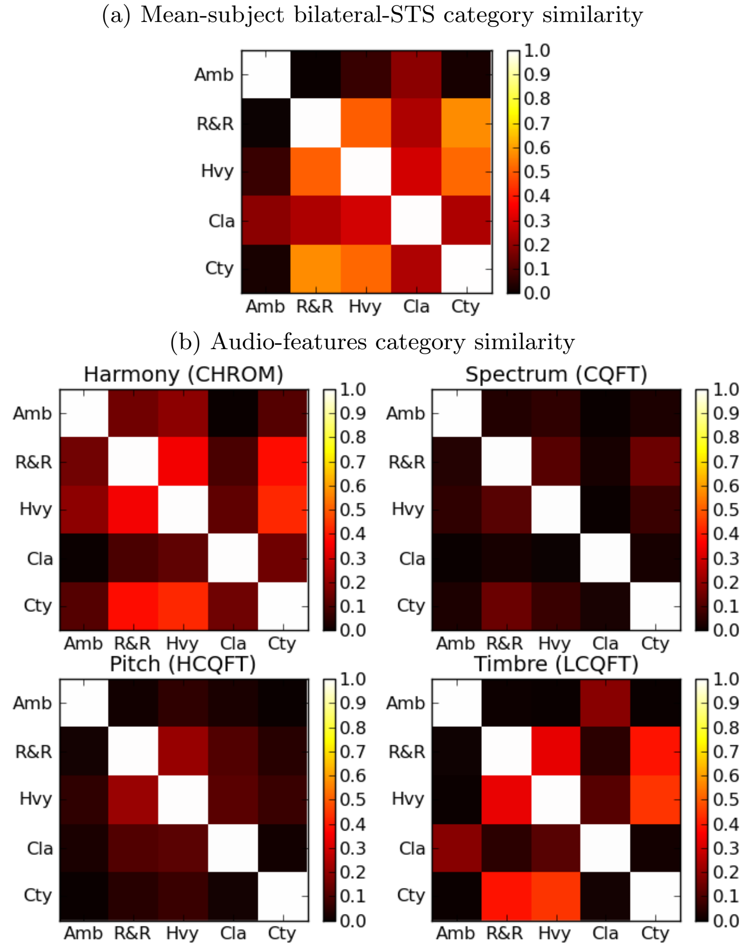
Category:	Amb	RR	Hvy	Cla	Cty
Predicted:					
Amb	**0.78** ±0.05	0.00 ±0.00	0.01 ±0.01	**0.12** ±0.02	0.00 ±0.00
RR	0.00 ±0.00	**0.49** ±0.04	**0.23** ±0.04	0.01 ±0.01	**0.30** ±0.05
Hvy	0.01 ±0.01	**0.14** ±0.03	**0.51** ±0.04	0.13 ±0.03	**0.21** ±0.03
Cla	**0.21** ±0.04	0.01 ±0.05	0.08 ±0.02	**0.74** ±0.04	0.02 ±0.01
Cty	0.01 ±0.01	**0.36** ±0.05	**0.18** ±0.03	0.01 ±0.01	**0.48** ±0.06

Amb=Ambient, RR=Rock & Roll, Hvy=Heavy Metal, Cla=Classical, Cty=Country.

3.2 Representational Similarity Analysis

We sought to verify, by similarity analysis of musical features of the audio, that the observed confusions were due to timbre and not other musical representations such as pitch, or harmony. Representational similarity analysis (RSA) has successfully been employed in previous studies to inspect cross-subject, and cross-species, neural representational spaces [13][14]. We used RSA to determine the similarity relationships between a set of candidate musical features, extracted from the audio stimuli, and the corresponding fMRI images. The mean per-category image over 8 runs was used to compute a per-subject similarity matrix. The mean subject similarity matrix, shown in Figure 1(a), was compared with per-category similarity matrices computed for the four audio features, each representing a different musical facet, see Figure 1(b).

Audio Feature Extraction We extracted audio features using the short-time Fourier transform, with 372ms analysis window advanced in 100ms hops (10Hz). Four feature sets were computed for each stimulus using the *Bregman* Toolkit [15]: (1) pitch-chroma-profiles (CHROM), 12-dimensional vectors representing the total energy attributed to each pitch folded into one octave and roughly corresponding to the harmony, or chord content, of musical stimuli [16];



(a) Mean-subject bilateral-STS category similarity

(b) Audio-features category similarity

Fig. 1. Representational similarity analysis of: (a) per-category means of subjects' images and (b) per-category means of audio features, showing a resemblance between fMRI image similarity and audio similarity for timbre (LCQFT) features.

(2) constant-Q Fourier transform (CQFT), perceptual frequency-warped Fourier spectra corresponding to a human-auditory model of frequency sensitivity and selectivity [17]; (3) high-pass constant-Q cepstral coefficients, extracted from the constant-Q Fourier transform and corresponding to fine-scale perceptual pitch-frequency and pitch-height information (HCQFT) [16]; and (4) low cepstral coefficients computed from the constant-Q Fourier transform (LCQFT) corresponding to timbre, i.e. the *way the stimulus sounds* [18]. The features were labeled by their associated stimulus category (1-5) and further processed by computing the category-mean vectors.

RSA results Figure 1 shows the average-subject between-category image similarity matrix and the between-category similarity matrices obtained using each of the four audio features. We computed the correlation coefficient between the image and audio feature similarity matrices. The highest correlation coefficient was achieved for the timbre features (LCQFT) with a coefficient of 0.99. To compute the significance of the result, and the robustness of the audio features to different temporal treatments, we further processed the features by 16 different temporal regularization algorithms: that is, combinations of mean vector in time, covariance matrix over time, vector stacking in time, per image-duration averaging ($3 \times 2s$ blocks verses $1 \times 6s$ block), and adding backward differences in time for derivatives. The timbre (LCQFT) set of regularized features had the highest mean correlation, 0.99, with $p < 1.06^{-7}$ using a one-way ANOVA.

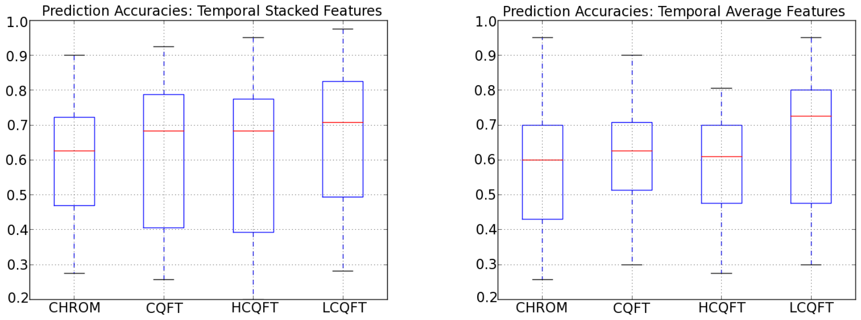
Overall we found that the similarity structure of our neurological data resembles the similarity structure of our timbre feature (LCQFT) more than any other feature. This supports our hypothesis that timbre, how sound is described independent of its loudness and pitch, is most important for the discrimination of different musical categories.

3.3 Multivariate Multiple Regression

It is natural to ask how accurately the image can be predicted by the auditory features of the stimulus. To this end, we performed a binary retrieval task using multivariate multiple regression between our four sets of audio features and the per-subject neural image data. A similar paradigm was used for a language study in [19] predicting neural images corresponding to different categories of visually-presented nouns. The audio features described in Section 3.2 were used for the regression-retrieval task.

For each stimulus as a target, holding out one run for testing, we chose a decoy stimulus from another category. The remaining runs were used to train a multivariate multiple regression model of the auditory representational space using audio-feature/image-feature pairs. The target and decoy predicted images were computed from their corresponding audio features using the trained regression weights. We evaluated the predictive performance of each audio feature by whether the target's predicted image was closer to the true target image or to the decoy's image. This procedure was repeated exhaustively for all 200 stimulus presentations for each subject.

Regression results Figure 2 shows that timbral features (LCQFT) were most accurate in predicting the image response. This was true for both temporal regularization treatments, with accuracies of 0.71 for temporal-stacking LCQFT features and 0.73 for temporal-averaging LCQFT features. The figure also shows the inter-quartile ranges for each feature set. Temporal stacking improved both auditory spectrum (CQFT) and pitch (HCQFT) features but made no improvement to harmony features (CHROM) or to the timbre result (LCQFT) which performed equally well with and without increased temporal context. This suggests that the auditory representational spaces corresponding to timbre, and also harmony, are more robust to differences of time scale than representations correlating with pitch and spectrum.


(a) Stacked temporal audio features

(b) Mean audio features over time

Fig. 2. Median and inter-quartile regression-prediction accuracies for audio features corresponding to harmony (CHROM), auditory spectrum (CQFT), pitch (HCQFT), and timbre (LCQFT). (a) Temporal context preserved by stacking feature vectors per stimulus. (b) No temporal context: mean over time of feature vectors per stimulus.

4 Conclusions

The discrimination of musical categories in our experiments is due to a timbre population code distributed in bilateral STS. This finding is supported by evidence from classification, similarity, and regression experiments between the audio and neuroimaging domains. Our results expand on previous studies, finding timbral specificity in STS, but our study shows this effect in greater detail and for more complex natural stimuli. Significantly worse results for pitch and spectrum features provided further evidence for a timbral code in our experiments.

Beyond neuroimaging, our results are consistent with computational systems that attempt to solve the same task: namely, high-level music classification, but using audio features alone. In previous studies, for example [18][20], timbral features similar to those used in our study were shown to be effective in the categorization task. Using different stimuli and computational tasks will likely reveal further population codes that are specific to aspects of musical stimuli other than timbre.

References

1. K. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–30, September 2006.

2. N. Staeren, H. Renvall, F. De Martino, R. Goebel, and E. Formisano. Sound categories are represented as distributed patterns in the human auditory cortex. *Current Biology*, 19(6):498–502, March 2009.
3. N. Kilian-Hütten, G. Valente, J. Vroomen, and E. Formisano. Auditory cortex encodes the perceptual interpretation of ambiguous sound. *The Journal of Neuroscience*, 31(5):1715–20, February 2011.
4. Y.-S. Lee, P. Janata, C. Frost, M. Hanke, and R. Granger. Investigation of melodic contour processing in the brain using multivariate pattern-based fMRI. *NeuroImage*, 57(1):293–300, July 2011.
5. S. Samson and R. J. Zatorre. Melodic and harmonic discrimination following unilateral cerebral excision. *Brain and Cognition*, 7(3):348–60, June 1988.
6. J. K. Bizley and K. M. M. Walker. Sensitivity and selectivity of neurons in auditory cortex to the pitch, timbre, and location of sounds. *The Neuroscientist*, 16(4):453–69, August 2010.
7. S. Samson and R. J. Zatorre. Contribution of the right temporal lobe to musical timbre discrimination. *Neuropsychologia*, 32(2):231–40, February 1994.
8. J. D. Warren, A. R. Jennings, and T. D. Griffiths. Analysis of the spectral envelope of sounds by the human brain. *NeuroImage*, 24(4):1052–7, February 2005.
9. M. Meyer, S. Zysset, D. Y. von Cramon, and K. Alter. Distinct fMRI responses to laughter, speech, and sounds along the human peri-sylvian cortex. *Cognitive Brain Research*, 24(2):291–306, July 2005.
10. G. T. Buracas and G. M Boynton. Efficient design of event-related fmri experiments using m-sequences. *NeuroImage*, 16(3):801–813, July 2002.
11. R. W. Cox. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29(3):162–73, June 1996.
12. V. Menon. Neural Correlates of Timbre Change in Harmonic Sounds. *NeuroImage*, 17(4):1742–1754, December 2002.
13. S. J. Hanson, T. Matsuka, and J. V. Haxby. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: Is there a face area? *Neuroimage*, (23):156–166, 2004.
14. N. Kriegeskorte, M. Mur, D. Ruff, P. Kiani, J. Bodurka, and H. Esteky. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, (60):1126–1141, 2008.
15. M. A. Casey. Bregman music and auditory python toolbox. *http://bregman.dartmouth.edu/bregman*, Jan 2012.
16. M. Müller, S. Ewert, and S. Kreuzer. Making chroma features more robust to timbre changes. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 1877–1880. IEEE, 2009.
17. J.C. Brown and M.S. Puckette. An efficient algorithm for the calculation of a constant q transform. *Journal of the Acoustical Society of America*, (92):2698–2701, 1992.
18. B. Logan. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the International Symposium on Music Information Retrieval*, 2000.
19. T.M. Mitchell, S.V. Shinkareva, A. Carlson, K.M. Chang, V.L. Malave, R.A. Mason, and M.A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.
20. G. Tzanetakis and P. R. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.