Capturing the musical brain with Lasso: Dynamic decoding of musical features from fMRI data

Petri Toiviainen^{a,*}, Vinoo Alluri^{a,b}, Elvira Brattico^{a,c,d}, Mikkel Wallentin^e, Peter Vuust^{e,f}

^a Finnish Centre of Excellence in Interdisciplinary Music Research, Department of Music, University of Jyväskylä, Finland

^b Department of Mathematical Information Technology, University of Jyväskylä, Finland

^c Brain & Mind Laboratory, Biomedical Engineering and Computational Science (BECS), Aalto University, Finland

^d Cognitive Brain Research Unit (CBRU), Institute of Behavioural Sciences, University of Helsinki, Finland

^e Center of Functionally Integrative Neuroscience, Aarhus University Hospital, Denmark

^f Royal Academy of Music, Aarhus, Denmark

ARTICLE INFO

Article history:

Accepted 10 November 2013

Available online 19 November 2013

Keywords:

Music

fMRI

Music Information Retrieval

Decoding

Time series

Naturalistic paradigm

ABSTRACT

We investigated neural correlates of musical feature processing with a decoding approach. To this end, we used a method that combines computational extraction of musical features with regularized multiple regression (LASSO). Optimal model parameters were determined by maximizing the decoding accuracy using a leave-one-out cross-validation scheme. The method was applied to functional magnetic resonance imaging (fMRI) data that were collected using a naturalistic paradigm, in which participants' brain responses were recorded while they were continuously listening to pieces of real music. The dependent variables comprised musical feature time series that were computationally extracted from the stimulus. We expected timbral features to obtain a higher prediction accuracy than rhythmic and tonal ones. Moreover, we expected the areas significantly contributing to the decoding models to be consistent with areas of significant activation observed in previous research using a naturalistic paradigm with fMRI. Of the six musical features considered, five could be significantly predicted for the majority of participants. The areas significantly contributing to the optimal decoding models agreed to a great extent with results obtained in previous studies. In particular, areas in the superior temporal gyrus, Heschl's gyrus, Rolandic operculum, and cerebellum contributed to the decoding of timbral features. For the decoding of the rhythmic feature, we found the bilateral superior temporal gyrus, right Heschl's gyrus, and hippocampus to contribute most. The tonal feature, however, could not be significantly predicted, suggesting a higher inter-participant variability in its neural processing. A subsequent classification experiment revealed that segments of the stimulus could be classified from the fMRI data with significant accuracy. The present findings provide compelling evidence for the involvement of the auditory cortex, the cerebellum and the hippocampus in the processing of musical features during continuous listening to music.

© 2013 Elsevier Inc. All rights reserved.

Introduction

Music, due to its inherent temporal nature, provides a natural means to investigate dynamic aspects of perception. Recent advances in Music Information Retrieval (MIR) allow dynamic extraction of perceptually relevant features from musical recordings using signal processing (Lartillot, and Toiviainen, 2007; Mueller, Ellis, Klapuri, and Richard, 2011). This provides a computational formalism for obtaining a non-linear mapping from input space to feature space as depicted by Naselaris et al. (Naselaris et al., 2011) and allows the use of real music to study the neural dynamics of musical feature processing.

To date, only a few studies have investigated neural processing of musical features using functional magnetic resonance imaging (fMRI) during a naturalistic listening condition (Alluri et al., 2012, 2013; Chapin, Jantzen, Kelso, Steinberg, and Large, 2010; Janata, 2009; Lehne et al., 2013). Alluri and colleagues (Alluri et al., 2012) recorded the brain responses of their participants while they were listening to a piece of Argentine tango. Using computational algorithms they extracted musical feature time series related to timbre, rhythm, and tonality from the stimulus. Using subsequent correlational analyses, they were able to locate areas in the auditory cortex (superior temporal gyrus), the somatosensory regions, the motor cortex and the default mode areas, and the cerebellum that responded to timbral features, in contrast to focal activity in motor and limbic regions, such as the amygdala, hippocampus and insula/claustrum, that responded to higher-order tonal and rhythmic features. In a subsequent study Alluri and colleagues (Alluri et al., 2013), developed an encoding model based on Principal Components regression to predict the dynamics of stimulus-related

* Corresponding author at: Department of Music, PO Box 35(M), 40014 University of Jyväskylä, Finland.

E-mail addresses: petri.toiviainen@jyu.fi (P. Toiviainen), vinoo.alluri@jyu.fi (V. Alluri), elvira.brattico@helsinki.fi (E. Brattico), mikkel@cfm.dk (M. Wallentin), pv@musikksn.dk (P. Vuust).

brain activation, which they cross-validated across two sets of musical stimuli to locate brain areas whose activation could be predicted across stimuli. Their findings evidenced a region of the superior temporal gyrus (encompassing the planum polare and the Heschl's gyrus) in the right hemisphere as well as a region of the orbitofrontal cortex, which were robustly activated by musical features during continuous listening.

A limitation of the aforementioned studies is that they rely exclusively on the encoding approach, thus attempting to understand how neural activation changes when stimulus features vary. Given the complexity of naturalistic music stimuli, it is crucial to extend the methodological palette in such studies to include decoding methods as well, thus predicting stimulus features from the neural activation. As Naselaris et al. (2011) put it, decoding models can be used to validate encoding models and provide a “sanity check” on the conclusion drawn from them. Accordingly, comparing results obtained from fMRI decoding studies on music to those from corresponding encoding studies, in terms of both feature predictability and associated spatial maps, will allow to pinpoint the areas that are crucial for the processing of musical features.

One of the challenges encountered when using decoding is that whole-brain fMRI data typically consists of a large number of features (voxels) and a relatively low number of observations (scans or trials). From a decoding perspective, this gives rise to various problems related to computational burden, overfitting, and difficult interpretability. Most fMRI decoding studies rely on theory-driven feature selection, in which the regions of interest (ROI) used in the decoder are defined based on previous knowledge about brain function (Cox and Savoy, 2003; Formisano, De Martino, and Valente, 2008). This approach, however, may be difficult if the knowledge about the ROIs is limited. This can be the case, for instance, with naturalistic stimuli, in which a multitude of different parameters are changing simultaneously, thus activating large brain areas (see, e.g., Alluri et al., 2012).

In this paper, we present a method for decoding musical feature dynamics from fMRI data obtained during listening to entire pieces of music. The method is purely data-driven, combining voxel selection by filtering, dimensionality reduction, and regression with embedded feature selection. At the first stage, a subset of voxels is selected from whole-brain data based on mean inter-participant correlation. Subsequently, PCA is applied to the group-level average of the selected voxels. Finally, the obtained PC projections are regressed against computationally extracted acoustic feature time series using the Least Absolute Shrinkage and Selection Operator (LASSO). Optimal model parameters (proportion of retained voxels and Lasso regularization parameter) are determined by maximizing the decoding accuracy using a leave-one-out cross-validation scheme. The method is deterministic and computationally and conceptually relatively simple.

We applied the method to previously published fMRI data that were collected using a naturalistic paradigm, in which participants' brain responses were recorded while they were listening to pieces of real music (Alluri et al., 2013). The dependent variables comprised musical feature time series that were computationally extracted from the stimulus. We were mainly interested in two questions: (1) the decoding accuracy of each musical feature; and (2) the areas that contribute to the optimal decoding models for each feature.

In general, we expected the results to agree with those obtained by Alluri et al. (2012, 2013), corroborating them with a decoding approach. More specifically, we had the following four hypotheses:

- As Alluri et al. (2012) found stronger responses to timbral features (maximal z values being in the range of 7–8) than to rhythmic and tonal features (maximal z values being in the range of 4–5), we expected timbral features to attain higher decoding accuracy than rhythmic and tonal features.
- As Alluri et al. (2012) observed high z values for timbral features in auditory cortical areas in the bilateral superior temporal gyrus (STG),

Heschl's gyrus (HG), middle temporal gyrus (MTG), and Rolandic operculum (RO) as well as areas in the cerebellum, we expected these areas to contribute significantly to the decoding models of timbral features.

- Based on the cross-validation study (Alluri et al., 2013), which found right-hemispheric asymmetry of processing instrumental music, we expected a hemispheric asymmetry effect in the decoding models, particularly for the auditory cortex areas.
- For the higher-order features, namely Pulse Clarity and Key Clarity, we expected, following the results in Alluri et al. (2012), to observe additional contributions of limbic areas, such as the amygdala, hippocampus, and insula, as well as motor and auditory areas.

In a subsequent classification experiment, we utilized the musical feature time series decoded from the individual participants' responses to classify segments based on their fMRI data. The purpose of this experiment was to get an overall estimate of the decoding accuracy.

Materials and methods

Data acquisition

Participants

Participants comprised fifteen healthy individuals (mean age: 25.7 ± 5.2 SD; 10 males), and were selected without regard to their musical education. None reported any neurological, hearing or psychological problems. All participants were right-handed. Permission for the study was obtained from the local ethics committee (Region Midtjylland, Denmark) and written informed consent was obtained from each participant. Each received a 100 DKK payment per hour for participation. All participants had Danish as their primary language. Further details on these subjects can be found from the paper reporting findings of cross-validation models by Alluri et al. (2013).

Stimulus

The musical stimulus comprised the B-side of the album *Abbey Road* by The Beatles (1969). The duration of the stimulus was approximately 16 min. The stimulus was played in mono with a sampling frequency of 22050 Hz and delivered through pneumatic headphones from Avotec (Stuart, FL USA). Participants were instructed to listen carefully to the music. To maintain their attention, we inserted in four places of the stimulation the voice stimulus “nu” (Danish word for “now” in English) and asked subjects to press a button whenever they heard it. This is a low-level task that requires only limited attention from the participants.

fMRI data acquisition

A 3T General Electrics Medical Systems (Milwaukee, WI USA) MR system with a standard head coil was used to acquire both T2-weighted gradient echo, echo-planar images (EPI) with Blood Oxygenation Level-Dependent (BOLD) contrast and T1-weighted structural images. 464 EPI volumes were acquired per participant. The first five volumes were discarded to allow for effects of T1 equilibrium. Whole brain coverage was achieved using 42 axial slices of 3 mm thickness with an in-plane resolution of 3×3 mm in a 64×64 voxel matrix (FOV 192 mm). Images were obtained with a TR of 2200 ms, a 30 ms TE and a 90° flip angle. A high-resolution 3D GR T1 anatomical scan was acquired for spatial processing of the fMRI data. It consisted of $256 \times 256 \times 134$ voxels with a $0.94 \text{ mm} \times 0.94 \text{ mm} \times 1.2 \text{ mm}$ voxel size, obtained with a TR of 6.552 ms, a 2.824 ms TE and a 14° flip angle.

Whole-brain image analysis was carried out using Statistical Parametric Mapping 8 (SPM8 – <http://www.fil.ion.ucl.ac.uk/spm>). For each subject the images were realigned, spatially normalized into the Montreal Neurological Institute template (12 parameter affine model, gray matter segmentation; realignment: translation components < 2 mm, rotation components $< 2^\circ$), and spatially smoothed (Gaussian filter with FWHM of 6 mm). Following this, a

high-pass filter with a cut-off frequency of .008 Hz, which conforms to the standards used to reduce the effects the scanner drift typically occurring at a timescale of 128 s (Smith et al., 1999), was used to detrend the fMRI responses. Next, temporal smoothing was performed as it provides a good compromise between efficiency and bias (Friston et al., 2000). The Gaussian smoothing kernel had a width of 5 s, which was found to maximize the correlation between the frequency response of the HRF and the smoothing kernel. Finally, the effect of the participants' movements was removed by modeling the 6 movement parameters as regressors of no interest.

Acoustic feature processing

We used the approach implemented by Alluri et al. (2012) for acoustic feature selection and extraction. Thereby twenty-five acoustic features capturing timbral, rhythmical and tonal properties were extracted from the stimulus using the MIR Toolbox (Lartillot and Toivainen, 2007). The features were extracted using a frame-by-frame analysis approach commonly used in the field of Music Information Retrieval (MIR). For the timbral features, the duration of the frames was 25 ms and the overlap between two adjacent frames 50% of the frame length, while for the rhythmical and tonal features the respective values were 3 s and 67% (see Alluri et al., 2012, for a comprehensive overview). For each feature, this resulted in a time series representing its temporal evolution. All the operations were performed in the MATLAB environment.

Following feature extraction, we performed a series of post-processing operations to make the data comparable to the fMRI data. First, the lag present in the fMRI data due to the hemodynamic response was accounted for by convolving each of the acoustic features with a double-gamma Hemodynamic Response Function (HRF) having the peak at 5 s and the undershoot at 15 s. Next, the convolved acoustic feature time-series were subjected to the same detrending operation in the post-processing stage of the fMRI data in order to eliminate those low-frequency components whose eventual brain correlates were eliminated during the preprocessing stage of the fMRI time-series. For subsequent analysis we excluded the first 26 s corresponding to the length of the HRF in order to avoid any artifacts due to the convolution operation. Following this, all the features were downsampled to match the sampling rate of the fMRI data (0.45 Hz). After this, both the acoustic features and the fMRI data had 441 time points.

To reduce the dimensionality, the acoustic features were subsequently subjected to Principal Components Analysis (PCA). Nine Principal components (PCs) were retained, as they accounted for >95% of the total variance. Following this, the retained PCs were subjected to Varimax rotation. This procedure was identical to the one applied by Alluri et al. (2012) and the obtained PCs were similar. In the present PCA solution, six of the PCs had loadings similar to the six PCs perceptually validated by Alluri et al. (2012), the correlations between the respective loadings being >.70. These PCs, subsequently referred to as acoustic components, were retained for subsequent analysis and were, according to the nomenclature used in Alluri et al. (2012), labeled as Fullness, Brightness, Activity, Timbral Complexity, Pulse Clarity, and Key Clarity.

Decoding

LASSO-PC regression

As the first part of the decoding approach, we trained models to predict the time course of evolution of each of the six acoustic components. The models comprised a combination of voxel selection, spatial PCA and Lasso regression. To avoid overfitting, a leave-one-out cross-validation scheme was utilized. Fig. 1 presents the decoding approach schematically. See Appendix A.2–A.4 for a mathematical description of the procedure. As the first step, one participant's data were taken out from the data pool, the remaining participants constituting the training set. Within the training set, the mean pairwise inter-participant correlation was calculated for each voxel, and a subset of voxels showing the highest mean inter-subject correlation was selected. The extent of this subset was varied,

comprising 1/4, 1/8, 1/16, 1/32, or 1/64 of the total number of voxels. For these voxels, the fMRI time series were averaged across the participants in the training set. Spatial PCA was applied to the averaged data and the thus obtained PC scores were subjected to Lasso regression, using each of the acoustic components in turn as the dependent variable. The Lasso was trained using the Least Angle Regression (LARS) algorithm (Efron, Hastie, Johnstone, and Tibshirani, 2004). The LARS produces a series of regression models, in which the value of the regularization parameter and consequently the sparsity (number of nonzero regression coefficients) varies across models. Subsequently, the respective acoustic component was predicted with the obtained regression models using the data from the participant not in the training set. The procedure was repeated for each participant. Optimal regularization parameters were estimated separately for each acoustic component by maximizing the decoding accuracy, as measured by the average Pearson correlation between actual and predicted acoustic components across participants.

Classification

To further investigate the decoding accuracy, a series of classification runs was carried out. To this end, the stimulus was divided to N segments of equal length, where $N = 2, \dots, 10$. For each number of segments, both the acoustic component data predicted from each participant's fMRI data using the optimal model and the actual acoustic component data were divided according to the respective segmentation. Subsequently, for each participant and segment the predicted data was classified as one of the segments using a similarity measure based on the average Pearson's correlation between actual and predicted acoustic component data (see Appendix A.6. for a mathematical description of the procedure).

Results

Model selection

Fig. 2 displays the decoding accuracy of the models for each acoustic component. The lines correspond to different proportions of voxels included in the models, and the horizontal axis corresponds to the number of nonzero components in the regression model. As can be seen, the decoding performance tends to reach its maximum with a relative low number of nonzero coefficients, the effect being most significant with Activity and Pulse Clarity. Furthermore, the proportion of included voxels has a significant effect on the performance, in particular with Brightness and Activity. As Key Clarity showed a low decoding accuracy, it was excluded from subsequent analyses.

Table 1 displays the optimal parameters for the Lasso models per each acoustic component. As can be seen, the optimal models were able to predict the acoustic components with moderate accuracy, with the exception of Key Clarity, which was consequently excluded from subsequent analyses. In terms of the optimal number of voxels, the sparsest optimal model was that for Fullness, while the least sparse was that for Pulse Clarity. On the average, in the optimal regression models ca. 12 (2.7%) of the PCs had non-zero beta coefficients.

Prediction accuracy

The distributions of Pearson correlations between actual and predicted acoustic features across participants are shown in the box plot of Fig. 3. As can be seen, the correlations are moderately high, the median correlations ranging between .38 and .52. Some of the components, in particular Fullness and Activity, show quite considerable between-participant variance in the prediction accuracy. Meanwhile, the acoustic component Timbral Complexity has the most consistent prediction accuracy across participants.

The significance of the correlations was estimated using a Monte Carlo simulation. Subsequently, multiple comparisons were corrected for with false discovery rate (FDR) control utilizing the Benjamini–

Fig. 1. Schematic presentation of the procedure for decoding acoustic components from one participant's fMRI data. See Appendix A.2–A.4 for a mathematical description. The symbols below and on the right side of each data matrix symbol indicate the dimensionality of the respective matrix. Explanation of symbols: A : acoustic feature matrix; $\bar{X}_v$: voxel time series data of participant s ; $\bar{p}_v$: mean intersubject correlation for voxel v ; p : voxel selection parameter; $\bar{X}_{-sp}$: voxelwise mean of fMRI data excluding participant s ; Y_{-sp} : loading matrix from PCA; C_{-sp} : score matrix from PCA; λ : Lasso regularization parameter; $\hat{\beta}_{cs}$: regression coefficients for participant s and acoustic component c ; $\hat{a}_{cs}$: acoustic feature c predicted from the response of participant s ; a_c : true (extracted) acoustic feature c .

Fig. 2. Goodness of prediction for the acoustic components Fullness, Brightness, Activity, Timbral Complexity, Pulse Clarity, and Key Clarity, as a function of the number of non-zero beta coefficients. In each subplot, the lines correspond to different proportions of voxels included in the model, as displayed in the legend. For each acoustic component, the best model is indicated with a red circle.

Table 1

The proportion of voxels included, the number of nonzero components and the value of regularization parameter λ for the optimal model for each acoustic component.

	Proportion of voxels (p)	Number of nonzero coefficients	$\hat{\lambda}_i$
Fullness	1/64	34	56.1
Brightness	1/16	2	235.7
Activity	1/16	12	105.0
Timbral complexity	1/16	5	164.3
Pulse clarity	1/8	5	188.4

Hochberg procedure. Table 2 displays, for each acoustic component, the mean correlation value and the number of participants who displayed a significant correlation according to different significance thresholds. As can be seen, almost all of the acoustic features could be significantly predicted in a majority of participants. According to these results, the component that could be most accurately predicted was Fullness, while the decoding accuracy for Timbral Complexity was the most consistent across participants.

To provide an example of the match between actual and predicted acoustic components, Fig. 4 displays the respective data for the components Fullness and Pulse Clarity. As can be seen, the predicted acoustic component time series follow the overall pattern of the respective actual component, while missing some detail of the fine structure, in particular more abrupt changes in the target signal.

Spatial maps

To investigate which brain regions contributed to the prediction of each acoustic component, Lasso models were trained for each acoustic component using the whole set of fMRI data with the optimal parameters obtained from the cross-validation. Subsequently, each set of obtained regression coefficients was projected back to voxel space by multiplying it by the transpose of the PC loading matrix (see Appendix A.5).

As some of the maps were expected to overlap with each other, the method by Valente et al. (2013) was used to remove for each acoustic component the contribution of the remaining acoustic components. To this end, for each acoustic component the voxel data used in training the decoder was augmented with the remaining acoustic components.

The significance of the regression coefficients was estimated for each acoustic feature using bootstrap resampling, following the procedure used by (McIntosh and Lobaugh, 2004). To this end, 14

Table 2

Median correlations between actual and predicted acoustic components and the number of participants with significant prediction at significance levels $p < .05$, and $p < .01$, using Benjamini–Hochberg procedure to control for false discovery rate (FDR).

Feature	Mean correlation	Number of participants with significant prediction ($p < .05$; total $N = 15$)
Fullness	0.49	9
Brightness	0.32	9
Activity	0.31	9
Timbral complexity	0.46	15
Pulse clarity	0.38	10

participants' data were first sampled randomly with replacement, following which the Lasso model was trained with the optimal parameters. This was repeated 1000 times for each acoustic feature, after which the voxel-wise standard error was calculated for each acoustic feature. Subsequently, a voxel was regarded significant if the absolute value of its regression coefficient exceeded 1.65 times the estimated standard error ($p < .05$). Following this, a cluster-size correction was performed with the same threshold as used by Alluri et al. (2013), that is, $n = 22$.

Subsequently, MarsBaR v0.43 (<http://marsbar.sourceforge.net>) was used to extract the regions of interest falling under each resulting cluster. For each cluster, the $x y z$ coordinates (in MNI space) of the voxel with the maximum Z -value were determined. Anatomical areas were determined using Automated Anatomical Labeling (AAL; Tzourio-Mazoyer et al., 2002). The thus obtained clusters are shown in Table 3. Additionally, Fig. 5 displays significant voxels for selected axial slices.

As can be seen from Table 3, the significant regions for all timbral features include areas in STG and HG. Moreover, RO is included in the models for Fullness, Brightness, and Activity. Finally, only Brightness and Activity contain significant areas in MTG and the cerebellum. In order to investigate the hemispheric symmetry of the regions in HG, RO, and MTG, Table 4 displays the number of significant voxels in these areas for each timbral feature and hemisphere. As can be seen, HG displays a clear asymmetry for all features, with the right hemisphere showing a higher number of significant voxels. For RO no consistent asymmetry is observed. Finally, for MTG a clear left-dominance of significant voxels is observed for all features (except Fullness which did not recruit the MTG).

Fig. 3. Distribution of decoding accuracy, as measured by Pearson's correlations between actual and predicted acoustic components, across participants. The red lines show the medians, the bottom and top of the blue boxes the 1st and 3rd quartile, and the whiskers the range of the data.

Fig. 4. Actual and predicted values for the acoustic components Fullness and Pulse Clarity. The blue line shows the prediction averaged across participants, and the black lines the average ± 1 standard deviation.

Table 3

Clusters (with size, location of maximum and maximal z value) showing significant ($p < .05$) coefficients in the decoding models of each acoustic feature. The coordinates are in MNI space. HG = Heschl's gyrus; RO = Rolandic operculum; STG = Superior temporal gyrus; MTG = Middle temporal gyrus; ITG = Inferior temporal gyrus; SFG = Superior frontal gyrus; IFG = Inferior frontal gyrus; MCG = Median cingulate gyrus; MPCG = Medial paracingulate gyrus; PCG = Posterior cingulate gyrus; PPCG = Posterior paracingulate gyrus.

Left hemisphere	N	x	y	z	Z	Right hemisphere	N	x	y	z	Z
<i>Fullness</i>											
Positive coefficients											
HG, RO, Insula, STG	121	−32	−30	16	3.17	HG, STG, Insula, RO	96	48	−14	4	2.45
						HG, Insula	43	36	−26	12	3.09
Negative coefficients						STG	43	58	−8	−8	2.72
<i>Brightness</i>											
Positive coefficients											
RO, Insula, HG	53	−32	−30	18	2.31	HG, Insula, STG, RO	328	38	−26	12	3.22
STG, HG	46	−38	−24	4	2.26	MCG, MPCG	36	10	−18	36	2.61
Negative coefficients											
MTG, STG	272	−60	−22	2	2.86	Lobules III–VI, Vermis I–II, IV–V, VI–VIII	154	20	−48	−20	2.81
SFG, orbital part, MFG, orbital part, Gyrus rectus	40	−16	32	−26	2.08	STG, MTG	51	50	−32	0	2.19
Caudate nucleus, Olfactory cortex	25	−2	8	−4	2.04						
<i>Activity</i>											
Positive coefficients											
RO, Insula, HG	53	−32	−30	16	2.35	HG, Insula, STG, RO	339	38	−26	12	3.42
STG, HG	47	−38	−24	4	2.34	MCG, MPCG (R)	32	10	−18	36	2.56
Negative coefficients											
MTG, STG	259	−60	−20	2	2.75	Lobules IV–VI	82	20	−48	−20	2.73
						STG, MTG	55	48	−32	0	2.20
						Lobules IV–V, Vermis IV–VIII	39	6	−60	−26	2.35
<i>Timbral Complexity</i>											
Positive coefficients											
						Caudate nucleus, Thalamus	75	10	4	6	1.73
Negative coefficients											
STG, MTG, HG	447	−60	−20	6	2.38	STG, HG, Insula, RO	530	56	−10	4	2.50
<i>Pulse Clarity</i>											
Positive coefficients											
Lingual gyrus, Lobules III–VI of cerebellum	240	−10	−44	−24	2.28	Vermis IV–VI, VIII, Lobules IV–VI, Lingual gyrus	225	8	−60	−20	2.58
ACG, APCG, IFG, medial orbital	174	−8	38	−4	2.60	ACG, APCG, IFG, medial orbital	28	4	36	−4	2.34
SFG, orbital part, Gyrus rectus	31	−10	36	−24	2.20	Insula, HG, STG, RO	283	46	−14	8	2.44
						Parahippocampal gyrus	27	24	−16	−26	2.50
Negative coefficients											
MTG	95	−56	−30	2	2.07	MTG, STG, ITG	381	60	−4	−4	2.60
STG, MTG	43	−54	−48	18	2.13	Globus pallidus, caudate nucleus, Putamen	24	10	4	6	2.29
						Insula, Temporal pole, STG, IFG, orbital part	90	48	18	−12	2.41
						MCG, MPCG, PCG, PPCG	27	4	−36	28	1.98
						Insula	25	32	18	−14	2.20

For Pulse Clarity a relatively large number of significant areas can be observed for bilateral STG (24 and 206 voxels in the left and right hemispheres, respectively) and MTG (114 and 381 voxels), right HG (100 voxels) and right ITG (39 voxels), as well as the bilateral cingulate gyrus (126 and 27 voxels). The amygdala, hippocampus, and supplementary motor areas, however, failed to display significance in the decoding model of this acoustic feature. Moreover, large bilateral areas of significant voxels are obtained in the cerebellum.

Classification

The mean correct classification rates across participants obtained for each segmentation scheme are displayed in Fig. 6. As can be seen from the figure, the mean correct classification rate is 2–3 times the chance level for all segmentations. According to a Monte Carlo permutation test, all classification results are significant ($p < .001$).

Fig. 7 displays the confusion matrices for each of the classification runs. In each of the subplot, each square displays the number of classification instances for the respective pair of actual and predicted segment, with dark shades corresponding to high numbers and vice versa. The diagonal elements correspond to correct classification of the respective segments. Consistent with Fig. 6, the elements on the diagonal tend to have higher values than other elements. There are, however, fairly significant differences between different segments in terms of their prediction pattern, as indicated by the uneven distribution of incorrectly classified instances.

Discussion

In this paper, we have presented a novel data-driven method for the decoding of musical feature dynamics from fMRI data, based on a combination of filtered feature selection, dimensionality reduction, and

Fig. 5. Voxels with Lasso regression coefficients significantly different from zero ($p < .05$) for each of the acoustic components, displayed for selected axial slices. Red and blue denote positive and negative coefficients, respectively. The z values are in MNI coordinates. The maps have been cluster-size corrected using a threshold of $n = 22$.

Lasso regression with embedded feature selection. We applied the method to fMRI data collected while participants were listening to a 16-minute excerpt of the album *Abbey Road* by the Beatles and to musical features that were computationally extracted from the same

stimulus. Using a between-participant cross-validation scheme, the model parameters were optimized to yield maximal decoding accuracy. Subsequently, the optimal models were used in a classification task, in which the similarity between actual and predicted acoustic components was used as the distance metric. The main findings can be summarized as follows. First, we found that most of the computationally extracted musical features could be significantly predicted for a majority of participants. Second, there was a relatively large between-participant variation in the prediction accuracy. Third, we found that optimal models were sparse, comprising less than 4% of the voxels with significant regression coefficients. Fourth, the areas that most contributed to the decoding resided mostly in auditory and motor areas for timbral features and additionally in limbic and frontal areas for the rhythmic feature. Finally, in a classification experiment we found that segments from the stimulus could be significantly classified for a number of different segmentation schemes, although there was great variation in the classification accuracy of different segments.

Table 4

Number of significant voxels ($p < .05$) in the optimal decoding models for the timbral features in the Heschl's gyrus (HG), Rolandic operculum (RO), and middle temporal gyrus (MTG) in each hemisphere.

	HG		RO		MTG	
	Left	Right	Left	Right	Left	Right
Fullness	40	87	39	1	0	0
Brightness	10	145	12	22	165	14
Activity	18	147	12	19	154	18
Timbral complexity	20	60	0	3	35	0

Fig. 6. Mean correct classification rate as a function of number of segments (blue line). The black line displays the chance level and the dashed red line the $p < .001$ limit, corresponding to the 99.9% point in the cumulative distribution obtained from a Monte Carlo permutation test.

The acoustic components under investigation were found to differ in terms of their prediction accuracy. In particular, and in accordance with our first hypothesis, the timbral components could be decoded more accurately than the rhythmic and tonal ones. One explanation for this could be that timbre as a musical element is more low-level than rhythm and tonality (see Alluri et al., 2012), and one could thus assume

a more consistent pattern of neural processing for timbre than for other musical elements.

With regard to the optimal models of timbral features, we found, in line with our second hypothesis, large areas in the superior temporal gyrus, Heschl's gyrus, and Rolandic operculum in all or most of the features, suggesting that these are core areas in the processing of timbre. While Brightness and Activity showed significant areas in the middle temporal gyrus and cerebellum, the Fullness and Timbral Complexity did not encompass these areas. This discrepancy calls for further study.

Regarding hemispheric specialization, we found, following our third hypothesis, clear asymmetry in Heschl's gyrus and middle temporal gyrus, but not in the Rolandic operculum. In particular, the right Heschl's gyrus and the left middle temporal gyrus showed larger significant regions than their homotopic counterparts. This asymmetry was particularly clear for Brightness and Activity, suggesting some degree of hemispheric specialization in the processing of those features, and being in line with the findings by Alluri et al. (2012).

For Pulse Clarity, we found, in accordance with our fourth hypothesis, significant regions in the bilateral superior temporal gyrus, right Heschl's gyrus, and hippocampus, again suggesting these regions to be the core areas for the processing of pulse. We failed, however, to find significant regions in the other hypothesized areas, such as the amygdala and putamen, suggesting that these areas bear less significance in the processing of that particular acoustic feature. The (negative) correlations of several limbic regions to pulse observed in Alluri et al. (2012) might be restricted to the stimulus used. In the current study a variety of musical pieces were utilized, whereas in Alluri et al. only one musical piece was adopted, perhaps allowing the development of affective responses to music, which are known to activate limbic and reward

Fig. 7. Confusion matrices obtained for each segmentation. The darkness of each square indicates the proportion of classification instances for the respective actual–predicted segment pair, black denoting 100% and white 0%.

areas of the brain (Brattico et al., 2011; Janata, 2009; Pereira et al., 2011).

We failed to decode Key Clarity with sufficient accuracy, and the feature was therefore left out from subsequent analyses. In addition to tonality being a high-level feature, which can be expected to display higher inter-subject variability than more low-level features, our failure to decode this feature could also be attributed to the relatively stable tonal structure of the stimulus (see Alluri et al., 2013), which prevented from inducing sufficiently strong tonality-related activation.

Overall, a notable proportion of the significant areas found in the present study were consistent with the respective areas found by Alluri et al. (2012). This latter study was conducted on another musical stimulus and a different subject pool of musicians, whereas here also non-musicians were included. Furthermore, it is interesting to note that here the decoding models for all features comprised a large region of high significance encompassing the right Heschl's gyrus, superior temporal gyrus, Rolandic operculum, and insula, agreeing with the results from the cross-validation study by Alluri et al. (2013), conducted on the same dataset as this one. This finding provides further evidence for the importance of these regions in musical feature processing.

The regression models with maximal decoding accuracy were overall sparse, comprising on average ca. 6.5% of the voxels. Furthermore, voxels with regression coefficient significantly different from zero comprised on average less than 2% of the scan volume. This suggests that although listening to real music activates relatively large areas in the brain (see Alluri et al., 2012), the areas in which the activation is sufficiently consistent across participants to allow decoding are sparser. The individual acoustic components differed with respect to model sparsity. Of all the acoustic components, pulse clarity had the highest number of voxels in the optimal model (1/8 of all voxels). This could be explained by the context-dependency of pulse perception, in particular that it requires temporal integration of information over a window of a few seconds (Fraisse, 1982). Consequently, one could assume that networks involved in the processing thereof are wider than for the other acoustic elements.

In the classification experiment a modest but significant performance was obtained by the method used. It must be noted that classification performance in general is dependent on class separability. In the present case, the segments to be classified belong to the same musical genre. In a classification task with musical examples from a wider range of music one could assume that a higher correct classification rate could be achieved. It was also observed that individual segments differed in their prediction accuracy and confusion patterns. This may be due to the difference in variation in the musical features within each segment. In particular if a segment contains a small amount of variation in terms of its musical content, it may be difficult to classify it correctly with the present correlation-based method. More work would be needed to account for this variability.

The present study utilized a between-participant cross-validation scheme using a single stimulus for model selection. This kind of cross-validation provides an estimate of the model's capability to generalize over new measurements using the same stimulus. While it certainly reduces the problem of overfitting to the data, it may still cause the trained models to learn relationships that are idiosyncratic to the particular stimulus used and do not generalize to new stimuli (Kriegeskorte, 2011). Therefore it would be crucial to perform such an experiment using a larger variety of stimuli and a cross-stimulus validation scheme. We plan to carry out such a study in the future.

Music is known to evoke strong emotions and affect the mood of the listeners. Despite active research in this area, neural correlates of musical emotion are still not well understood (Koelsch, 2010). In particular, the dynamics of neural processing of musical emotions has not been investigated. Therefore an interesting future direction of research would be to extend the present approach to include subjective measures of emotion. This would allow elucidating the musical and neural correlates of subjective experience of music.

Acknowledgments

This research was supported by the Academy of Finland (Centre of Excellence program, project number 141106; post-doctoral researcher grant, project number 133673), the University of Helsinki (three-year grant, project number 490083), by the TEKES (Finland) grant 40334/10 'Machine Learning for Future Music and Learning Technologies', and by the MindLab grants from the Danish Ministry of Science, Technology and Innovation. We wish to thank Andreas Højlund Nielsen and Anders Dohn for their help during the data collection and Peter M. Rasmussen for his help in designing the model.

Appendix A

A.1. Acoustic feature processing

Acoustic feature processing was carried out following the procedure used by Alluri et al. (2012). The 25 extracted features were convolved with canonical hemodynamic response function (double gamma), resampled to match the fMRI sampling rate, and subjected to Principal Components Analysis. This resulted in the feature matrix $\mathbf{F} \in \mathbb{M}(T, 25)$, where T is the number of time points ($T = 441$). Nine Principal Components, accounting for 95% of the variance, were retained and subjected to Varimax rotation, yielding the loading matrix $\mathbf{L} \in \mathbb{M}(25, 9)$. Six components corresponded to features extracted and perceptually validated by Alluri et al. (2012) and were consequently retained for subsequent analysis. The resulting score matrix is denoted by $\mathbf{A} = [\mathbf{a}_c] \in \mathbb{M}(T, 6)$. The vectors $\mathbf{a}_c$ will be subsequently referred to as the acoustic component time series.

A.2. fMRI data preprocessing

Let Σ denote the set of participants. The fMRI time series of participant $s \in \Sigma$ is denoted by the matrix $\mathbf{X}_s \in \mathbb{M}(T, V)$, where V is the number of voxels (in the present study, $V = 228, 453$). Each voxel time series was normalized to unit variance.

Voxel selection

The first step comprises selecting a subset of voxels based on the mean intersubject correlation in the training set. The mean intersubject correlation $\bar{\rho}_v$ for voxel v is denoted by

$$\bar{\rho}_v = \frac{1}{N-1} \sum_{k, j \in \Sigma} \rho_v(k, j)$$

where $\rho_v(k, j)$ denotes the Pearson correlation between the time series in voxel v for participants k and j . For the training set excluding participant s , we denote by $\psi_s(p)$ the subset of voxels whose mean intersubject correlation within the training set belongs to the top 100p % of all voxels,

$$\psi(p) = \{v | \rho_{v-s} > Q(1-p, \rho)\}, 0 \leq p \leq 1,$$

where

$$\bar{\rho}_{v-s} = \frac{1}{N-1} \sum_{\substack{k, j \in \Sigma \\ k, j \neq s}} \rho_v(k, j),$$

and Q is the quantile function of the correlation distribution. The fMRI data of participant s comprising voxels in $\psi(p)$ is denoted by $\mathbf{X}_{sp}$.

The voxelwise mean of fMRI data excluding participant s for voxels in $\psi(p)$ is denoted by

$$\bar{X}_{-sp} := \frac{1}{N-1} \sum_{k \in \Sigma, k \neq s} X_{kp},$$

where N is the number of participants. In the present study we used the values $p \in \{1/64, 1/32, 1/16, 1/8, 1/4, 1/2\}$.

Dimensionality reduction

To reduce dimensionality, each of the matrices $\bar{X}_{-sp}$ was subjected to spatial Principal Components Analysis. As calculating the covariance matrix $\bar{X}_{-sp}^T \bar{X}_{-sp} \in M(V, V)$ was not feasible, given the high number of voxels, the PCA was performed on the transpose matrix using the method described in (Turk and Pentland, 1991). Each of the rows in the data matrix was centered to have a zero mean. Assuming that $\mathbf{v}$ is an eigenvector of $\bar{X}\bar{X}^T \in M(T, T)$, then

$$\begin{aligned} \bar{X}\bar{X}^T \mathbf{v} &= \lambda \mathbf{v} \Rightarrow \\ \bar{X}(\bar{X}\bar{X}^T) \mathbf{v} &= \bar{X}(\lambda \mathbf{v}) \Rightarrow \\ (\bar{X}^T \bar{X}) \bar{X}^T \mathbf{v} &= \lambda (\bar{X}^T \mathbf{v}) \end{aligned}$$

and thus $\bar{X}^T \mathbf{v}$ is an eigenvector of $\bar{X}^T \bar{X}$. Furthermore, all of the variance in the original data is included in the first T PCs. For each participant s , the PCA yields the loading matrix $\mathbf{Y}_{-sp} \in M(T, V)$ and the score matrix $\mathbf{C}_{-sp} = \bar{X}_{-sp} \mathbf{Y}_{-sp}^T \in M(T, T)$.

A.3. Training the Lasso model

For each participant s and each acoustic component c , $\mathbf{C}_{-sp}$ was regressed against $\mathbf{a}_i$ using Lasso regression, in which the regression coefficients are estimated using the following minimizer:

$$\hat{\beta}_{cs}(p, \lambda) = \arg \min \left(\left\| \mathbf{a}_i - \mathbf{C}_{-sp} \beta \right\|^2 + \lambda \|\beta\|_1 \right),$$

where $\|\beta\| = \sum_{k=1}^T |\beta_k|$ and λ is the regularization coefficient. The regression was performed using the Least Angle Regression (LARS) algorithm (Efron et al., 2004). The algorithm produces the entire solution path, in which the value of λ and the number of non-zero regression coefficients vary.

A.4. Prediction and model selection

For each participant s , each acoustic component c was predicted according to

$$\hat{\mathbf{a}}_{cs}(p, \lambda) = \mathbf{X}_{sp} \mathbf{Y}_{-sp}^T \hat{\beta}_{cs}(p, \lambda)$$

Subsequently, for each acoustic component, the optimal parameters p and λ were estimated by maximizing the fit between actual and predicted time series across participants:

$$(\hat{p}_c, \hat{\lambda}_c) = \arg \max (\bar{\rho}_{cp\lambda}) = \arg \max \left(\frac{1}{N} \sum_{s \in \Sigma} (\mathbf{a}_c, \hat{\mathbf{a}}_{cs}(p, \lambda)) \right),$$

where $\rho(\cdot, \cdot)$ denotes the Pearson correlation coefficient.

A.5. Back projection of regression coefficients

To obtain an estimate of each voxel's contribution to the prediction of each acoustic component, the regression coefficients were projected back to voxel space as follows. The fMRI data averaged across all participants $\bar{X}$, was subjected to PCA, yielding the loading matrix $\mathbf{Y}$ and the

score matrix $\mathbf{C}$. The score matrix was regressed against each acoustic component c in turn, using the Lasso model with the optimized parameter values $\hat{p}_c$ and $\hat{\lambda}_c$. Subsequently, to obtain the spatial maps corresponding to the regression coefficients, the obtained regression coefficients were multiplied by the loading matrix according to

$$\mathbf{B}_c = \mathbf{Y}^T \hat{\beta}(\hat{p}_c, \hat{\lambda}_c).$$

A.6. Classification

The stimulus was partitioned into N_κ segments Λ_κ of equal length:

$$\Lambda_\kappa = \left[(\kappa-1) \frac{T}{N_\kappa}, \kappa \frac{T}{N_\kappa}, \kappa = 1, \dots, N_\kappa, \right]$$

where T is the stimulus length. For each segment, the time points in the actual and predicted acoustic components were extracted:

$$\mathbf{a}_c^\kappa = (a_{ic}^\kappa), c = 1, \dots, 6; \kappa = 1, \dots, N_\kappa$$

where t_i is the time corresponding to the i 'th element in the time series.

For subject s and segment κ , the respective fMRI data was classified as the segment that maximized the sum of Pearson correlations between predicted and actual acoustic components:

$$\hat{\kappa}_{sk} = \arg \max_x \left[\sum_c \rho(\mathbf{a}_c^x, \mathbf{a}_{sc}^\kappa) \right]$$

References

- Alluri, V., Toivainen, P., Jääskeläinen, I.P., Glerean, E., Sams, M., Brattico, E., 2012. Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage* 59 (4), 3677–3689.
- Alluri, V., Toivainen, P., Lund, T.E., Wallentin, M., Vuust, P., Nandi, A.K., Brattico, E., 2013. From Vivaldi to Beatles and back: predicting lateralized brain responses to music. *NeuroImage* 83, 627–636.
- Brattico, E., Alluri, V., Bogert, B., Jacobsen, T., Vartiainen, N., Nieminen, S., Tervaniemi, M., 2011. A functional MRI study of happy and sad emotions in music with and without lyrics. *Front. Psychol.* 2. <http://dx.doi.org/10.3389/fpsyg.2011.00308> (Article 308).
- Chapin, H., Jantzen, K., Kelso, J.A.S., Steinberg, F., Large, E., 2010. Dynamic emotional and neural responses to music depend on performance expression and listener experience. *PLoS One* 5 (12), e13812.
- Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* 19 (2), 261–270.
- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. *Ann. Stat.* 32 (2), 407–451.
- Formisano, E., De Martino, F., Valente, G., 2008. Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. *Magn. Reson. Imaging* 26 (7), 921–934.
- Fraisse, P., 1982. Rhythm and tempo. In: Deutsch, D. (Ed.), *The psychology of music*. Academic Press, New York, pp. 148–180.
- Friston, K.J., Josephs, O., Zarahn, E., Holmes, A.P., Rouquette, S., Poline, J.B., 2000. To smooth or not to smooth? Bias and efficiency in fMRI time-series analysis. *NeuroImage* 12 (2), 196–208.
- Janata, P., 2009. The neural architecture of music-evoked autobiographical memories. *Cereb. Cortex* 19 (11), 2579–2594.
- Koelsch, S., 2010. Towards a neural basis of music-evoked emotions. *Trends Cogn. Sci.* 14 (3), 131–137. <http://dx.doi.org/10.1016/j.tics.2010.01.002>.
- Kriegeskorte, N., 2011. Pattern-information analysis: from stimulus decoding to computational-model testing. *NeuroImage* 56 (2), 411–421.
- Lartillot, O., Toivainen, P., 2007. MIR in Matlab (II): a toolbox for musical feature extraction from audio. Proceedings of 8th International Conference on Music Information Retrieval 2007 (Available online at http://ismir2007.ismir.net/proceedings/ISMIR2007_p127_lartillot.pdf).
- Lehne, M., Rohrmeier, M., Koelsch, S., 2013. Tension-related activity in the orbitofrontal cortex and amygdala: an fMRI study with music. *Soc. Cogn. Affect. Neurosci.* <http://dx.doi.org/10.1093/scan/nst141>.
- McIntosh, A., Lobaugh, N., 2004. Partial least squares analysis of neuroimaging data: applications and advances. *NeuroImage* 23, S250–S263.
- Mueller, M., Ellis, D.P.W., Klapuri, A., Richard, G., 2011. Signal processing for music analysis. *IEEE J. Sel. Top. Sig. Process.* 5 (6), 1088–1110.
- Nasrilaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. *NeuroImage* 56 (2), 400–410.

- Pereira, C.S., Teixeira, J., Figueiredo, P., Xavier, J., Castro, S.L., Brattico, E., 2011. Music and emotions in the brain: familiarity matters. *PLoS One* 6 (11), e27241.
- Smith, A.M., Lewis, B.K., Ruttimann, U.E., Ye, F.Q., Sinnwell, T.M., Yang, Y., Duyn, J.H., Frank, J.A., 1999. Investigation of low frequency drift in fMRI signal. *NeuroImage* 9, 526–533.
- Turk, M., Pentland, A., 1991. Eigenfaces for recognition. *J. Cogn. Neurosci.* 3 (1), 71–86.
- Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* 15 (1), 273–289. <http://dx.doi.org/10.1006/nimg.2001.0978>.
- Valente, G., Castellanos, A.L., Vanacore, G., Formisano, E., 2013. Multivariate linear regression of high-dimensional fMRI data with multiple target variables. *Hum. Brain Mapp.* <http://dx.doi.org/10.1002/hbm.22318>.