DECODING DRUMS, INSTRUMENTALS, VOCALS, AND MIXED SOURCES IN MUSIC USING HUMAN BRAIN ACTIVITY WITH FMRI

Vincent K.M. Cheung [1] **Lana Okuma** [2] **Kazuhisa Shibata** [2]
Kosetsu Tsukuda [3] **Masataka Goto** [3] **Shinichi Furuya** [1]
[1] Sony Computer Science Laboratories, Tokyo, Japan
[2] RIKEN Center for Brain Science, Japan
[3] National Institute of Advanced Industrial Science and Technology (AIST), Japan

ABSTRACT

Brain decoding allows the read-out of stimulus and mental content from neural activity, and has been utilised in various neural-driven classification tasks related to the music information retrieval community. However, even the relatively simple task of instrument classification has only been demonstrated for single- or few-note stimuli when decoding from neural data recorded using functional magnetic resonance imaging (fMRI). Here, we show that drums, instrumentals, vocals, and mixed sources of naturalistic musical stimuli can be decoded from single-trial spatial patterns of auditory cortex activation as recorded using fMRI. Comparing classification based on convolutional neural networks (CNN), random forests (RF), and support vector machines (SVM) further revealed similar neural encoding of vocals and mixed sources, despite vocals being most easily identifiable. These results highlight the prominence of vocal information during music perception, and illustrate the potential of using neural representations towards evaluating music source separation performance and informing future algorithm design.

1. INTRODUCTION

The goal of brain decoding is to infer mental states and perceptual information from neural activity [1, 2]. Common neuroimaging techniques such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) allow data acquisition in a non-invasive manner, which has resulted in rapid developments in brain-computer interfaces (BCI) [3, 4].

Although fMRI- and EEG-based models both make use of neural activity for decoding, the form of information retrieved is substantially different. That is because fMRI offers (sub-)millimetre spatial resolution at the cost of low temporal resolution, whilst EEG provides a millisecond-level temporal resolution at the expense of poor spatial resolution [2, 5]. Consequently, fMRI-based decoders typically rely on spatial representations of neural activation as features, whilst EEG-based decoders exploit the temporal dynamics of neural activity.

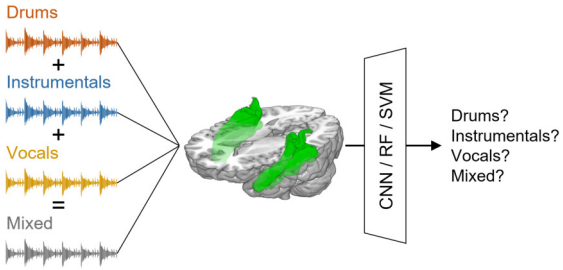

Figure 1. We compared the decoding performance of convolutional neural networks (CNN), random forests (RF), and support vector machines (SVM) in classifying drums, instrumentals, vocals, and mixed naturalistic musical sources based on human auditory cortex activation (highlighted in green) as recorded using fMRI.

In the context of music information retrieval (MIR), both fMRI- and EEG-based decoders have been employed for a variety of classification/estimation tasks, such as genre [6–9], pitch [6, 10–12], rhythm [13, 14], musical emotion classification [15–21], song identification [22–25], music composition [26], beat and note onset detection [27,28], and acoustic feature extraction [29], as well as reconstruction from heard and imagined melodies [30–35].

However, a problem that has remained under-studied is the decoding of different instruments within a song based on brain activity. This is despite its intimate relation to the standard MIR task of music source separation, which seeks to decompose a musical sound mixture into a linear sum of instrumental sources [36, 37]. Although music and speech source separation share the same goals, the key difference is that sound sources from multiple musical instruments are more correlated in music than in speech [36].

The most relevant literature on neural-driven music source separation is the work by Cantisani et al. [38, 39]. Their initial work showed that EEG can be used to decode listeners' attention deployment to a particular instrument

from naturalistic polyphonic music mixtures [38]. The approach was to first record listeners' EEG as they were presented with solo instrumental sources. A temporal response function was then trained to reconstruct the solo instrumental sources from EEG. This response function was later applied to the EEG signal when subjects listened to the polyphonic mixtures. The attended instrument was identified as the one that showed the highest correlation with the reconstructed source. In their subsequent work [39], they showed that the reconstructed sources from their EEG attention decoding model can be used as contrastive priors to inform a non-negative matrix factorisation-based source separation model.

On the other hand, relevant work based on fMRI data seems to be lacking. While existing studies have identified the role of the auditory cortex in processing timbre [40, 41] via correlational approaches, even the relatively simple task of decoding musical instrument category has only been restricted to single- or few-note stimuli [19, 42].

In this paper we address this gap by showing that distinct musical sources, namely drums, instrumentals, and vocals from naturalistic musical stimuli, as well as their mixtures, can be decoded from spatial representations of neural activation recorded using fMRI on the single-trial level. We report that decoding performance was the highest when detecting the presence of vocal information in the auditory stimulus, and we explain our model decisions in terms of patterns of neural activations. Importantly, unlike most existing decoding studies which have relied on a single classification algorithm, we additionally compared performance across three decoders, convolutional neural networks (CNN), random forests (RF), and support vector machines (SVM), to enhance the generalisability of our findings. In the last section of this paper, we also discuss how brain activity could be used in the future to evaluate music source separation and inform algorithm design.

2. METHODS

2.1 Experimental stimuli

Experimental stimuli consisted of 15-second musical audio excerpts derived from the beginning of the chorus section of 24 unreleased pop and rock songs within an in-house music dataset created by professional musicians.

Four versions of each song—*drums*, *instrumentals*, *vocals*, and *mixed*—were compiled, resulting in a total of 96 stimuli. The versions were produced by first separating the original song into `bass`, `drums`, `other`, and `vocals` using a state-of-the-art music source separation model, Demucs-v4 [43]. Due to the frequency response of MRI-compatible noise-isolating earphones (Sensimetrics S15), `bass` and `other` were linearly combined to form an *instrumentals* version. A 100-ms fade-out was then applied to the *drums*, *instrumentals*, and *vocals* versions, followed by loudness normalisation to the EBR U 128 standard. Finally, the normalised *drums*, *instrumentals*, and *vocals* versions of each song were linearly combined to form the *mixed* version, which was also normalised for loudness. We chose to use the *mixed* version rather than the original song to ensure that decoding was not biased by differences in loudness from the underlying versions. We made sure that each song actually included drums, vocals, and other instruments before source separation, and we checked our resulting stimuli after source separation to ensure that they were free from audible artefacts and separation errors, and that they did not contain silences at the start and end that would shorten the stimuli.

2.2 Data acquisition

Data were collected from 24 healthy, normal-hearing adults aged between 19-34 with their written informed consent. The 96 music stimuli were presented over eight runs whilst functional gradient echo planar images (TR/TA/TE = 2/2/0.025 s, voxel size = 3×3×3 mm^3, 33 slices, flip angle = 77°, 188 volumes per run) were acquired using a Siemens Prisma 3T MRI scanner. Each run lasted approximately 6 minutes, and was separated by a short break of around one minute. Stimulus presentation was counter-balanced across runs, with the constraint that each run contained three samples of the four versions, all stimuli came from different songs, and that each song (regardless of version) appeared only once every other run. Stimulus presentation within a run was randomised. To maintain attention, subjects were also asked to rate their preference on a 1-9 scale within a 4-second time window using a button box after each stimulus presentation. Our study was approved by the Ethics Committee at RIKEN.

2.3 fMRI data preprocessing

Functional MRI data for each subject were preprocessed using fMRIprep [44]. Functional images were first corrected for slice-timing differences, motion artefacts, and susceptibility distortions, then co-registered to subjects' anatomical image, and then normalised to standard MNI-space using the ICBM 152 Nonlinear Asymmetrical template. Next, for each subject, we fitted a general linear model in each voxel to estimate the blood oxygen level-dependent (BOLD) response on the single-trial level using SPM [45] following a 'least-squared all' approach [46]: each stimulus was modelled as a separate regressor in the design matrix, and a parametric modulator that varied by subjects' stimulus rating was also added to control for differences in preference. Another regressor was included to account for variance during the rating period. These regressors were modelled as boxcar functions and convolved with the canonical haemodynamic response function (HRF). Six motion, one cardiac, and one respiratory regressors were further added to the design matrix to control for motion- and physiology-induced artefacts. Model parameters were estimated using restricted maximum likelihood, and the resulting parameter estimates at each voxel provided a spatial representation (i.e., *beta maps*) of neural activations for each stimulus separately, which we used for subsequent decoding.

As we were interested in stimulus differences in the neural-perceptual level, we considered voxels in the hu-

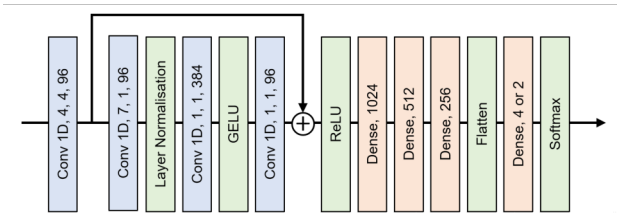

Figure 2. Architecture of our CNN-based decoder.

man auditory cortex (see Figure 1) as decoding features. These were obtained by applying a mask to the bilateral `early-auditory` and `auditory-associative` regions in the HCP-MM1 brain atlas [47,48], and then flattened into a 1D-vector using nilearn [49].

2.4 Decoding analyses

We performed two decoding analyses. The first was a four-way classification task, whose goal was to classify which of the four versions a stimulus belonged to based on subjects' brain activation as summarised by its beta map. The second was a binary recognition task, whose goal was to detect the presence of drums, instrumentals, or vocals in the stimulus from brain activation. As an example, for drum-recognition, *drums* and *mixed* versions would be assigned a positive label, whilst *instrumentals* and *vocals* versions would be assigned a negative label.

We also examined whether decoding performance depended on neural information encoded in the left, right, or both auditory cortices. This was motivated by neuroscientific findings suggesting a right-lateralised hemispheric dominance to musical stimuli [50], and that the left auditory cortex may be more sensitive to rapid temporal features in an auditory stimulus whilst the right may be more sensitive towards spectral features [51].

To enhance the generalisability of our findings, we performed leave-one-subject-out cross-validation, where each decoder was trained on data from 23 subjects and tested on 1 remaining subject. Note that brain decoding between subjects is generally harder than decoding within subjects, because the decoder must additionally overcome individual differences in structural and functional organisation of the brain when predicting on an unseen subject [52].

2.5 Implementation

We trained three types of classifiers—convolutional neural networks (CNN), random forests (RF), and support vector machines (SVM)—for our two decoding tasks. While classical approaches such as SVMs and RFs remain popular [53], CNNs have also been recently used to decode visual objects [54], vocal emotions [55], and musical pitch [10] from fMRI data. We implemented CNN decoders on TensorFlow2, whilst RF and SVM were implemented on scikit-learn [1].

Training data were first put through a variance threshold to remove features that gave identical outputs (e.g., at the boundary of the brain), and then scaled using a robust scaler, before decoders were fitted.

Our CNN decoders (see Figure 2) were inspired by ConvNeXt [56], which is a family of purely convolutional neural networks that recently achieved state-of-the-art performance in image classification. Input features first passed through a 1D-convolution layer (96 units, kernel size = 4), and a ConvNeXt-like residual block. This block comprised a 1D-convolution layer (96 units, kernel size = 7), followed by layer normalisation, 1D-convolution (384 units, kernel size = 1), GELU activation, another 1D-convolution (96 units, kernel size = 1), and a residual connection layer followed by ReLU activation. Outputs of the residual block then passed through three dense layers (1024, 512, and 256 units, respectively), a flattening layer, and finally a dense layer with softmax output. All convolution layers had a stride length of 1 (except for the first, which had a length of 4) and `same`-padding. Each model was trained to minimise categorical entropy loss for 200 epochs, and early-stopped if validation performance did not improve after 25 epochs (with best weights restored). Data from two random subjects ($\sim$10%) in the training set were held-out for validation, and we selected a batch size of 512, and an AdamW optimiser [57] with learning rate = 0.001 and weight decay = 0.0001 for training.

RF decoders were trained with bootstrapping using 100 trees in the forest, at least 1 sample per leaf, and 2 samples per split. Quality of split was assessed with Gini impurity.

SVM decoders were trained using regularisation parameter of 1 with squared-L2 penalty and a linear kernel for a maximum of 10,000 iterations.

3. RESULTS AND DISCUSSION

3.1 Four-way classification

Table 1 and Figure 3 show the leave-one-subject-out cross-validation performance of CNN, RF, and SVM decoders in classifying whether a stimulus belonged to *drums*, *instrumentals*, *vocals*, or *mixed* versions of a song, given auditory cortex activation. To test the statistical significance of our results (see Table 2), we fitted linear mixed models with the interaction between classifier and hemisphere (and lower order terms) as fixed effects and a maximal random effects structure with subject as a grouping factor.

All classifiers showed significantly above-chance performance (all $p < 2.2\times10^{-16}$, see Supplementary Information [1]) in decoding accuracy and Area Under the Receiver Operating Characteristic Curve (ROC AUC), suggesting that despite the slow temporal resolution of fMRI, correlated sources of the same song can be decoded and classified from spatial representations of brain activation.

For all classifiers, accuracy and ROC AUC were highest when decoding from the bilateral auditory cortex, followed by the right and left hemispheres. Resolving significant interaction effects between classifier and hemisphere for accuracy and ROC AUC furthermore revealed that accuracy was significantly worse when decoding from the left compared to the right and bilateral auditory cortices for CNN,

[1] For data, code, and Supplementary Information, please refer to `https://github.com/vkmcheung/neuromusic-decoding/`

	CNN		RF		SVM	
	acc	**auc**	**acc**	**auc**	**acc**	**auc**
Four-way classification						
l AC	.506	.799	.523	.802	.524	.810
r AC	.588	.858	.536	.817	.563	.843
l+r AC	**.604**	.863	.554	.824	.597	**.863**
l+r PV	.301	.560	.319	.554	.253	.510
l+r SM	.304	.547	.332	.550	.276	.554
Drums recognition						
l AC	.595	.638	.603	.637	.550	.559
r AC	.622	.677	.586	.638	.559	.591
l+r AC	**.630**	**.683**	.599	.655	.553	.601
l+r PV	.507	.505	.528	.533	.526	.531
l+r SM	.517	.544	.530	.545	.490	.500
Instrumentals recognition						
l AC	.656	.726	.638	.688	.615	.679
r AC	.642	.723	.666	.727	.627	.687
l+r AC	**.680**	**.762**	.657	.712	.652	.703
l+r PV	.577	.593	.585	.611	.495	.509
l+r SM	.558	.576	.580	.600	.517	.553
Vocals recognition						
l AC	.794	.891	.799	.913	.746	.847
r AC	.816	.926	.841	.937	.836	.936
l+r AC	.839	.946	.829	.936	**.843**	**.950**
l+r PV	.527	.527	.525	.541	.495	.502
l+r SM	.516	.544	.563	.581	.516	.552

Table 1. Mean brain decoding performance with leave-one-subject-out cross-validation. *acc* = accuracy; *auc* = ROC AUC; *l/r/l+r* = left/right/bilateral; *AC* = auditory, *PV* = primary visual, *SM* = somatosensory-motor cortices.

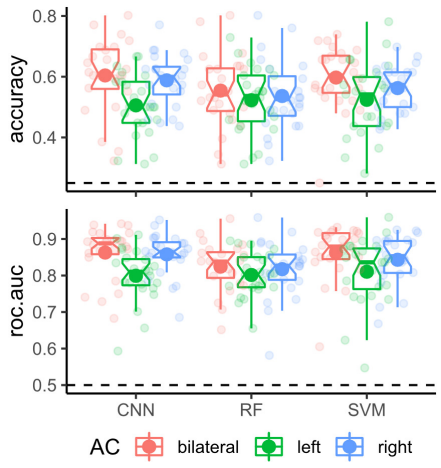

Figure 3. Box plots showing four-way classification performance when decoding from voxels in the left and/or right auditory cortex using CNN, RF, and SVM. Light circles indicate test performance on each held-out subject. Filled circles indicate mean. Dashed lines indicate chance.

and compared to the right for SVM. Likewise, ROC AUC was significantly lower when decoding from the left compared to the bilateral auditory cortex for all classifiers, and compared to the right for CNN and SVM. Although these results corroborate previous findings (e.g., [50]) that support a dominant role of the right auditory cortex in processing musical stimuli, they nevertheless show that both auditory cortices were engaged and provided useful information for decoding.

Four-way classification	χ^2	*df*	*p*
Accuracy			
hemisphere	37.3	2	8.08×10^{-9} ***
classifier	6.81	2	.0331 *
hemisphere:classifier	14.6	4	.00551 **
ROC AUC			
hemisphere	55.3	2	9.58×10^{-13} ***
classifier	11.8	2	.00278 **
hemisphere:classifier	14.3	4	.00625 **
Recognition task	χ^2	*df*	*p*
Accuracy			
hemisphere	3.30	2	.192
classifier	14.5	2	.000720 ***
task	59.9	2	9.78×10^{-14} ***
hemisphere:classifier	3.60	4	.463
task:classifier	9.76	4	.0447 *
hemisphere:task	2.90	4	.574
hemisphere:classifier:task	11.3	8	.184
ROC AUC			
hemisphere	4.89	2	.0866
classifier	13.6	2	.00114 **
task	94.2	2	$< 2.2\times10^{-16}$ ***
hemisphere:classifier	2.07	4	.722
task:classifier	10.9	4	.0275 *
hemisphere:task	2.68	4	.612
hemisphere:classifier:task	9.54	8	.299

Table 2. ANOVA table evaluating the statistical significance of hemisphere and classifier in four-way classification and recognition task performance. * = p < .05; ** = p < .01; *** = p < .001.

Confusion matrices in Figure 4(A) provide further insight on decoding performance. We notice that across the three decoders trained on both hemispheres, recall was the highest for *drums* and the lowest for *mixed*. The high recall for *drums* could be because they were the most temporally regular and had limited pitch possibilities. Furthermore, *mixed* and *vocals*, as well as *drums* and *instrumentals*, were often misclassified as the other. These suggest a similar neural representation between *mixed* and *vocals*, as well as *drums* and *instrumentals*. Whether this pairing is contingent on the stimulus set, the part of a song used (here, our stimulus excerpts were taken from the beginning of the chorus section), and the experimental design remains to be verified in future studies.

3.2 Neural representations

To explain the impact of each voxel towards classification, we turned to SHapley Additive exPlanations (SHAP) [58]. SHAP decomposes a model prediction into the additive contribution of each feature from the mean using game theory. Figure 4(B) shows the mean-averaged contribution of voxels in the bilateral auditory cortex towards classifying a stimulus as belonging to the four versions in Subject 4 using a CNN. We notice that the pattern of contributions were quite similar for *mixed* and *vocals*, which could explain the misclassification of the two labels observed above. Furthermore, there were substantial contributions from both

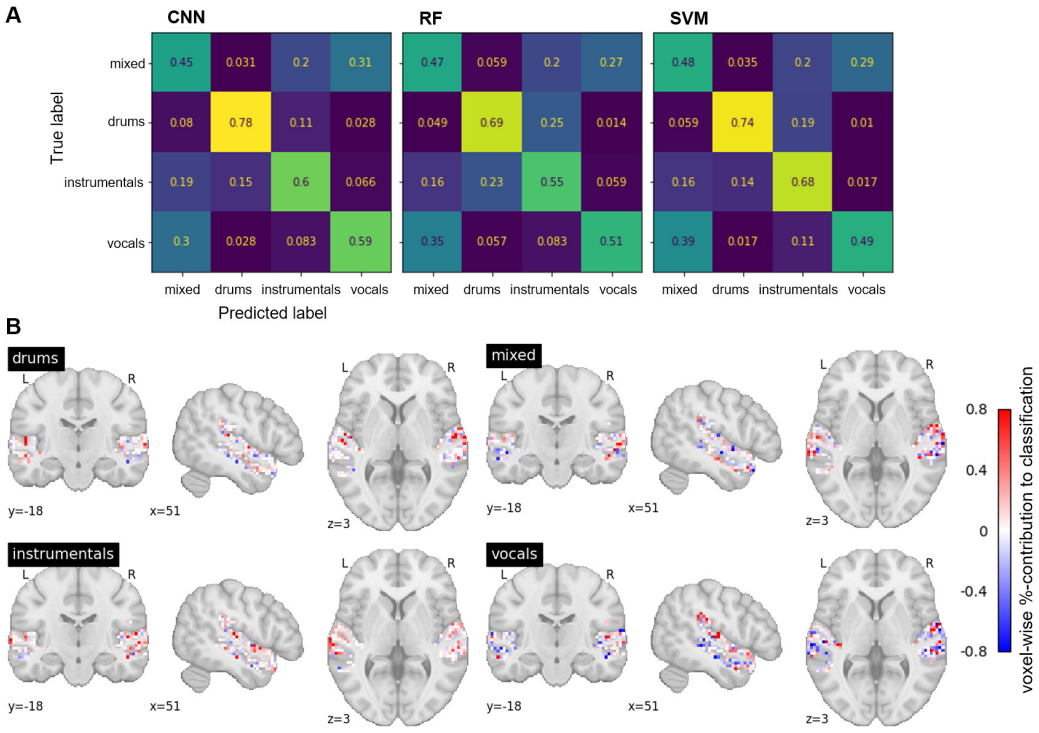

Figure 4. (A) Confusion matrix (normalised along rows) for each decoder when trained on the bilateral auditory cortices, pooled across all subjects. Notice that *drums* recall was highest, and a consistent misclassification between *mixed* and *vocals*, as well as *drums* and *instrumentals*. (B) Mean additive contribution of each voxel in the bilateral auditory cortex towards classifying a given label for one subject using a CNN decoder derived using SHAP [58].

auditory cortices, again indicating a bilateral engagement during music processing.

3.3 Recognition task

We next tested whether decoding performance in recognising the presence of drums, vocals, or instrumentals varied from the left and/or right auditory cortex. Results from leave-one-subject-out cross-validation are summarised in Figure 5 and Tables 1 and 2.

Resolving significant main effect of tasks for accuracy and ROC AUC revealed substantially higher decoding performance across CNN, RF, and SVM in recognising vocals compared to drums and instrumentals (see Supplementary Information [1]). That the presence of vocal information was most robustly encoded from neural activation patterns is very interesting, as it suggests that listeners show an enhanced sensitivity towards perceiving human voice in music. This finding is in line with the view that singing vocals play a prominent and powerful role in communicating and expressing meaning and emotion during music listening [59, 60]. We speculate that the presence of vocal information might have additionally engaged neural populations involved in language processing, which consequently increased its dissimilarity amongst other labels.

Significant main effects of classifier for accuracy and ROC AUC also indicated superior performance of CNN and RF over SVM when averaged across recognition tasks. However, significant task-by-classifier interactions for both measures suggest that performance varied according to recognition task. Resolving the interaction revealed significantly lower accuracy and ROC AUC for SVM compared to CNN and RF in drums recognition. Significantly higher ROC AUC was also observed for CNN compared to SVM in recognising instrumentals. Nevertheless, we were not able to detect any meaningful differences in laterality across tasks or classifiers.

3.4 Feature-encoding specificity

Thus far, we relied on neural activations in the auditory cortex as input features for our decoding models. To assess the specificity of information encoding, we repeated the above analyses in two other sensory processing brain regions, namely the bilateral `primary-visual`, and the `somatosensory-and-motor` regions as derived from the HCP-MM1 brain atlas [47, 48]. As before, we assessed the statistical significance of decoding performance using linear mixed models. However, rather than comparing the effects of hemisphere within the auditory cortex, we now compare performance across the bilateral auditory, primary visual, as well as somatosensory-motor cortices.

In the four-way classification task, we observe in Table 1 and the Supplementary Information [1] that decoding from the bilateral auditory cortex resulted in significantly higher accuracy and ROC AUC compared to the two other sensory cortices across all classifiers (all $p < 2.2 \times 10^{-16}$).

Interestingly, decoding accuracy and ROC AUC were also significantly above chance when CNNs and RFs were trained using features from the visual and somatosensory-motor regions (with no significant differences between these two regions). Furthermore, resolving signifi-

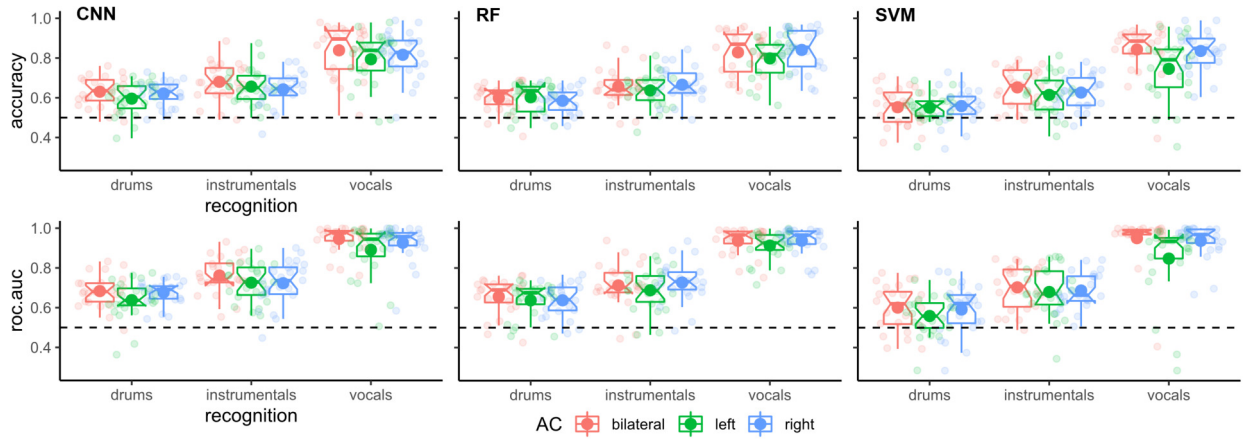

Figure 5. Box plots comparing performance in recognising the presence of *drums*, *instrumentals*, or *vocals* in a musical stimulus using left and/or right auditory cortex activation as decoding features. Significantly higher decoding performance was detected in the *vocals* recognition. Dashed lines indicate chance performance.

cant cortex-by-classifier interactions showed significantly lower accuracy and ROC AUC when decoding from the primary visual cortex using SVM compared to CNN and RF, and from the auditory cortex using RF compared to CNN, as well as lower accuracy when decoding from the somatosensory-motor cortex using SVM compared to RF.

A similar picture could be seen in recognition performance. Significant main effects of cortex for recognition accuracy and ROC AUC indicate superior performance when decoding from the auditory compared to visual or somatosensory-motor regions. Resolving significant cortex-by-task interactions further revealed that the significantly higher performance in recognising vocals compared to drums or instrumentals was specific to the auditory cortex. By contrast, accuracy and ROC AUC for instrumentals were significantly higher than drums in the auditory and somatosensory-motor areas, as well as in the primary visual cortex (ROC AUC only).

Engagement of the primary visual cortex during music has been suggested to be related to mental imagery [61,62], which is thought to be an important way through which music evokes emotions [63]. Likewise, the somatosensory cortex has been said to encode the emotional percept or feeling states associated with music [15], whilst auditory-motor interactions during music perception is thought to be related to the integration and updating of hierarchical predictions of the musical beat [64, 65]. Combined with the substantially higher performance observed when decoding from the auditory cortex, these suggest that while musical sources could be decoded from visual and somatosensory-motor regions, the information encoded is unlikely to be related to the auditory content itself. Rather, such representations might encode affective or metrical information from associated cognitive processes that arise when perceiving the four different musical sources.

4. CONCLUSION AND FUTURE PERSPECTIVES

In this paper, we demonstrated that drums, instrumentals, vocals, and mixed sources of naturalistic music can be decoded from human auditory cortex fMRI data on the single-trial, between-subject level. While decoding performance was the highest for CNN, performance across all classifiers—CNN, RF, and SVM—were above chance and suggested similar neural representations for vocals and mixed sources. An especially high performance in vocals recognition across all classifiers further pointed towards an enhanced perceptual sensitivity towards vocal information during music listening. Taken together, our results show that despite the low temporal resolution of fMRI, the high spatial resolution it offers could still provide relevant information for decoding in neural-driven MIR tasks.

Although our specificity analyses highlighted the auditory cortex in encoding stimulus-relevant information compared to other sensory areas, the perception of different musical sources is a hierarchical process that engages higher-order brain regions in the prefrontal cortex via dorsal and ventral pathways [66, 67]. Future work could examine differences in representations along these two pathways to shed light on neural mechanisms involved in auditory-object processing.

In the context of music source separation, one future possibility is to use neural data for evaluation. While current subjective evaluation of music source separation algorithms typically rely on explicit ratings such as MUSHRA or mean opinion scores, ratings are known to be prone to response biases [68–70] and might consequently fail to adequately reflect subjects' perception. This could be overcome by directly evaluating performance on the neural-perceptual level. Future work could, for example, compare the neural representations of source-separated stimuli from different algorithms or hyperparameters. Separation quality could be determined by identifying the algorithm that maximises dissimilarity in neural activation across the different sources. Another possibility is to assess sensitivity to each instrument by examining neural activation in response to different mixing proportions. This would provide perceptual priors that could be used to constrain the parameter space in future music source separation algorithms. While these prospects may seem too challenging at this time, we envision that our work will help pave the way in that direction.

5. ACKNOWLEDGMENTS

This work was supported in part by JST CREST Grant Number JPMJCR20D4 and JSPS KAKENHI Grant Number JP21H04917, Japan.

6. REFERENCES

[1] N. Kriegeskorte and P. K. Douglas, "Interpreting encoding and decoding models," *Current Opinion in Neurobiology*, vol. 55, pp. 167–179, 2019.

[2] B. Kaneshiro and J. P. Dmochowski, "Neuroimaging methods for music information retrieval: Current findings and future prospects," in *Proceedings of the 16th International Society for Music Information Retrieval Conference*, ser. ISMIR 2015, 2015, pp. 538–544.

[3] M. Zhuang, Q. Wu, F. Wan, and Y. Hu, "State-of-the-art non-invasive brain–computer interface for neural rehabilitation: A review," *Journal of Neurorestoratology*, vol. 8, no. 1, pp. 12–25, 2020.

[4] R. Sitaram, A. Caria, R. Veit, T. Gaber, G. Rota, A. Kuebler, and N. Birbaumer, "fMRI brain-computer interface: a tool for neuroscientific research and treatment," *Computational Intelligence and Neuroscience*, vol. 2007, 2007.

[5] B. He and Z. Liu, "Multimodal functional neuroimaging: Integrating functional MRI and EEG/MEG," *IEEE Reviews in Biomedical Engineering*, vol. 1, pp. 23–40, 2008.

[6] M. A. Casey, "Music of the 7ts: Predicting and decoding multivoxel fMRI responses with acoustic, schematic, and categorical music features," *Frontiers in Psychology*, vol. 8, p. 1179, 2017.

[7] T. Nakai, N. Koide-Majima, and S. Nishimoto, "Encoding and decoding of music-genre representations in the human brain," in *Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics*, ser. SMC 2018, 2018, pp. 584–589.

[8] T. Nakai, N. Koide-Majima, and S. Nishimoto, "Correspondence of categorical and feature-based representations of music in the human brain," *Brain and Behavior*, vol. 11, no. 1, p. e01936, 2021.

[9] J. S. Rahman, T. Gedeon, S. Caldwell, and R. Jones, "Brain melody informatics: Analysing effects of music on brainwave patterns," in *Proceedings of the 2020 International Joint Conference on Neural Networks*, ser. ICJNN 2020, 2020, pp. 1–8.

[10] V. K. Cheung, Y.-P. Peng, J.-H. Lin, and L. Su, "Decoding musical pitch from human brain activity with automatic voxel-wise whole-brain fMRI feature selection," in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP 2023, 2023.

[11] K. Tsekoura and A. Foka, "Classification of EEG signals produced by musical notes as stimuli," *Expert Systems with Applications*, vol. 159, p. 113507, 2020.

[12] V. De Angelis, F. De Martino, M. Moerel, R. Santoro, L. Hausfeld, and E. Formisano, "Cortical processing of pitch: Model-based encoding and decoding of auditory fMRI responses to real-life sounds," *NeuroImage*, vol. 180, pp. 291–300, 2018.

[13] P.-C. Chang, J.-R. Chang, P.-Y. Chen, L.-K. Cheng, J.-C. Hsieh, H.-Y. Yu, L.-F. Chen, and Y.-S. Chen, "Decoding neural representations of rhythmic sounds from magnetoencephalography," in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP 2021, 2021, pp. 1280–1284.

[14] S. Stober, D. J. Cameron, and J. A. Grahn, "Classifying EEG recordings of rhythm perception," in *Proceedings of the 15th International Society for Music Information Retrieval Conference*, ser. ISMIR 2014, 2014, pp. 649–654.

[15] S. Koelsch, V. K. Cheung, S. Jentschke, and J.-D. Haynes, "Neocortical substrates of feelings evoked with music in the acc, insula, and somatosensory cortex," *Scientific Reports*, vol. 11, no. 1, p. 10119, 2021.

[16] M. E. Sachs, A. Habibi, A. Damasio, and J. T. Kaplan, "Decoding the neural signatures of emotions expressed through sound," *NeuroImage*, vol. 174, pp. 1–10, 2018.

[17] V. Putkinen, S. Nazari-Farsani, K. Seppälä, T. Karjalainen, L. Sun, H. K. Karlsson, M. Hudson, T. T. Heikkilä, J. Hirvonen, and L. Nummenmaa, "Decoding music-evoked emotions in the auditory and motor cortex," *Cerebral Cortex*, vol. 31, no. 5, pp. 2549–2560, 2021.

[18] I. Daly, D. Williams, F. Hwang, A. Kirke, E. R. Miranda, and S. J. Nasuto, "Electroencephalography reflects the activity of sub-cortical brain regions during approach-withdrawal behaviour while listening to music," *Scientific Reports*, vol. 9, no. 1, p. 9415, 2019.

[19] S. Paquette, S. Takerkart, S. Saget, I. Peretz, and P. Belin, "Cross-classification of musical and vocal emotions in the auditory cortex," *Annals of the New York Academy of Sciences*, vol. 1423, no. 1, pp. 329–337, 2018.

[20] X. Cui, Y. Wu, J. Wu, Z. You, J. Xiahou, and M. Ouyang, "A review: Music-emotion recognition and analysis based on EEG signals," *Frontiers in Neuroinformatics*, vol. 16, p. 997282, 2023.

[21] D. S. Naser and G. Saha, "Influence of music liking on EEG based emotion recognition," *Biomedical Signal Processing and Control*, vol. 64, p. 102251, 2021.

[22] S. Hoefle, A. Engel, R. Basilio, V. Alluri, P. Toiviainen, M. Cagy, and J. Moll, "Identifying musical pieces from fMRI data using encoding and decoding models," *Scientific Reports*, vol. 8, no. 1, pp. 1–13, 2018.

[23] D. Sonawane, K. P. Miyapuram, B. Rs, and D. J. Lomas, "Guessthemusic: Song identification from electroencephalography response," in *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, ser. CODS-COMAD '21, 2021, pp. 154–162.

[24] P. Pandey, G. Sharma, K. P. Miyapuram, R. Subramanian, and D. Lomas, "Music identification using brain responses to initial snippets," in *Proceedings in the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP 2022, 2022, pp. 1246–1250.

[25] R. S. Schaefer, J. Farquhar, Y. Blokland, M. Sadakata, and P. Desain, "Name that tune: decoding music from the listening brain," *NeuroImage*, vol. 56, no. 2, pp. 843–849, 2011.

[26] D. Wu, C. Li, Y. Yin, C. Zhou, and D. Yao, "Music composition from the brain signal: Representing the mental state by music," *Computational Intelligence and Neuroscience*, vol. 2010, 2010.

[27] S. Stober, T. Prätzlich, and M. Müller, "Brain beats: Tempo extraction from EEG data." in *Proceedings of the 17th International Society for Music Information Retrieval Conference*, ser. ISMIR 2016, 2016, pp. 276–282.

[28] I. Sturm, M. Treder, D. Miklody, H. Purwins, S. Dähne, B. Blankertz, and G. Curio, "Extracting the neural representation of tone onsets for separate voices of ensemble music using multivariate EEG analysis." *Psychomusicology: Music, Mind, and Brain*, vol. 25, no. 4, p. 366, 2015.

[29] N. Gang, B. Kaneshiro, J. Berger, and J. P. Dmochowski, "Decoding neurally relevant musical features using canonical correlation analysis," in *Proceedings of the 18th International Society for Music Information Retrieval Conference*, ser. ISMIR 2017, 2017, pp. 131–138.

[30] I. Daly, "Neural decoding of music from the EEG," *Scientific Reports*, vol. 13, no. 1, pp. 1–17, 2023.

[31] L. May, A. R. Halpern, S. D. Paulsen, and M. A. Casey, "Imagined musical scale relationships decoded from auditory cortex," *Journal of Cognitive Neuroscience*, vol. 34, no. 8, pp. 1326–1339, 2022.

[32] S. Ntalampiras and I. Potamitis, "A statistical inference framework for understanding music-related brain activity," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 275–284, 2019.

[33] G. M. Di Liberto, G. Marion, and S. A. Shamma, "Accurate decoding of imagined and heard melodies," *Frontiers in Neuroscience*, vol. 15, p. 673401, 2021.

[34] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Towards music imagery information retrieval: Introducing the OpenMIIR dataset of EEG recordings from music perception and imagination." in *Proceedings of the 16th International Society for Music Information Retreival Conference*, ser. ISMIR 2015, 2015, pp. 763–769.

[35] A. Ofner and S. Stober, "Modeling perception with hierarchical prediction: Auditory segmentation with deep predictive coding locates candidate evoked potentials in EEG," in *Proceedings of the 21th International Society for Music Information Retrieval Conference*, ser. ISMIR 2020, 2020, pp. 566–573.

[36] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, and F.-R. Stöter, "Musical source separation: An introduction," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 31–40, 2019.

[37] E. Manilow, P. Seetharman, and J. Salamon, "Open source tools & data for music source separation," 2020. [Online]. Available: https://source-separation.github.io/tutorial

[38] G. Cantisani, G. Trégoat, S. Essid, and G. Richard, "MAD-EEG: an EEG dataset for decoding auditory attention to a target instrument in polyphonic music," in *Proceedings of the Speech, Music and Mind (SMM), Satellite Workshop of Interspeech 2019*, 2019.

[39] G. Cantisani, S. Essid, and G. Richard, "Neuro-steered music source separation with eeg-based auditory attention decoding and contrastive-nmf," in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP 2021, 2021, pp. 36–40.

[40] E. J. Allen, M. Moerel, A. Lage-Castellanos, F. De Martino, E. Formisano, and A. J. Oxenham, "Encoding of natural timbre dimensions in human auditory cortex," *NeuroImage*, vol. 166, pp. 60–70, 2018.

[41] V. Alluri, P. Toiviainen, I. P. Jääskeläinen, E. Glerean, M. Sams, and E. Brattico, "Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm," *NeuroImage*, vol. 59, no. 4, pp. 3677–3689, 2012.

[42] M. Ogg, D. Moraczewski, S. E. Kuchinsky, and L. R. Slevc, "Separable neural representations of sound sources: Speaker identity and musical timbre," *NeuroImage*, vol. 191, pp. 116–126, 2019.

[43] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP 2023, 2023.

[44] O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, H. Oya, S. S. Ghosh, J. Wright, J. Durnez, R. A. Poldrack, and K. J. Gorgolewski, "fmriprep: a robust preprocessing pipeline for functional MRI," *Nature Methods*, vol. 16, no. 1, pp. 111–116, 2019.

[45] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[46] J. A. Mumford, T. Davis, and R. A. Poldrack, "The impact of study design on pattern estimation for single-trial multivariate pattern analysis," *NeuroImage*, vol. 103, pp. 130–138, 2014.

[47] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. Van Essen, "A multi-modal parcellation of human cerebral cortex," *Nature*, vol. 536, no. 7615, pp. 171–178, 2016.

[48] A. Horn, "HCP-MMP1.0 projected on MNI2009a GM (volumetric) in NIfTI format," 2016. [Online]. Available: https://figshare.com/articles/dataset/HCP-MMP1_0_projected_on_MNI2009a_GM_volumetric_in_NIfTI_format/3501911

[49] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux, "Machine learning for neuroimaging with scikit-learn," *Frontiers in Neuroinformatics*, p. 14, 2014.

[50] S. Koelsch, "Neural substrates of processing syntax and semantics in music," *Current Opinion in Neurobiology*, vol. 15, no. 2, pp. 207–212, 2005.

[51] R. J. Zatorre, P. Belin, and V. B. Penhune, "Structure and function of auditory cortex: music and speech," *Trends in Cognitive Sciences*, vol. 6, no. 1, pp. 37–46, 2002.

[52] Y. Zhang, H. Ruan, Z. Yuan, H. Du, X. Gao, and J. Lu, "A learnable spatial mapping for decoding the directional focus of auditory attention using eeg," in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP 2023, 2023.

[53] A. Floren, B. Naylor, R. Miikkulainen, and D. Ress, "Accurately decoding visual information from fMRI data obtained in a realistic virtual environment," *Frontiers in Human Neuroscience*, vol. 9, p. 327, 2015.

[54] T. Horikawa and Y. Kamitani, "Generic decoding of seen and imagined objects using hierarchical visual features," *Nature Communications*, vol. 8, no. 1, p. 15037, 2017.

[55] Y.-T. Wu, H.-Y. Chen, Y.-H. Liao, L.-W. Kuo, and C.-C. Lee, "Modeling perceivers neural-responses using lobe-dependent convolutional neural network to improve speech emotion recognition." in *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, ser. INTERSPEECH 2017, 2017, pp. 3261–3265.

[56] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[57] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of the 7th International Conference on Learning Representations*, ser. ICLR 2019, 2019.

[58] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, 2017, p. 4768–4777.

[59] E. J. Humphrey, S. Reddy, P. Seetharaman, A. Kumar, R. M. Bittner, A. Demetriou, S. Gulati, A. Jansson, T. Jehan, B. Lehner, A. Krupse, and L. Yang, "An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 82–94, 2018.

[60] C. Gupta, H. Li, and M. Goto, "Deep learning approaches in topics of singing information processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2422–2451, 2022.

[61] W. Trost, T. Ethofer, M. Zentner, and P. Vuilleumier, "Mapping aesthetic musical emotions in the brain," *Cerebral Cortex*, vol. 22, no. 12, pp. 2769–2783, 2012.

[62] S. Koelsch and S. Skouras, "Functional centrality of amygdala, striatum and hypothalamus in a "small-world" network underlying joy: An fmri study with music," *Human Brain Mapping*, vol. 35, no. 7, pp. 3485–3498, 2014.

[63] P. N. Juslin, "From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions," *Physics of Life Reviews*, vol. 10, no. 3, pp. 235–266, 2013.

[64] R. J. Zatorre, J. L. Chen, and V. B. Penhune, "When the brain plays music: auditory–motor interactions in music perception and production," *Nature reviews neuroscience*, vol. 8, no. 7, pp. 547–558, 2007.

[65] C. L. Gordon, P. R. Cobb, and R. Balasubramaniam, "Recruitment of the motor system during music listening: An ale meta-analysis of fmri data," *PloS ONE*, vol. 13, no. 11, p. e0207213, 2018.

[66] J. K. Bizley and Y. E. Cohen, "The what, where and how of auditory-object perception," *Nature Reviews Neuroscience*, vol. 14, no. 10, pp. 693–707, 2013.

[67] V. K. M. Cheung and S. Sakamoto, "Separating uncertainty from surprise in auditory processing with neurocomputational models: Implications for music perception," *Journal of Neuroscience*, vol. 42, no. 29, pp. 5657–5659, 2022.

[68] S. K. Zieliński, P. Hardisty, C. Hummersone, and F. Rumsey, "Potential biases in MUSHRA listening tests," in *Proceedings of the 123rd Audio Engineering Society Convention*, ser. AES 123, 2007.

[69] A. Furnham, "Response bias, social desirability and dissimulation," *Personality and Individual Differences*, vol. 7, no. 3, pp. 385–400, 1986.

[70] G. Kalton and H. Schuman, "The effect of the question on survey responses: A review," *Journal of the Royal Statistical Society: Series A (General)*, vol. 145, no. 1, pp. 42–57, 1982.