Robust inter-subject audiovisual decoding in functional magnetic resonance imaging using high-dimensional regression

NeuroImage 163 (2017) 244–263

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Robust inter-subject audiovisual decoding in functional magnetic resonance imaging using high-dimensional regression

Gal Raz [a,b,c,*], Michele Svanera [d], Neomi Singer [a,c,e], Gadi Gilam [a,e], Maya Bleich Cohen [a], Tamar Lin [a], Roee Admon [f], Tal Gonen [a,g], Avner Thaler [a,c,h,i], Roni Y. Granot [j], Rainer Goebel [k], Sergio Benini [d], Giancarlo Valente [k]

[a] Functional Brain Center, Wohl Institute for Advanced Imaging, Tel Aviv Sourasky Medical Center, 64239 Tel Aviv, Israel
[b] Film and Television Department, Tel Aviv University, 69978 Tel Aviv, Israel
[c] Sagol School of Neuroscience, Tel Aviv University, 69978 Tel Aviv, Israel
[d] Department of Information Engineering, University of Brescia, 38, 25123 Brescia, Italy
[e] School of Psychological Sciences, Tel Aviv University, 69978 Tel Aviv, Israel
[f] Department of Psychology, University of Haifa, 3498838 Haifa, Israel
[g] Department of Neurosurgery, Tel Aviv Sourasky Medical Center, 64239 Tel Aviv, Israel
[h] Movement Disorders Unit, Neurological Institute, Tel-Aviv Sourasky Medical Center, 64239 Tel Aviv, Israel
[i] Sackler Faculty of Medicine, Tel Aviv University, 69978 Tel Aviv, Israel
[j] Musicology Department, Hebrew University of Jerusalem, 9190501 Jerusalem, Israel
[k] Department of Cognitive Neuroscience, Maastricht University, 6211 LK Maastricht, The Netherlands

ARTICLE INFO

Keywords:
fMRI
Audiovisual decoding
Motion pictures
Kernel ridge regression
Sound loudness
Optical flow
Face
Motion pictures

ABSTRACT

Major methodological advancements have been recently made in the field of neural decoding, which is concerned with the reconstruction of mental content from neuroimaging measures. However, in the absence of a large-scale examination of the validity of the decoding models across subjects and content, the extent to which these models can be generalized is not clear. This study addresses the challenge of producing generalizable decoding models, which allow the reconstruction of perceived audiovisual features from human magnetic resonance imaging (fMRI) data without prior training of the algorithm on the decoded content. We applied an adapted version of kernel ridge regression combined with temporal optimization on data acquired during film viewing (234 runs) to generate standardized brain models for sound loudness, speech presence, perceived motion, face-to-frame ratio, lightness, and color brightness. The prediction accuracies were tested on data collected from different subjects watching other movies mainly in another scanner.

Substantial and significant (Q_{FDR}<0.05) correlations between the reconstructed and the original descriptors were found for the first three features (loudness, speech, and motion) in all of the 9 test movies ($\overline{R}$=0.62, $\overline{R}$ = 0.60, $\overline{R}$ = 0.60, respectively) with high reproducibility of the predictors across subjects. The face ratio model produced significant correlations in 7 out of 8 movies ($\overline{R}$=0.56). The lightness and brightness models did not show robustness ($\overline{R}$=0.23, $\overline{R}$ = 0). Further analysis of additional data (95 runs) indicated that loudness reconstruction veridicality can consistently reveal relevant group differences in musical experience.

The findings point to the validity and generalizability of our loudness, speech, motion, and face ratio models for complex cinematic stimuli (as well as for music in the case of loudness). While future research should further validate these models using controlled stimuli and explore the feasibility of extracting more complex models via this method, the reliability of our results indicates the potential usefulness of the approach and the resulting models in basic scientific and diagnostic contexts.

* Corresponding author. Kfar Daniel 253, 7312500, Israel.
E-mail address: galraz@post.tau.ac.il (G. Raz).

https://doi.org/10.1016/j.neuroimage.2017.09.032
Received 3 May 2017; Received in revised form 14 September 2017; Accepted 17 September 2017
Available online 20 September 2017

1. Introduction

"Mind reading" based on neural decoding is an ambitious line of research within contemporary neuroscience. Assuming that certain psychological processes and mental contents may be encoded in the brain as specific and consistent patterns of neural activity, researchers in this field aim to decode and reconstruct them given only the neuroimaging data. In order to "read" stimuli out of one's brain, researchers adopt different machine learning approaches and apply various pattern analysis methods that link local or distributed neural activity patterns with specific audiovisual features.

Neural decoding refers to the prediction of a stimulus features from measured brain activity (Schoenmakers et al., 2013) (fMRI data in our case). Several notable neural decoding achievements have been reported so far, mainly in studies employing functional magnetic resonance imaging (fMRI), but also in intracranial recording and electro- and magneto-encephalography experiments (for review, see Chen et al., 2013; Haxby, 2012). Reported decoding classification accuracies for out-of sample data commonly range between 70 and 90% (see Poldrack et al., 2009), and correlation as high as 0.99 between predicted and observed continuous descriptors was demonstrated (Valente et al., 2011). Decoding targets vary and include mental states such as action intentions (Haynes et al., 2007), reward assessment (Kahnt et al., 2011) and response inhibition (Cohen et al., 2010; Poldrack et al., 2009); low-level features such as visual patterns in dynamic video (Nishimoto et al., 2011), geometrical patterns, text (Fujiwara et al., 2009; Miyawaki et al., 2008; van Gerven et al., 2010), and optical flow acceleration in a video game (Chu et al., 2011; Valente et al., 2011); and semantic elements such as animal and objects categories (Connolly et al., 2012; Haxby et al., 2001, 2011), objects and actions in a hierarchical semantic space (Huth et al., 2012), visual imagery content during sleep (van Gerven et al., 2010), and actions and events in a video game (Chu et al., 2011; Valente et al., 2011).

This productive stream of research supports the appealing vision of generating a repertoire of "fMRI fingerprints" for a wide range of mental states and perceptual processes (or "cognitive ontology", see Poldrack et al., 2009). Ideally, such repertoire will facilitate robust and rich neural decoding for any subject independently of prior training and using any standard MRI scanner. The generation of a reliable repertoire of this kind is valuable both in terms of basic science (providing a reproducible and comprehensive ground truth for brain-function mapping) and applicable technology (in diagnosis, for instance; see Cohen et al., 2011).

Strong evidence for the generalizability of such repertoire of functional models of the brain can be gained by demonstrating their performance under conditions of high heterogeneity across the training and the test data. For this end, it is necessary to show that these models facilitate successful decoding also when analyzing stimuli that are different from those on which the algorithm was trained. However, eminent neural decoding achievements were gained using a within-subject design including only five subjects or less (e.g., Horikawa et al., 2013; Huth et al., 2012; Miyawaki et al., 2008; Nishimoto et al., 2011), which limits the examination of the reproducibility of the results. Thus, the key aspect of inter-subject generalizability of neural decoding has yet to be systematically investigated (Chen et al., 2013).

Confronting the limitations of the within-subject design in neural decoding, Haxby and colleagues have recently demonstrated the feasibility of between-subject classification. This group developed methods for cortical anatomy alignment for different subjects based on the maximization of the inter-subject similarity of blood oxygen level dependent (BOLD) reaction patterns (Haxby et al., 2011; Sabuncu et al., 2010) and functional connectivity structures (Conroy et al., 2013). These studies have demonstrated that between-subjects classification may yield success rates equivalent to those of within-subject classification. Successful decoding of data of out-of-sample individual was also reported in few other studies that did not implement inter-subject alignment methods that rely on functional data. Shinkareva et al. (2008) and Poldrack et al. (2009), reached average accuracy rates of ~80% in classifying visual input and task type, respectively. Cohen and colleagues (Cohen et al., 2011) decoded response inhibition related variables with above-chance correlation values of 0.4–0.5 between the predicted and real parametric values.

In keeping with the notion that a compelling validation of neural decoding method relies on its success under highly heterogeneous conditions, the current work introduces a markedly increased variability across several experimental dimensions. First, we aimed to decode continuous time-varying features, which change on a moment-to-moment basis. Second, we tested the decoding reliability on a set of different movies, comprising highly heterogeneous, complex, and naturalistic stimuli. Lastly, the validation of the function-brain models was performed using movies that were not employed in the training procedure with data collected in a different MRI scanner from un-tested individuals. An overview of the study is given in Fig. 1.

We combined data from 234 movie-viewing sessions (with 5 different clips) for the training of our algorithm and the cross-validation of the resulting model (Table 1). The validity of the models was tested using an independent sample of 63 sessions (with 9 other clips). We selected relatively coarse features across three elementary perceptual domains: audio, vision and motion. The selected features were sound loudness (loudness), speech presence (speech), detected motion (motion), face-to-frame dimension ratio (face ratio), perceived lightness, and brightness. These audiovisual features were extracted using both manual and automatic annotation tools.

In order to decode these continuous features from the fMRI data we used linear kernel ridge regression (KRR) with generalized cross-validation (GCV) (Golub et al., 1979). We chose a kernel version since it is particularly efficient when the number of data points is considerably lower than the number of measurable properties, or features (in our case time repetitions and number of voxels, respectively; see Golub et al., 1979). The combination of L2-norm penalization with GCV is relatively computationally inexpensive when compared with iterative kernel methods such as Relevance Vector Regression and Gaussian Processes, while still achieving good performance, and it allows for fast permutations in order to ascertain non-parametrically statistical significance (Valente et al., 2014). We used a linear kernel for several reasons. First, by using a linear model we allow for the reconstruction of descriptors of specific features as linear combinations the weighted BOLD time series, with the advantage of a straightforward interpretation of the fMRI models relative to non-linear kernels and other complex pattern recognition methods such as artificial neural networks. Second, the optimization of non-linear kernel hyperparameters would increase the computational time of several orders of magnitude. Finally, the large number of dimensions, compared to the available samples in our problem, makes it difficult to exploit the increased flexibility provided by non-linearities, increasing the risk of overfitting. An alternative to linear kernel ridge regression could be to use a linear ridge regression after projecting the data onto the subspace spanned by the principal components, which would result in similar computational burden if all the principal components are retained.

In addition to the extraction of spatial brain models of continuous audiovisual features, we temporally optimized the models. In specific, we applied time-lag optimization following evidence that multi-voxel pattern analysis (MVPA) classification may be improved by fitting different temporal hemodynamic response models to different brain regions (Kohler et al., 2013). Due to the high dimensionality of the problem we used simulated annealing (Kirkpatrick et al., 1983), a heuristic algorithm based on thermodynamic principles, to optimize the temporal parameters of the decoding models. This procedure was performed on a cross-validation subset of the data.

Thus, we produced spatio-temporal decoding maps, which assign optimized lag and weight values to every voxel in the brain to reconstruct specific features. Our study examined the extent to which various audiovisual features can be robustly and reliably reconstructed by these

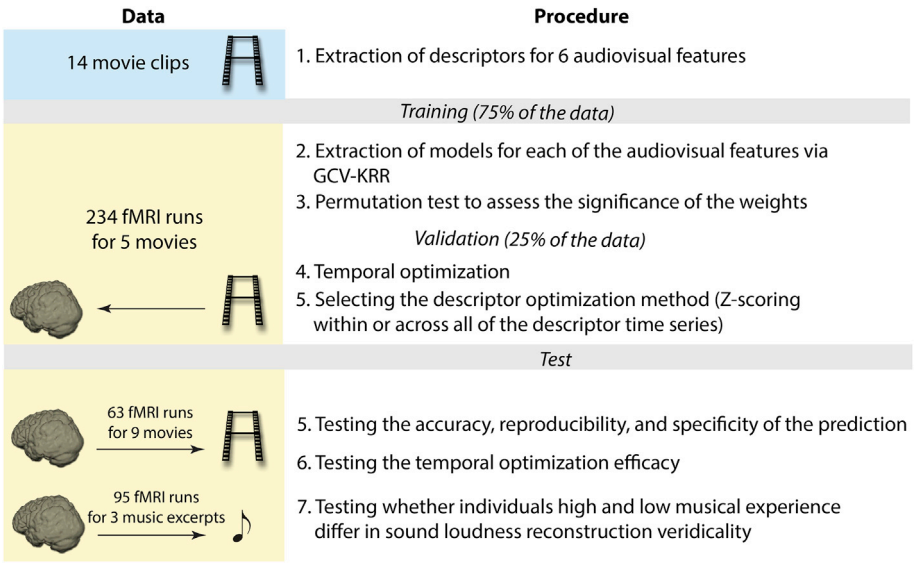

Fig. 1. The study outline.

methods for cinematic stimuli given different subjects and experimental context. We further explored the functional significance of the specific spatio-temporal brain models resulting from our analysis.

Finally, we were interested in the potential of generalized fMRI models to reveal functionally meaningful inter-group differences. In specific, we hypothesized that the loudness descriptor will be reconstructed more accurately when applying the loudness model to data obtained from a group of individuals with higher musical experience. This hypothesis relies on the assumption that musical training may increase the sensitivity to loudness modulations since musicians are commonly required to perform subtle manipulations of this feature (Bishop et al., 2013), which has a key role in musical expression (Juslin and Laukka, 2003). Indeed, musicians show higher ability to detect transient silent gaps (~3ms) in a quasi-continuous sound (Rammsayer and Altenmüller, 2006; Zendel and Alain, 2012), indicating exquisite sensitivity to the envelop of sound. Moreover, recent evidence suggests that loudness time courses can be reconstructed more accurately by expert musicians during a musical imagery task (Bishop et al., 2013). We hypothesized that (a) a reliable predicted descriptor will be generated based on the loudness model and fMRI data recorded during music listening (i.e., generalization of the model from audiovisual to auditory paradigms); (b) greater veridicality of the predicted loudness descriptor will be obtained for the group of individuals with higher musical experience (i.e., the modeling will be sensitive to group differences).

2. Materials and methods

2.1. Stimuli and data collection

All data were collected from healthy volunteers without known history of neurological or psychiatric disorder. The participants had at least 12 years of education with Hebrew as their spoken language. They signed a consent form approved by the ethical committees of the Tel Aviv Sourasky Medical Center. For demographic details, see Table 1. For an overview of the study, see Fig. 1.

In the movie conditions, the participants were instructed to passively view excerpts from commercial movies that were presented to them using an LCD projector. In the music condition, the participants passively listened to musical excerpts. We trained the neural decoding algorithm and tested its validity using two independent samples (Table 1):

2.1.1. Training set

To generate the decoding models for the audiovisual features of interest, a training data set was created by pooling 234 fMRI scans from different studies (including Raz et al., 2013, 2012; few scans were added after publication). In several cases, the same participant watched two of the movie clips. Thus, for 23 subjects we included data recorded during the viewing of both Sophie's Choice and Stepmom clips. Valid data for The Ring and The X-Files were collected from the same individual in four cases. Due to technical failures (lagging video playback and playback errors) and exaggerated head motions (deviations higher than 1.5 mm and 1.5° from the reference point), 27, 20, 21, 3, and 6 runs were discarded in the cases of Avenge, Sophie, Stepmom, The Ring 2, and The X-Files, respectively. Table 1 summarizes relevant details on the participants, movie contents, and sample sizes.

2.1.2. Testing set

(i) **Movie viewing**: The testing data set included 63 scans that were obtained during the viewing of 9 movie clips. Data for the clips Black Swan and The Fly were collected from two independent groups including 8 and 20 volunteers, respectively. Five and two scans were discarded due to exaggerated head motions (threshold: 1.5 mm and 1.5°), respectively. In this case, the acquisition was performed using the same scanner that was used to obtain the training data. We selected 7 additional clips specifically for the validation procedure to test the robustness of the neural decoding across various contexts. For this aim, we relied on a database of movie clips classified into different genres based on emotional annotation by a large sample group (Schaefer et al., 2010). One clip was selected from each of the six emotional categories examined in this study. These clips were taken from the films Dead Poet Society, Forrest Gump, Saving Private Ryan, Se7en, The Shining, and There is Something About Mary, representing the categories Sadness, Tenderness, Disgust, Anger, Fear, and Amusement, respectively. A seventh excerpt, taken from the documentary film Denali, was added as a neutral clip (following Rottenberg et al., 2007). These seven clips were displayed to five participants in another MRI scanner.

(ii) **Music listening**: The participants passively listened to three recorded piano pieces: (1) Ricercatas 1&2 of György Ligeti's Musica Ricercata (hereafter referred to as LM; 7:41 min), (2) Philip Glass's "Hours" (GH, 7:03 min) from the soundtrack of the film *The Hours*. (3) Modest Mussorgsky's *Night on Bald Mountain* (MN; 10:57 min). Data for the first two pieces were previously analyzed in a different context (Singer et al., 2016). Four, two, and

Table 1
Details on materials and samples used in the study. The sample size indicates the net number of scans after dropout due to technical reasons and head motions.

Movie Data							
Movie details				Sample details			
Training set							
Film title	**Duration (min)**	**Theme**	**Relevant Reference**	**Sample size**	**Dropout**	**Average ± std age (years)**	**Females/ males**
Avenge But One of My Two Eyes (Mograbi, 2006)	5:27	A political activist confronts with Israeli soldiers	(Raz et al., 2016)	74	27	19.51 ± 1.45	0/74
Sophie's Choice (Pakula, 1982)	10:00	A mother is forced to choose which of her two children will be taken from her	(Raz et al., 2012)	44	20	26.73 ± 4.69	25/19
Stepmom (Columbus, 1998)	8:21	A mother talks with her children about her future death	(Raz et al., 2012)	53	21	26.75 ± 4.86	21/32
The Ring 2 (Nakata, 2005)	8:15	A child is lost in a bazaar; The child and his mother are attacked by deer	(Raz et al., 2016)	27	3	26.41 ± 4.12	11/16
The X-Files, the episode "Home" (Manners, 1996)	5:00	Zombies attack a couple in their home	(Raz et al., 2016)	36	6	23.70 ± 1.23	14/22
Testing set							
Film title	**Duration (min)**	**Theme**	**Relevant Reference**	**Sample size**	**Dropout**	**Average ± std age (years)**	**Females/ males**
Alaska's Wild Denali (Thomas, 1997)	5:00	Nature documentary with narration	(Rottenberg et al., 2007)	5	0	26.6 ± 4.33	4/1
Black Swan (Aronofsky, 2010)	9:00	A ballet dancer experiences a series of hallucinations	NA	8	2	31.63 ± 8.1	3/5
Dead Poet Society (Weir, 1989)	5:18	Parents find out that their son committed a suicide	(Schaefer et al., 2010)	5	0	26.6 ± 4.33	4/1
Forrest Gump (Zemeckis, 1994)	5:21	The protagonist is introduced to his unknown son for the first time	(Schaefer et al., 2010)	5	0	26.6 ± 4.33	4/1
Saving Private Ryan (Spielberg, 1998)	6:18	American troops landing on Omaha Beach during World War II.	(Schaefer et al., 2010)	5	0	26.6 ± 4.33	4/1
Se7en (Fincher, 1995)	6:18	A murdered tells a detective that he beheaded his pregnant wife	(Schaefer et al., 2010)	5	0	26.6 ± 4.33	4/1
The Fly (Cronenberg, 1986)	8:15	A man is transformed into a giant fly after conversation with his former lover and attacking her friend	NA	20	5	42.55 ± 7.47	8/12
The Shining (Kubrick, 1980)	5:21	A man pursues his wife with an axe	(Schaefer et al., 2010)	5	0	26.6 ± 4.33	4/1
There is Something About Mary (Farrelly and Farrelly, 1998)	5:00	A man fights with his girlfriend's dog	(Schaefer et al., 2010)	5	0	26.6 ± 4.33	4/1
Music Data							
Music piece title	**Duration (min)**	**Author**	**Relevant Reference**	**Sample size**	**Dropout**	**Average ± std age (years)**	**Females/ males**
Musica Ricercata	7:41	György Ligeti	(Singer et al., 2016)	32	8	25.8 ± 3.71	18/14
Hours	7:03	Philip Glass	(Singer et al., 2016)	33	7	25.56 ± 3.67	19/14
Night on Bald Mountain	10:57	Modest Mussorgsky	(Singer et al., 2016)	30	10	25.82 ± 3.7	17/13

six scans were discarded due to excessive head motions (<1.5 mm or 1.5°) in LM, GH, and MN, respectively. Data of two subjects were excluded due to cyclic noise, probably resulting from a technical failure (as in Singer et al., 2016). The scanning was terminated for two other subjects due to claustrophobia. One GH session was not completed due to technical issues.

2.2. Image acquisition and preprocessing

The data for this study were collected using two scanners located at the Tel-Aviv Sourasky Medical Center. The training and the testing data for two of the movies (Fly and Black Swan) and all of the music conditions were acquired by a 3 T Signa Excite scanner (GE Medical Systems, WI, USA) with an 8-channel head coil. All other testing data were obtained via 3T Siemens system (MAGNETOM Prisma, Germany) with a 20-channel head coil. Active noise canceling headphones (Optoacoustics) were used during the scans.

2.2.1. Movie data

Identical scanning parameters were set across the scans in both scanners. For structural scanning, we used a T1-weighted 3D axial spoiled gradient echo (SPGR) pulse sequence with the following parameters: TR/TE = 7.92/2.98 ms, slice thickness = 1 mm, flip angle = 15°, pixel size = 1 mm, FOV = 256 × 256 mm. Functional whole-brain scans were performed in interleaved order with a T2*-weighted gradient echo planar imaging pulse sequence (time repetition [TR]/TE = 3000/35 ms, flip angle = 90, pixel size = 1.56 mm, FOV = 200 × 200 mm, slice thickness = 3 mm, 39 slices per volume).

2.2.2. Music data

The scans were performed using a GE 3T scanner as specified above. T1-weighted SPGR parameters were TR/TE = 8.9/3.5 ms, flip angle = 13°, voxel size = 1 × 1 × 1 mm, FOV = 256 × 256 mm, slice thickness = 1 mm. The recording and preprocessing of the T2*-weighted images were identical to the procedures applied on the movie data except for a FOV of 220 × 220 mm, and a slice number of 38 for four scans (of 2 participants) due to technical errors.

We pre-processed and registered all data to a standardized Talairach anatomical template via Brainvoyager QX version 2.3 (Brain Innovations, Maastricht, Netherlands) with manual verification of the automated corregistration. For the detection and correction of head motions we used trilinear and sinc interpolations respectively, applying rigid body transformations with three translation and three rotation parameters. High pass filtering of 3 cycles per time course and spatial smoothing with a 6 mm FWHM kernel were applied. Physiological noise correction was performed by regressing out the mean ventricles signal from the blood-oxygen-level dependent (BOLD) time course. An ICBM 452 probability map (http://www.loni.usc.edu/atlases) for the ventricles (thresholded at 99%) was used to generate the mask. We confined all analyses to 42309 voxels included in a gray matter mask. The mask was created by thresholding ICBM 452 map to exclude voxels with probability lower than 80% of being classified as gray matter (thus encompassing both cortical and brain stem regions).

Data were discarded due to exaggerated head motions in case of deviations higher than 1.5 mm and 1.5° from the reference point. For the total numbers of scans that were dropped out due to head motions and various reasons, see Table 1.

2.3. Extracting audiovisual descriptors

Six features were annotated for all movies: color brightness, perceived lightness, face-to-frame ratio, motion detection, sound loudness, and speech presence. Loudness was annotated for the three music pieces as well.

2.3.1. Color brightness

Brightness measures the perception elicited in a human by the luminance emitted or reflected by a visual target, meaning that there is a distinction between the physical intensity of the light by an object (i.e., luminance), and the description of how the human visual system perceives it (i.e., brightness). Brightness is then a subjective measure in the sense that it describes the color "sensation" with respect to a standard human observer. Brightness directly appears in different representations of color spaces: for example in HSV color space (hue, saturation, and brightness or value) it appears as the third color coordinate or in the RGB color space as the arithmetic mean of the red, green, and blue color coordinates. In this work it is computed as the mean value of pixel intensities extracted from images obtained at one frame per second and then averaged to match the BOLD temporal resolution.

2.3.2. Perceived lightness

Even when they share the same luminance, colored lights seem brighter than white light with the same luminance. This perceptual effect, known as the Helmholtz-Kohlrausch effect, states the difference between brightness and lightness: brightness is the intensity of the object independently from the light source (bottom row of Fig. S1), while lightness is the brightness of the object in respect to the light reflecting on it (top row of Fig. S1). In order to capture this perceptual amplification of color, we make use of the Retinex theory (Land and McCann, 1971; Morel et al., 2010), which aims at reproducing the sensory response to color stimuli by the human visual system. This model was developed starting from the assumption that actual color sensations are related to the intrinsic reflectance of objects rather than to the radiance values captured by the eyes (Bertalmío et al., 2009). Therefore the application of Retinex should turn a generic input picture into an image closer to what a human observer would perceive if she were looking at the same scene when the picture was taken. In this work we use a fast implementation of Retinex proposed by Limare et al. (2011) to derive a perceived lightness descriptor which accounts for the chi-squared distance between image histograms computed on video frames before and after the application of Retinex algorithm.

2.3.3. Motion detection

To transmit a sensation of speed and dynamism or a feeling of quiet and tranquility in a movie scene, directors may rely on camera and object motion. In order to capture motion dynamics we extract a map which is descriptive of the moving objects by adopting the algorithm by Barnich and Droogenbroeck (2011). This work extends the idea of background subtraction, which requires the pixel-wise comparison between the static background with the current frame, by storing for each pixel a set of values taken in the past at the same location and in its neighbourhood. While other analyses in this work are carried out on individual frames, the evaluation of motion is conducted directly on video data. The background model is updated every frame, and a binary motion map is obtained every five frames.

2.3.4. Face-to-frame ratio

To estimate face-to-frame ratio, we used a method proposed by Zhu and Ramanan in (Zhu and Ramanan, 2012), which is highly effective in capturing global elastic deformation of faces, ensuring high recall rates also from non-frontal views. This algorithm employs a tree structure composed by a set of parts (e.g., histogram of oriented gradients descriptor) modeling every facial landmark as a part, and uses global mixtures to model topological changes due to viewpoint. Authors allow different mixtures to share part templates, so that it is possible to model a large number of views with low complexity. Finally, all parameters of this model, including part templates, modes of elastic deformation, and view-based topology, are discriminatively trained in a max-margin framework.

We annotated only information about the ratio of the largest face in the image to the total area of the frame, assuming that the viewer

attention is mainly captured by the character at the closest camera distance. Faces were detected each second, and the final descriptor was averaged over 3-s intervals to fit the BOLD signal resolution. Since the excerpt from the film Denali included no human face, a face-to-frame ratio descriptor was not produced for this film.

2.3.5. Sound loudness

Sound loudness, originally defined by Fletcher and Munson (1933), describes the relationship between the measured sound pressure level and the perceived intensity of the sound from a "standard" human ear system. As the ear is less sensitive to low audio frequencies, A-weighting filtering (http://soundmetersource.com/iec-61672-1.html, retrieved 2013-04-29) is applied to the frequency spectrum of the audio signal.

The frequency attenuation of the A-weighting filter corresponds to an empirical average obtained across a broad sample of perceptual experiments. Fast Fourier transform (FFT) is applied on the audio signal to estimate the frequency spectrum, which is then weighted using a closed-form expression for the A-weighting filter. In order to determine the signal level in dBA, the total signal energy is then integrated by calculating the energy within each data window by applying Parseval's relation in the frequency-domain.

2.3.6. Speech presence

Despite considerable advancements in automatic speech detection and discrimination (e.g., Panagiotakis and Tziritas, 2005; Pikrakis and Theodoridis, 2014) our attempts to apply such procedures to the complex soundtracks of the selected clips yielded disappointing results. Therefore, speech presence was manually annotated to produce binary indices for the existence of an utterance within 1-s time bins. The annotated time courses were down-sampled to fit the BOLD temporal resolution by computing the median value for each three subsequent time bins (with no overlap). The resampled annotation, which was produced by a single rater, showed considerably high reliability compared with an independent rater's annotation for seven of the testing set clips: identical values between annotations were found in 93.7% of the time points. Therefore, further analyses relied on annotations by the first rater.

2.4. Data analysis

The training data set was randomly split into two subsets. Spatial decoding and permutation test were performed on the larger subset, which included 75% of the data. The other subset was used for temporal optimization, and the selection of normalization method.

2.4.1. Generating spatial decoding model

Given the high dimensions of the problem, we applied high-dimensional multi-voxel regression using a linear kernel (Valente et al., 2011). The linear kernel remaps the original temporally concatenated fMRI data X, of size n (time-points) × s (subjects)-by-p (voxels), into a kernel matrix K:

$$K = XX^T \tag{1}$$

of dimension n × s-by-n × s, where the regression is carried out. The original problem of linking targets and fMRI data is therefore recast in the following linear regression:

$$y = K\mathrm{w} + \varepsilon \tag{2}$$

where y is the target n × s dimensional vector and K is the kernel matrix. In our case, y is the target descriptor (e.g., loudness time series), and K is built from X, the temporal concatenation across subjects of Z-scored time-series, as shown in equation (1); w is an n × s dimensional vector that weights the different fMRI volumes to predict y. With this approach, the regression problem is more tractable, since it is done in a space of dimension n × s, whereas in the original space regression would be performed in a space of dimension p, with p > n × s (in our work, p and n × s are roughly 43,000 and 25,000 respectively).

When new data X_1 are used, the predicted target y_1 is estimated by applying the linear model w on the new kernel $K_1 = X_1X^T$, resulting in:

$$y_1 = K_1\mathrm{w} = X_1X^Tw = X_1\beta \tag{3}$$

where the p-dimensional vector $\beta = X^TW$ is used to predict a target based on the original fMRI data and can be used to create predictive brain maps.

We assessed the similarity between the reconstructed time series and the target descriptor using Pearson's correlation coefficient:

$$R_f = r\left(y_f\ , X\beta_f\right) \tag{4}$$

where y_f and β_f are the time series and the weights vector for a specific feature f, respectively. Since the prediction is carried out using new subjects, new movies and in some cases different scanners, the first and second order statistics of both the BOLD time series and the targets could be different between training and testing data. We therefore z-scored training and testing fMRI data and targets separately and did not aim at reconstructing the descriptor in absolute terms, hence the choice of a similarity measure such as Pearson's correlation coefficient (that was employed in the context of multivariate regression in the Pittsburgh Brain Activity Interpretation Competition, PBAIC 2007 (Valente et al., 2011)), complemented by a non-parametric permutation test to determine significant differences from chance.

In the training and the validation data sets, X and y_f were concatenated over scans so that the total number of time points n ranged from 23723 to 25482 and from 7014 to 8773 (depending on the random assignment of scans to the groups), respectively. The descriptors could be Z-scored either before or after concatenation. The normalization method that yielded higher R_f in the cross-validation group was selected. Thus, Z-scoring was performed after concatenation in the case of motion and face ratio descriptors, and before concatenation in all other cases.

The estimation of w with least squares is prone to overfitting (Bishop, 2007), and regularization should be employed. Among the possible choices, ridge regression is a computationally attractive and yet powerful method to estimate a weighting w that achieves good generalization performance; in the estimation a quadratic term is added to penalize solutions with large weights (which likely result from overfitting):

$$y = K\mathrm{w} + \lambda||\mathrm{w}||^2 + \varepsilon \tag{5}$$

where $||\mathrm{w}||^2$ denotes the L2-norm of w. The estimation of this new model leads to a close-form solution dependent on the regularization strength λ:

$$\widehat{\mathrm{w}}(\lambda) = (K^TK + \lambda I)^{-1}K^T\mathrm{y} \tag{6}$$

where I is a n × s by n × s identity matrix.

To optimize the regularization parameter λ we used generalized cross validation (GCV-KRR); (Golub et al., 1979), which is based on the minimization of a cost function weighting simultaneously both data fit (numerator) and model complexity (denominator):

$$V(\lambda) = \frac{\frac{1}{n}\left|\left| I - A(\lambda)y \right|\right|^2}{\left[\frac{1}{n}\ Tr(I - A(\lambda))\right]^2} \tag{7}$$

where Tr denotes the trace and

$$A(\lambda) = K(K^TK + \lambda I)^{-1}K^T \tag{8}$$

We empirically tested $\widehat{\lambda}$ in the range of $10^7 \le \widehat{\lambda} \le 10^{12}$ with $\widehat{\lambda}$ changing logarithmically in increments of 0.5. For all of the audiovisual features we found the global minimum of $V(\lambda)$ to fall within this range. It

is worth mentioning that the GCV procedure uses all the training data and does not necessitate of a nested loop to optimize $\hat{\lambda}$. This has the advantage of selecting the suitable regularization for the amount of examples available: since the amount of regularization depends on the available training data, a nested cross-validation procedure would tend to select a stronger regularization since fewer examples are available for training, whereas using all the training data does not suffer from this problem.

To assess the quality of the decoding, we first reconstructed a target time series $X_{s,m}\beta_f$ for every subject *s*, movie *m*, and feature *f* in the test set. We then computed Pearson's correlations to compare the individual and the average prediction over subjects $(\overline{X_m}\beta_f)$ with the target time series.

We performed a voxel-wise testing of the weights in the resulting maps using a non-parametric permutation test. For each descriptor we generated 10,000 corresponding scrambled time series by implementing a Daubechies mother wavelet (Daubechies et al., 1992) on the signal and decomposing it into 7 levels with 4 vanishing moments. This type of decomposition, applied on the descriptor after it was convolved with an estimated HRF, keeps the autocorrelation structure of the original signal (Bullmore et al., 2001). In each level, we resampled the details coefficients and reconstructed the time series using inverse wavelet transform. By applying GCV-KRR on the training data with the permuted descriptors as the target time series and the optimal λ per descriptor, we generated 10,000 models whose weights served as null-distribution for each voxel. Based on these null distributions, we thresholded the original weight map using a double-sided test and correcting for multiple comparisons at $q_{FDR}<0.05$.

2.4.2. Testing the decoding accuracy

To estimate the significance of the correlation between the reconstructed predictors and the "ground truth" descriptors, we created a background distribution of Pearson's coefficients. For each feature *f* and each of the 10,000 models generated in the previous step we computed a predicted time series $\overline{X}_c\beta_f$ and $X_{s,c}\beta_f$, and correlated them with the descriptor time series $y_{f,c}$. The p value was defined as the proportion of Pearson's coefficients in this null distribution that are higher or equal to the coefficient obtained for the original comparison between the predicted and the descriptor's time series (adding one to both numerator and denominator to avoid zero p-values when no permutation equals or exceeds the observation). To assess the reproducibility of the reconstruction over movies, we conducted a partial conjunction analysis (Heller et al., 2007) considering all of the eight (in the absence of human face ratio annotation) or nine p-values for each of the features. This analysis tests the proportion of conditions (in our case – movies) that show a real effect in a way which is valid under dependence between conditions. The results were FDR corrected (Heller et al., 2007).

2.4.3. Examining the relations between decoding accuracy and training data parameters

We conducted further analyses to examine the accuracy of GCV-KRR that relies on smaller chunks of the training data. These analyses can be informative when considering the size of data required for future applications of the algorithm in fMRI.

First, we repeated the training and testing procedure described above (i.e., generating models based on the training data without temporal optimization and testing them using the nine testing movies), but with a varying size of training data. A window of N sequential time points was randomly selected with N increasing from 10 to 20,000 time points. The incremental increase was of 10 time points within the range of 10-2000 time points, but it was increased to 200 in the range of 2200–22,000 due to heavy computational costs. We repeated this procedure five times for each window size and averaged the correlation coefficients across movies and iterations. Since this analysis is computationally expensive, we performed it only for the four models that yielded the most reliable decoding, namely loudness, motion, speech, and face ratio. For illustrative purposes, we computed and plotted a regression curve based on a logarithmic least square fitting (Weisstein, 2017). For each feature *f*, we fitted a function in the form of

$$y = a + b \ln x \tag{9}$$

with coefficients computed as follows:

$$b = \frac{n\sum_{i=1}^{n}\left(\overline{R}_i \ln Ws_i\right) - \sum_{i=1}^{n}\overline{R}_i\sum_{i=1}^{n}\ln Ws_i}{n\sum_{i=1}^{n}(\ln Ws_i)^2 - \left(\sum_{i=1}^{n}\ln Ws_i\right)^2} \tag{10}$$

$$a = \frac{\sum_{i=1}^{n}\overline{R}_i - b\sum_{i=1}^{n}\ln Ws_i}{n} \tag{11}$$

n refers to the total number of time windows, $\overline{R}_i$ is the coefficient for the correlation between the predicted and the observed descriptors averaged over movies and iterations per time window, and Ws_i is the window size. Since we had twice as many time windows in the range of 10-2,000 time points, we doubled the weights given to $\overline{R}_i$ in the range of 2,200–22,000 by creating the $\overline{R}$ and *Ws* vectors with a duplication of the values $Ws_{2,200-22,000}$ and $\overline{R}_{\ 2,200-22,000}$.

Second, we examined the advantage of training our algorithm on data obtained from five different movies in comparison to reliance on data from each of the single movies. We repeated the training and testing procedure (without temporal optimization) while limiting the training to data from a single movie. To equalize the size of the training data, we randomly selected an identical number of time points for each of the training movies. According to the training movie for which we had the minimal data, we included 2,576 time points in each single-movie training data set. Five corresponding data sets with the same number of time points were generated by randomly selecting these points from the entire training data including all movies.

We computer the vector $\Delta R_{tm,f,sm}$ so that $\Delta R_{tm,f,sm} = R_{tm,f,all} - R_{tm,f,sm}$, where $R_{tm,f,sm}$ refers to the Pearson correlation between the predicted and observed descriptor for feature *f* at the testing movie *sm* with a model was trained on the training movie *tm*. $R_{tm,f,all}$ is similar except for the fact that the model was trained on all five training movies and the correlations were averaged over the five iterations. A 95% confidence interval of the mean difference was generated around $\Delta R_{tm,f,sm}$ by sampling this vector with replacement (bootstrapping).

In addition, to assess the similarity between the models that were created based on data from the different single training movies, we computed the correlations between the weight vectors.

2.4.4. Temporal optimization

The optimal temporal alignment between the BOLD time course and the target descriptor may vary across voxels and descriptors. This phenomenon can result from inter regional difference in hemodynamic response properties or more interestingly, it may reflect the progression of a functional process. Optimizing the alignment between the signals may therefore improve prediction and provide significant functional information.

However, a full voxel-wise combinatorial fitting is highly computationally demanding (for example, 4^{400} combinations in a case in which 4 possible lags are tested in a mask containing 400 voxels). To reach an estimated solution, we used simulated annealing optimization (Kirkpatrick et al., 1983), which simulates the generation of crystals (i.e., minimum energy structures) when liquids slowly freeze. In this process, the system is more tolerant to perturbations when the temperature is high, but its tolerance is gradually decreasing when it is cooling down. A formal description of the implementation of simulated annealing in our study is provided in the Supplementary Materials.

Using Pearson correlation we compared the temporally optimized predictions with the observed descriptors. The resulted Pearson coefficients were pooled across features and Z-transformed (to approximate

normal distribution). We used one sided paired Wilcoxon signed rank test to examine whether the temporal optimization increased the correlation in an independent data set. This comparison was performed for every individual subject with FDR correction for dependent tests (Benjamini and Yekutieli, 2001).

To identify simple spatio-temporal patterns in the resulting optimized lag maps, we conducted an exploratory analysis including only the model maps that improved the prediction of the target features. Within each cluster in the thresholded maps (with size > 5 voxels) we correlated the optimal lag value (ranging from −1 to 2) with its X, Y, or Z coordinates to detect sagittal, coronal, and axial temporal gradients. In addition, we correlated the lag values with the Euclidean distance of the voxel from the cluster's centroid to identify centrifugal and centripetal patterns. Spearman correlation coefficients were computed in this test, and FDR correction was applied to control for all of the comparisons within each family of hypotheses (i.e., about either X, Y, and Z coordinates or the distance from the centroid).

2.4.5. Specificity and reproducibility of the predictors

To further examine confounding effects across models, we tested whether the temporally optimized target descriptor can be better predicted based on a model generated for other descriptors. Thus, for each movie we compared Pearson's coefficients obtained for the comparison of the target descriptor with either its corresponding average predictor or the predictor of all other descriptors.

The measure of Inter-Subject Correlation (ISC) (Hasson et al., 2004) has been widely used to assess the reliability of BOLD response patterns in the context of naturalistic stimuli such as movies. We expected that our method will yield predictors that not only significantly correlate with the target descriptors, but are also highly similar across subjects. Therefore, we used ISC to examine the similarity between individual predictors for each of the features and movies.

Individual ISC (Hasson et al., 2004) was computed for subject *s*, movie *m*, and feature *f*, as Pearson correlation between the predictor computed for this subject and the average predictor over all other subjects: $ISC_{s,m} = r\,(Pr_{s,m}, \overline{Pr_{all-s,m}})$. For each of the audiovisual features and movies we also computed $ISC_{f,m} = \overline{ISC_{s,f,m}}$. Relying on a previously published protocol for estimating the significance of $ISC_{s,m}$ and $ISC_{f,m}$, we applied a bootstrapping procedure that included phase randomization for each Fourier component of the predictor time series and then the inverse Fourier transformation (Silbert et al., 2014).

We generated 10,000 scrambled time series for each of the individual predictors and computed $ISC_{s,m}$ and $ISC_{f,m}$ for these randomized signals as described above. P values for the empirical $ISC_{s,m}$ and $ISC_{f,m}$ were determined based on a comparison with the resulting null distribution.

2.4.6. Interpreting the weight vectors

The weight vector β cannot be neuroscientifically interpreted *as is* since it reflects not only the signal of interest, but also other information that improves prediction but is not directly related to the studied psychological process (Haufe et al., 2014). In Valente et al. (2014) procedure to remove the confounding effects of other targets has been proposed, which is suited for large scale problems. However, even within a single predictive model, some features can have a large weight to improve prediction by canceling out elicited activity in other areas. To tackle this problem, we followed a procedure that transforms β into a "forward model", which explains how X (the n × s-by-p concatenated matrix of the training data) was generated from neural sources (Haufe et al., 2014).

This method, however, is based on the estimation of covariance between features (in our case, voxels), which is ill-posed since we have more features than samples (n × s < p) and cannot be applied as is to our problem. To generate forward models we therefore reduced the number of voxels by selecting only the voxels that were deemed significant in at least one out of k (k = 6) predictive maps (with a total number of voxels $p_1 = 2,758$). We first show that a prediction using only this subset of voxels is still significant (Fig. S2), and then we apply the transformation suggested in (Haufe et al., 2014). In specific, we computed the activation patterns vector Ac, whose values indicate both the strength and the direction of the effect of the observed features on the data, as follows:

$$Ac = \Sigma_X \beta_{p_1} \Sigma_y^{-1} \tag{12}$$

where $\Sigma_X = E[X^T X]$ and $\Sigma_y = E[y^T y]$. β_{p_1} is an $p_1 x\, k$ matrix with the β weight vectors (predictive maps) computed for each of the features as described above as columns (k is the number of descriptors and in this case, k = 6). *y* is a matrix, which includes the observed descriptors.

The $p_1 x\, k$ matrix Ac was thresholded based on a permutation test. In this test, activation pattern vectors were computed (Equation (11)) for each of the 10,000 weight vectors that were generated using the randomly shifted time series (see above). We used FDR correction for multiple testing under dependency (Benjamini and Yekutieli, 2001) with Q = 0.05.

2.4.7. Music experience and loudness reconstruction veridicality

Finally, we tested whether a generalized fMRI model can be used to reveal functionally meaningful intersubject differences. In specific, we hypothesized that the loudness descriptor will be more accurately reconstructed when applying the loudness model to data obtained from individual with high relative to low musical experience.

The participants rated their acquaintance with the specific musical pieces LM, GH, and MN (5-point scale) and filled a short questionnaire regarding their musical training. The fMRI data were divided into two groups based on the mean reported music-playing experience, which was 5.07, 5.86, and 5.78 years in LM, GH, and MN, respectively. An identical division would have been obtained if a cutoff of 5 years of experience had been selected as in previous neuroimaging studies of musical experience (Chapin et al., 2010; Singer et al., 2016). The high-musical experience group included 10 (4 females, 25.8 ± 4.05 years), 12 (5 females, 25.67 ± 3.77 years), and 11 (4 females, 26 ± 3.77 years) in LM, GH, and MN. The corresponding low-experience groups included 22 (14 females, 25.8 ± 3.65 years), 21 (14 females, 25.67 ± 3.77 years), and 19 (13 females, 25.71 ± 3.75 years).

Temporally-optimized predicted descriptors were computed for each musical piece and were compared with the average observed descriptors as described above. We compared the veridicality of the predicted loudness predictors between the high and low musical experience groups. To note, these groups significantly differed in reported musical experience (LM: Wilcoxon $Z = 4.48$, $p < 1 \times 10^{-5}$, GH: $Z = 4.3$, $p < 5 \times 10^{-6}$, MN: $Z = 4.5$, $p < 1 \times 10^{-5}$), but not in age (LM: Wilcoxon $Z = 0.12$, $p = 0.9$, GH: $Z = 0.31$, $p = 0.75$, MN: $Z = 0.45$, $p = 0.65$), male-female ratio (LM: $\chi^2 = 1.56$, $p = 0.21$, GH: $\chi^2 = 1.95$, $p = 0.16$, MN: $\chi^2 = 2.92$, $p = 0.09$), and acquaintance with the musical pieces (Wilcoxon $Z = 0.65$, $p = 0.52$, GH: $Z = -0.34$, $p = 0.73$, MN: $Z = 0.32$, $p = 0.75$).

Individual Pearson correlation coefficients were computed for the comparison between the predicted and observed loudness descriptors. We tested the hypothesis that musical experience is associated with reconstruction accuracy for each of the musical conditions by comparing the accuracy obtained for the two groups using two alternative methods: (a) one-sided Wilcoxon test; (b) a permutation test with 10,000 random divisions of the data into two groups whose sizes are identical to those of the original groups.

3. Results

Based on GCV-KRR analysis of the training data, we produced 3-dimensional weight maps in a standardized Talairach space for each of the features. These models include linear weights per voxel. These models are available in nifti format as Supplementary Material. Thus, a predicted value for a specific feature in a given volume can be computed by simply summing the results of the voxel-wise multiplications of these weights with the corresponding voxel-wise Z-scored BOLD values. We

generated continuous descriptors that predict each of the six features for each of the nine movie clips on a moment-to-moment basis. We next describe the results of the testing of these models using an independent data set, the effects of the temporal optimization of these models, and the reliability and specificity of features of these models. Finally, we test whether loudness reconstruction veridicality covaries with group difference in musical experience.

3.1. Prediction models for audiovisual features and their accuracy

In terms of correlation between the predicted and observed descriptors, we found three features with significant results in all of the nine new testing movies: loudness ($\overline{R}$=0.6; 0.48 ≤ R ≤ 0.81), speech ($\overline{R}$=0.58, 0.40 ≤ R ≤ 0.81), and motion ($\overline{R}$=0.59; 0.44 < R < 0.81). A partial conjunction analysis (Heller et al., 2007) indicated that a real effect is indeed obtained across all of the nine tested movies for these features (Q_{FDR}<0.05). In the case of face ratio, we found significant correlations for 7 out of 8 movies ($\overline{R}$=0.56, 0.36 ≤ R ≤ 0.69), and a marginally significant correlation (R = 0.4, p < 0.06) for the eighth movie (the overall $\overline{R}$ = 0.54). This proportion of significant effects survived an FDR corrected conjunction analysis.

On the other hand, lightness decoding produced a less reliable model as 6 out of 9 predicted descriptors significantly correlated with the observed descriptors ($\overline{R}$=0.22, −0.12 ≤ R ≤ 0.43). In the case of the brightness model, the predicted and the observed descriptors were significantly correlated only in two movies ($\overline{R}$=0.18; −0.14 ≤ R ≤ 0.41). In both cases, the FDR corrected conjunction analysis indicated that a significant effect appears in at least one movie.

We further applied partial conjunction analysis to test the correlation between the predicted and the observed descriptors for each of the individual runs (for raw data, see Supplementary Data). After FDR correction we found a significant effect (Q_{FDR}<0.05) in 47.6%, 49.2%, 49.2%, 39.7%, 1.6% and 3.2% of the runs in the cases of loudness, speech, motion face ratio, lightness, and brightness, respectively. The average Pearson's correlation between the individual predicted and observed descriptors over runs were 0.46, 0.45, 0.49, 0.39, 0.15, and 0.15, respectively. As evident in the right charts in Fig. 2, the predicted descriptors showed high similarity across subjects. A quantitative analysis, which is described in the Supplementary Materials, indicates a high intersubject reproducibility of the predicted descriptors (Table S1).

3.2. Relation between prediction accuracy and size and type of the training data

A pattern of logarithmic growth in the average prediction accuracy with the increase in the number of data points in the training set was observed in all four tested models (loudness, speech, motion, and face ratio; Fig. 3). We compared these values to the averages of the correlation coefficients reported above, which were obtained using training data sets of ~24,000 time points. Thus, for instance, based on the logarithmic model fitted to the data, we estimate that in order to reach 90%, 95%, and 99% of the average correlations reported above, one would need approximately 4,250, 8,000, 13,000; 2,900, 7,700, 16,700; 2,561, 5,400, 9,800; and 4,800, 7,800, and 11,512 time points in the case of loudness, speech, motion, and face ratio, respectively.

We also examined the benefit of including data for five different movies in the training set rather than relying on data from a single movie. When assessing the effect on the four more successful models, we found that in 40% of the cases, the prediction accuracy was improved with the inclusion of all five movies. The results are described in details the Supplementary Materials (Fig. S4).

3.3. Temporal optimization

We next conducted a voxel-wise optimization of the temporal alignment between the BOLD time series and the observed descriptor in the training set. The optimization procedure increased the correlation between the predicted and observed descriptors in the cross-validation subset (i.e., a subset of the training data that was kept aside for optimization) in 11.27%, 7.69%, 10.3%, 18.3%, 48.9%, and 86.22% for loudness, speech, motion, face ratio, lightness, and brightness, respectively. Its efficiency was assessed in an independent test set (Figs. S3a–f). The pattern of decoding accuracy across movies was similar to the results of the non-optimized procedure, but with higher average correlations (except for brightness; see Fig. 4): loudness – $\overline{R}$ = 0.62 (0.42 ≤ R ≤ 0.85), speech – $\overline{R}$ = 0.60 (0.39 ≤ R ≤ 0.81), motion – $\overline{R}$ = 0.60 (0.43 ≤ R ≤ 0.83), face ratio – $\overline{R}$ = 0.56 (0.34 ≤ R ≤ 0.71), lightness – mean $\overline{R}$ = 0.23 (0.06 ≤ R ≤ 0.38), brightness – $\overline{R}$ = 0.15 (−0.13 ≤ R ≤ 0.42). The correlations between the average predicted and the observed descriptors were significant in all movies for loudness, speech, and motion, and in all movies but one (for which p = 0.054) for face ratio (Fig. 4). These results survived an FDR corrected conjunction analysis. In the case of brightness and lightness, 3 and 4 out of 9 correlations were significant respectively, and none of them survived the FDR corrected conjunction analysis.

A paired one-sided Wilcoxon signed rank test, which was performed for every individual session in the test data set (i.e., n = 63 for all features but face ratio for which n = 58), indicated that temporal optimization significantly increased the accuracy of loudness (Z = 5.94, $Q_{FDR}<5\times10^{-8}$), speech (Z = 3.42, Q_{FDR}<0.005), motion (Z = 4.2, Q_{FDR}<0.0005), and face ratio (Z = 4.42, Q_{FDR}<0.0001) models, but not of lightness (Z = 1.6, NS) and brightness (Z = −1.05, NS) models. Fig. 6 visualizes the resulting spatiotemporal maps of the four more successfully optimized models (the original models are presented in Fig. 5).

At the individual level, a conjunction analysis indicated significant correlation between the temporally optimized predicted and observed descriptors (Q_{FDR}<0.05) in 61.9% of the runs for loudness, speech, and motion, 50% for face ratio, 1.6% for lightness, and 4.8% for brightness, respectively. Apart from the case of brightness, the average individual Pearson's coefficients were all higher than in the non-optimized prediction: 0.48, 0.47, 0.51, 0.42, 0.16, and 0.14, respectively.

3.4. Specificity of the predictive models

We examined whether the correlation of a specific observed descriptor with its corresponding predicted descriptor is higher than its correlation with predicted descriptors generated based on models of other features (Fig. 7; cf. Fig. S6 for specificity of the observed descriptors). Overall, we found high prediction specificity for the predictors of loudness, speech, motion, and face ratio. Considering all of the reciprocal correlations between these features for all movies, only in three cases (out of 102 possible), the observed descriptor was not best predicted by its corresponding model (e.g., speech better predicted loudness in the movie Denali). In one case (out of other 102 possible cases), the predicted descriptor had higher correlation with other observed descriptors than with its own target. In one other case, motion and loudness were equally predicted by the motion predictor. The two other models showed lower prediction specificity. Lightness and brightness were the best predictors of lightness and brightness observed descriptors only in 5 and 3 out of 9 movies, respectively. The predicted descriptors of lightness and brightness had highest correlations with the corresponding observed descriptors only in 2 and 0 movies, respectively.

3.5. Loudness reconstruction veridicality and group differences in musical experience

To examine the efficiency of our method in capturing meaningful inter-group differences, we tested whether music loudness can be more accurately decoded from data of individuals with high musical experience. We used data obtained during the listening to Ligeti's Musica

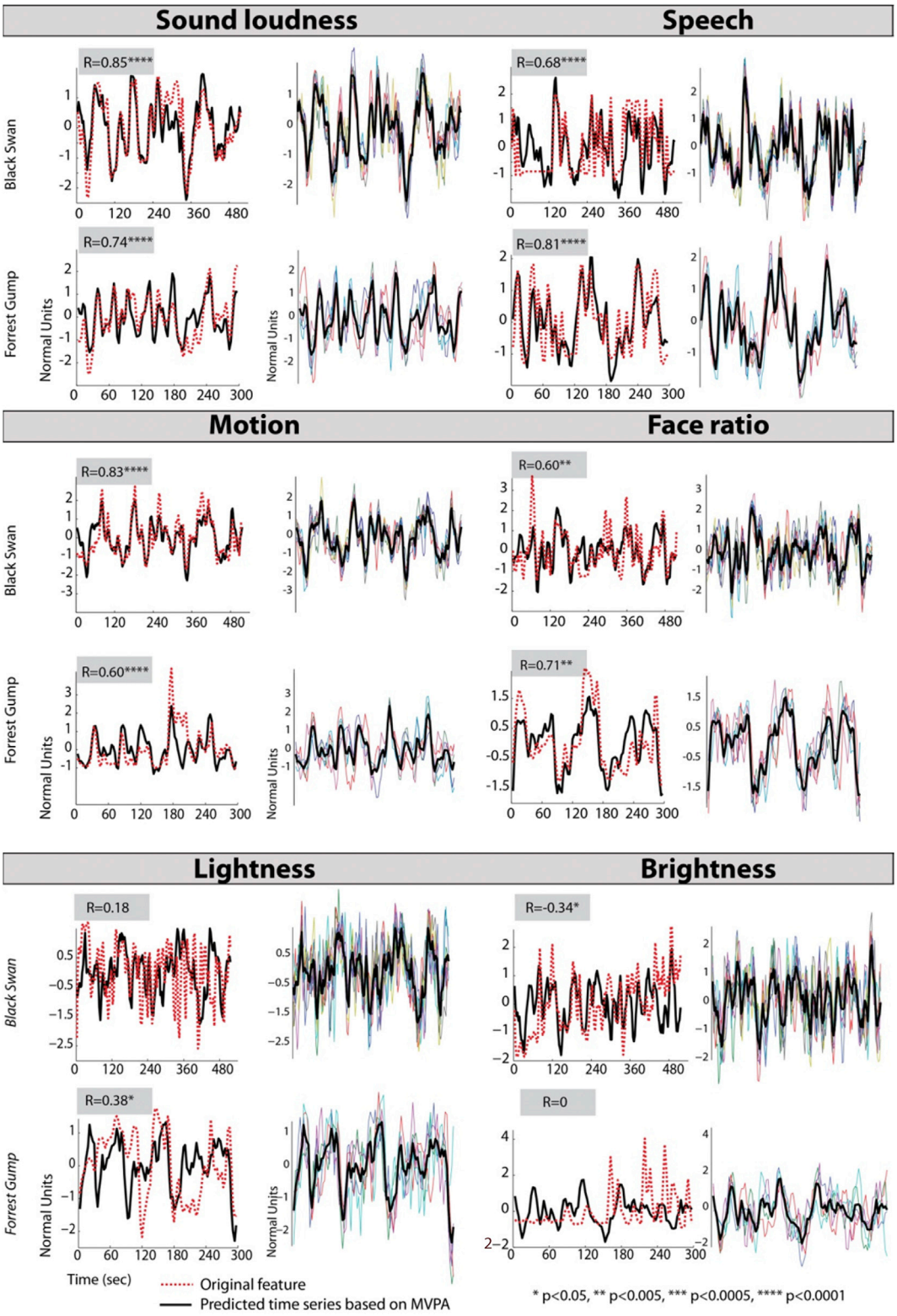

Fig. 2. Reconstructed time series (predicted descriptors) and observed descriptors for two movies: Black Swan (N=8), and Forrest Gump (N=5). The observed descriptors (dashed red lines) and the average predicted descriptors (black line) are presented in the left chart for each of the panels. The right charts indicate the similarity of the reconstructed time series across subjects, as the different individual predicted descriptors are presented in different colors. The predictions were derived from the temporally optimized models. For other movies, see Figs. S3a–f. The presented predictors are derived from temporally optimized models.

Ricercata (LM), Glass's "Hours" (GH), and Mussorgsky's Night on Bald Mountain (MN; see Supplementary Materials). Although the loudness model was generated using multimodal movie data, the average predicted descriptor significantly correlated with the observed loudness descriptor of the unimodal musical pieces LM ($R = 0.71$, $p < 0.001$) and MN ($R = 0.55$, $p < 0.005$). In the case of GH, the observed loudness descriptor significantly correlated with the average predicted descriptor of the high-musical experience group ($R = 0.29$, $p < 0.05$), but only

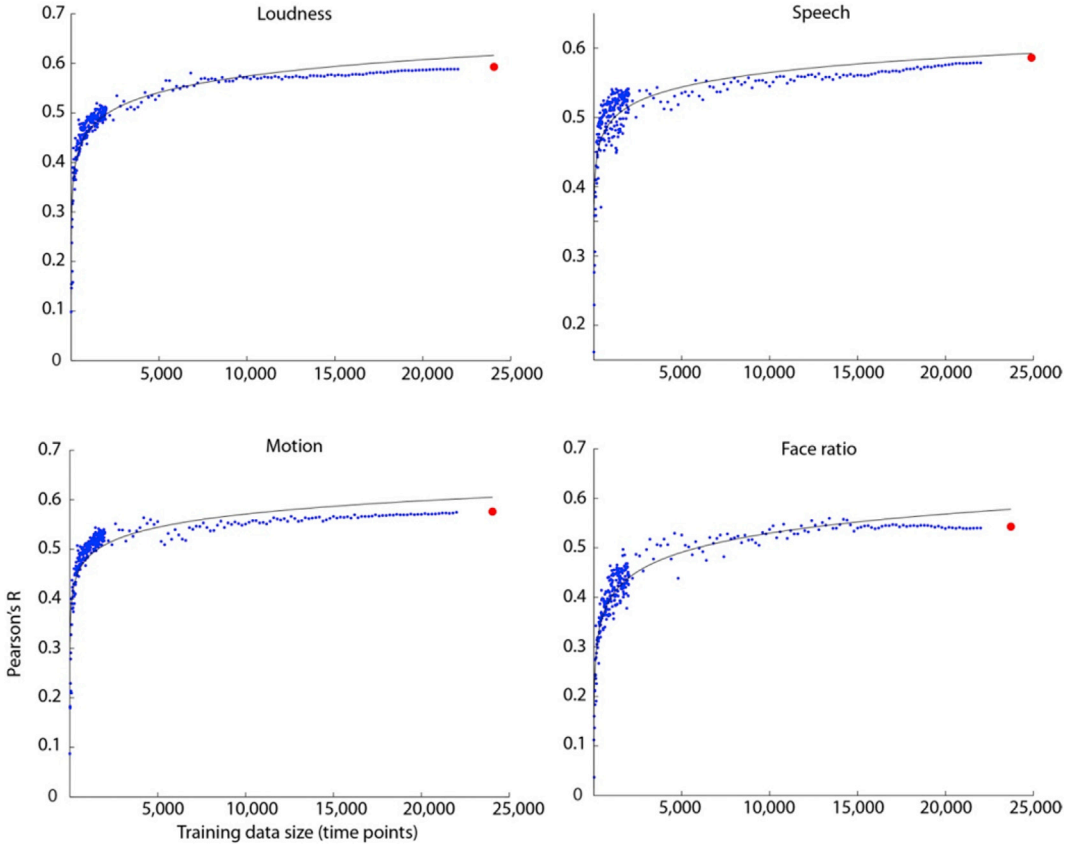

Fig. 3. Relations between prediction accuracy and training data size. The y-axis denotes the correlation between the predicted and observed descriptors averaged over the testing movies. The x-axis denotes the number of points in the data that were used for GCV-KRR training. A logarithmic least square fitting curve is presented in black. The red dots represent the average prediction obtained by using 75% of all training data as reported in the text.

marginally correlated with the overall average predictor ($R = 0.15$, $p = 0.08$). Furthermore, significantly higher correlation between the predicted and observed loudness descriptor was found for high-relative to low-musical experience group in all three conditions as indicated by both permutation test (random assignment of participants to the groups) and non-parametric Wilcoxon test (LM: $Z = 2.05$, p(Wilcoxon) < 0.02, p(permutations) < 0.03; GH: $Z = 1.96$, p(Wilcoxon) < 0.03, p(permutations) < 0.05, MN: $Z = 2.11$, p(Wilcoxon) < 0.02, p(permutations) < 0.02; Fig. 8). FDR corrected conjunction analysis confirmed a significant effect in any of the three conditions (for both testing methods).

3.6. Components of the standard brain models for audiovisual features

Our version of GCV-KRR analysis allows for the interpretation of the linear contribution of every voxel in the model to the audiovisual feature of interest (predictive maps). However, for this aim, the multiplicity of target features has to be accounted for (Valente et al., 2014) and the multi-voxel models should be transformed into mass univariate models (Haufe et al., 2014). In our application, we first generated the predictive maps, then re-generated the models based on a subset of voxels included in the thresholded union mask, and finally transformed the predictive models into "forward models" following the procedure described in Haufe et al. (2014).

The thresholded predictive maps of the six models are visualized in Fig. 5 (and are provided in the Supplementary Material). Details on the weight clusters are presented in Table 2. While large positive and negative clusters (mainly cortical) survived the test when examining the loudness, speech, motion and face ratio models, in the cases of lightness and brightness we found only a few occipital clusters, and two small pulvinar and cuneus clusters, respectively. The predicted descriptors that were derived from models computed over the subset of voxels in the union mask (Fig. S7) were slightly less correlated with the observed descriptors (Wilcoxon $Z = 1.97$, $p = 0.0499$) relative to the original predictors (computed over all gray matter voxels). The difference between the median correlations across movies was small: 0.45 versus 0.49, respectively (see Fig. S2).

Based on these constrained predictive models, we next produced generative forward models, whose entries may be interpreted as relevant activations and deactivations (Fig. 5B, Table 3) for loudness, speech, and motion. In the case of the face size model, the transformed weights did not survive FDR-corrected permutation test, although a bilateral cluster of negative parahippocampal values with $p = 1 \times 10^{-5}$ at its peak was observed.

3.7. Spatiotemporal maps of the models

Since slow propagation of neural processing may be reflected in lagging BOLD signals (see Formisano and Goebel, 2003), we were interested in patterns within our spatiotemporal models. We looked for axial, coronal, and sagittal gradients, as well as centrifugal (higher lags for voxels with greater distance from the centroid) and centripetal patterns. The exploration was limited to 54 clusters pooled over the lag optimization maps of loudness, speech, motion, and face ratio. The results are presented in Table 4. Two selected spatiotemporal patterns are visualized in Fig. 9.

4. Discussion

Our findings suggest that the combination of GCV-KRR and temporal optimization via simulated annealing facilitates the production of fairly reliable predictive models for continuous audiovisual features as

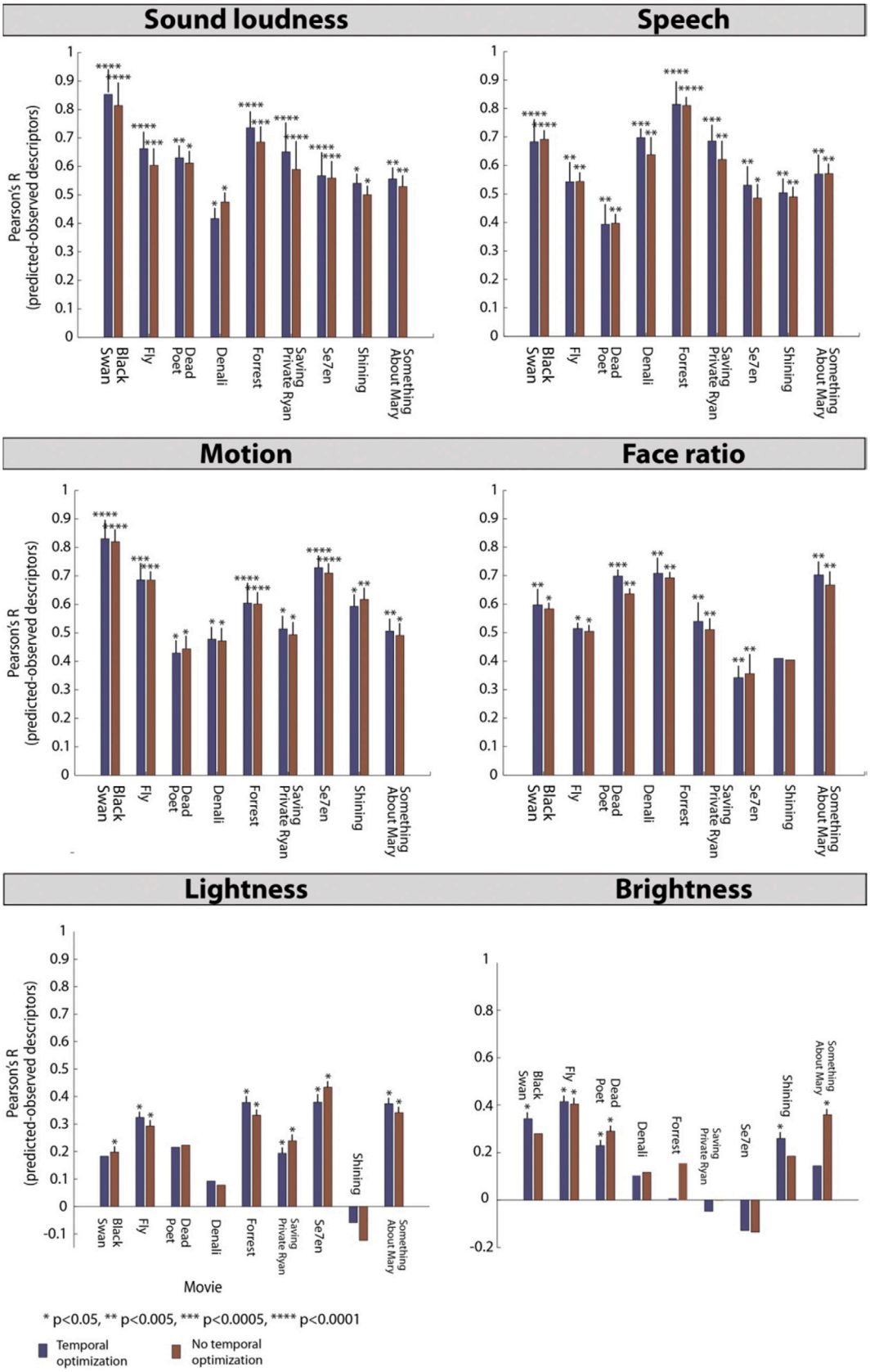

Fig. 4. Accuracy and reliability of the reconstruction of audiovisual features. Pearson's R coefficients for the correlation between the predicted and the target descriptors are presented for each of the movies and audiovisual features. Note that since the movie Denali contained no images of human face, face ratio prediction was not tested in this case.

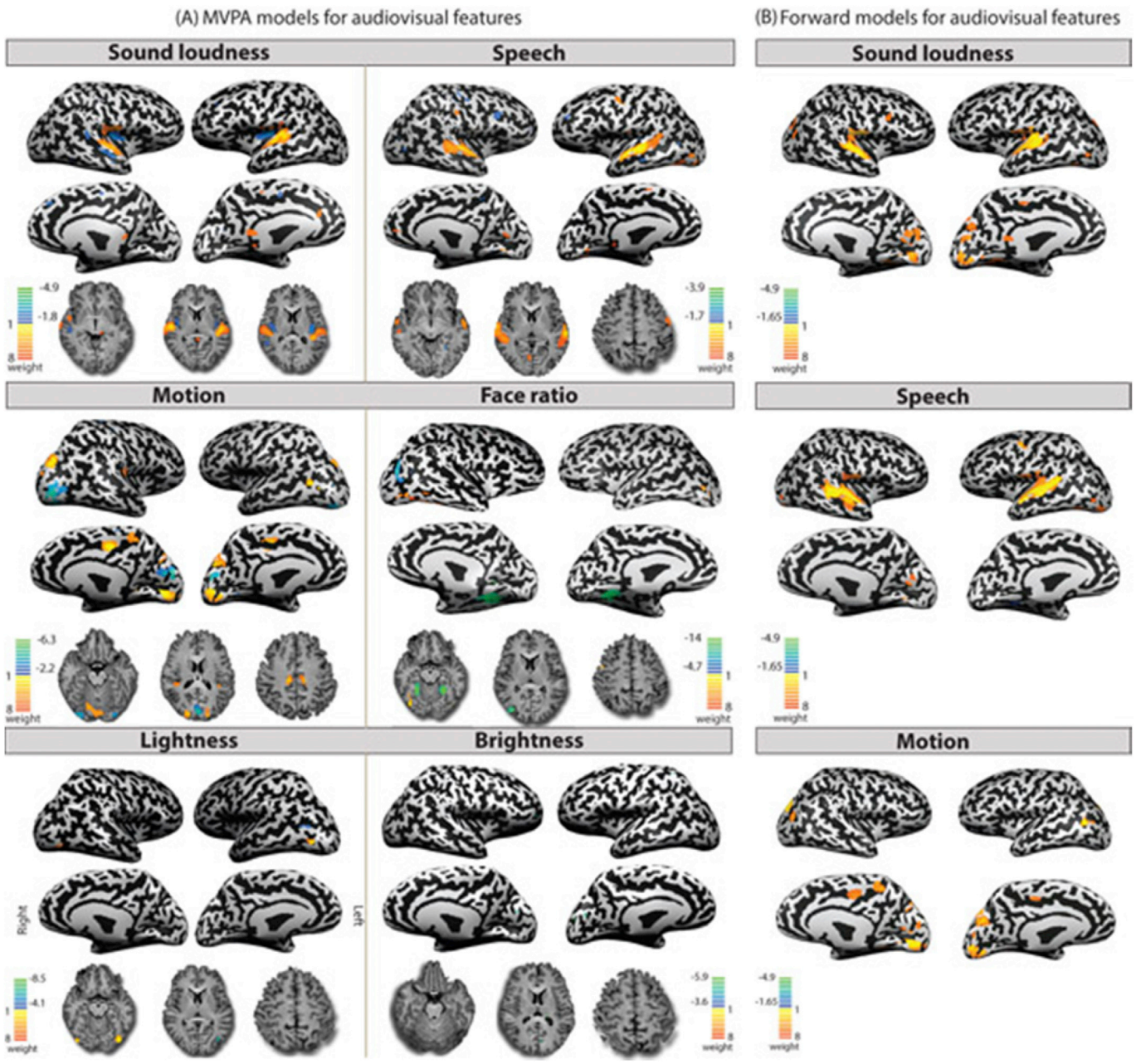

Fig. 5. Brain models extracted via GCV-KRR for the audiovisual features of interest. (A) MVPA maps thresholded at $Q_{FDR}<0.05$ following a permutation test. (B) Forward models computed based on the MVPA maps following the transformation method described in (Haufe et al., 2014).

observed in uncontrolled cinematic content. Four of these models – loudness, motion, speech, and face ratio - show good generalization over contents, subjects, and scanning conditions both at the individual and the group levels. The decoding robustness is further supported by the high similarity of the predictor time series across subjects, as suggested by our ISC analysis (Table S1).

Apart from the reliability of the reconstruction, our method also yields interpretable brain models for the examined audiovisual features. In general, the spatial constellations of these models are highly congruent with the relevant neuroscientific literature. The loudness model contains bilateral high positive weights for the primary auditory cortex and the left auditory thalamic medial geniculate nucleus (as well as a cluster in its right homologue, but with lower weights), which are key components in the auditory pathway (Hudspeth et al., 2013). The speech model (after accounting for multiple targets) includes large clusters of high positive weights across the posterior superior temporal regions overlapping with the Wernicke area and its right homologous, which are implicated in language processing (Bigler et al., 2007). The motion model consists of focal positive weights in the classical motion processing region MT\V5 (Howard et al., 1996), as well as in the dorsal cuneus/V6 and cingulate sulcus visual area (CSv). Both of these regions have been implicated in the perception of wide-field, complex, and coherent movement, and especially in self-motion (see Cardin and Smith, 2011; Pitzalis et al., 2015; Fischer et al., 2012; Wall and Smith, 2008, respectively). Finally, while the face ratio model did not survive the transformation into a forward model, it includes bilateral positive occipital face area weights, and bilateral negative parahippocampal place area weighted, possibly reflecting a parametric trade-off between face and scene perception (see Vanduffel and Zhu, 2015). PH1 and PH2, which are part of the face ratio model, respond preferentially to scenes and landmarks and show peripheral bias (Baldassano et al., 2016). Thus, as the face ratio is associated with the dominance of the face relative to the background in the

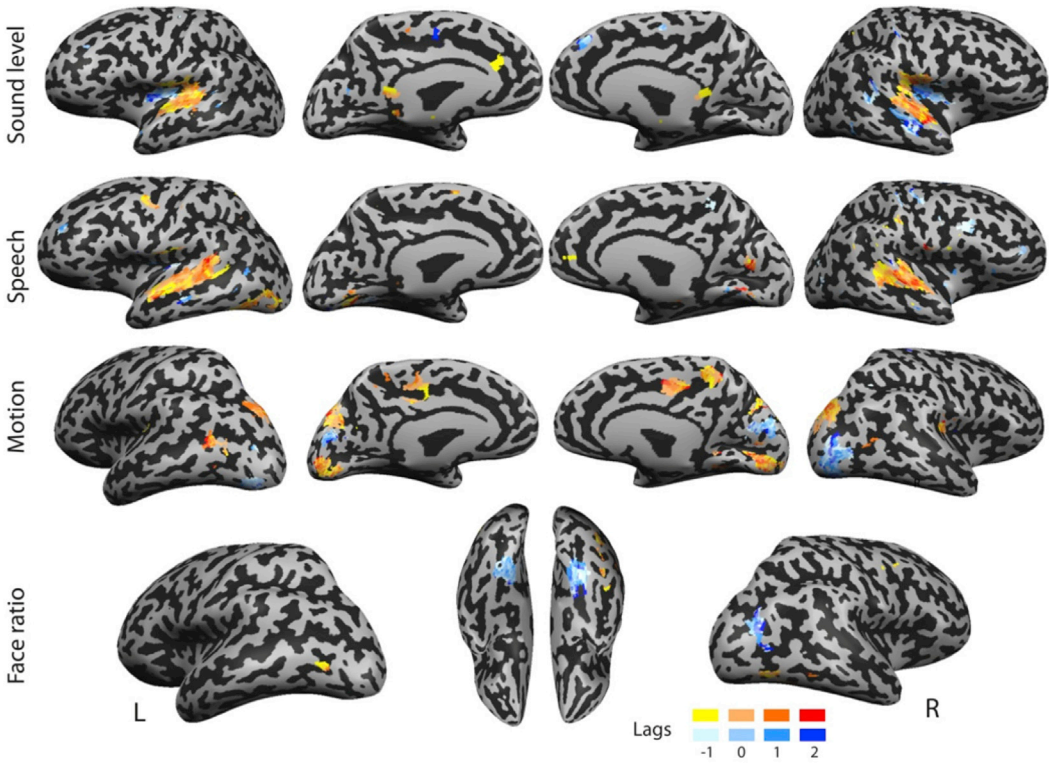

Fig. 6. Successfully optimized spatiotemporal models for audiovisual features. Optimal lags (3 s for lag) are visualized in different color schemes for positive and negative weights. The lag maps are presented after masking them by the thresholded maps presented in Fig. 5.

visual field, the negative weights of the PHC probably reflect the enhanced processing of scene-related elements when the face size is small.

An apparently surprising finding is the cluster of positive weights in the primary visual cortex and adjacent sections of the lingual gyrus (Table 3, Fig. 5B) in the thresholded generative (forward) model of sound loudness. In fact, this finding is in line with substantial evidence that has accumulated over the past decade, pointing to the responsiveness of the primary visual cortex to auditory input (for a review, see Murray et al., 2016). These works support the notion that regions that were considered as unimodal visual areas are in fact implicated in multisensory processing. Considering the type of stimuli used in our study, our findings point to the possibility that loudness consistently affects visual processing in the primary visual cortex in natural perception.

Notable differences can be observed between the predictive and the forward models (Fig. 5). The comparison between the maps points to intriguing possibilities that are not evinced by standard univariate analyses alone. Thus, for instance, while the loudness predictive model includes large bilateral clusters of negative posterior insula/parietal operculum weights, these voxels are positively weighted in the forward model. It is possible that while loudness-related activation is measured both in A1 and the posterior operculum, the contrast between the signals of these adjacent regions filter the signal and yields a more accurate loudness decoding.

On the other hand, the differences between the models may also reflect functional rather than technical aspects. This may be the case with our findings on the dorsal V3B and ventral LO1 and LO2 weights in the motion intensity model. Both of these regions were previously functionally defined as “kinetic occipital” (KO) areas, which responds more strongly to motion boundaries (an important cue for object-background separation) than to transparent motion (Dupont et al., 1997; Larsson et al., 2010; Larsson and Heeger, 2006). However, in the motion model we found positive weights for the dorsal region, but negative ventrolateral occipital weights. In the generative motion model, dorsal V3B is positively weighted, while no activation is observed in LO1 and LO2. Previous studies already reported on differences between the ways in which these KO regions process motion. For example, LO1 and LO2 showed increased responses to images of objects created by dot movement, while an area that overlaps with our dorsal V3B cluster was sensitive to moving edges but not to shapes (Vinberg and Grill-Spector, 2008). These observations are congruent with a hierarchical model of motion boundary processing, which posits that local moving edges are identified by a set of linear filters in the first step and the resulting information is then integrated to generate higher-order perception in the second step (Larsson et al., 2010). In line with this model, it is possible that a high rate of motion boundary cues will trigger increased processing in regions implicated in the first-order filtering, but will also be too rapid to elicit a coherent integration, which implicates high-level perception regions. This interpretation is congruent with evidence suggesting that the lateral occipital cortex (LOC) is the highest purely visual area in the ventral stream hierarchy (Lehky and Tanaka, 2016).

Thus, the ratio of LOC and V3B activity levels may be functionally meaningful as an index of motion perception acuity. In other words, the contrast between the activity levels of these (non-adjacent) regions may not only technically allow for better analytical prediction, but also reflect a functional status, which is interpreted by higher-level systems in the brain. The advantage of multi-voxel modeling is evident in this case as the LOC-V3B interplay is not manifested as LOC deactivation following intensive motion in the univariate generative model. Our results point to the potential of a focused analysis on the interplay between the activity levels of these regions in the context of motion perception.

The temporal optimization, which was applied as part of our procedure, significantly improved the prediction of the four best predictive models generated by GCV-KRR. The interpretation of the resulted spatiotemporal models (with the optimized lags) in psychophysiological terms should be taken with caution, since the spatiotemporal gradients may be caused by factors such as regional variability in neurovascular coupling parameters (i.e., hemodynamic rather than neural latencies). However, these patterns may nevertheless reflect slow gradual recruitment of neural populations (Formisano and Goebel, 2003). At least in one

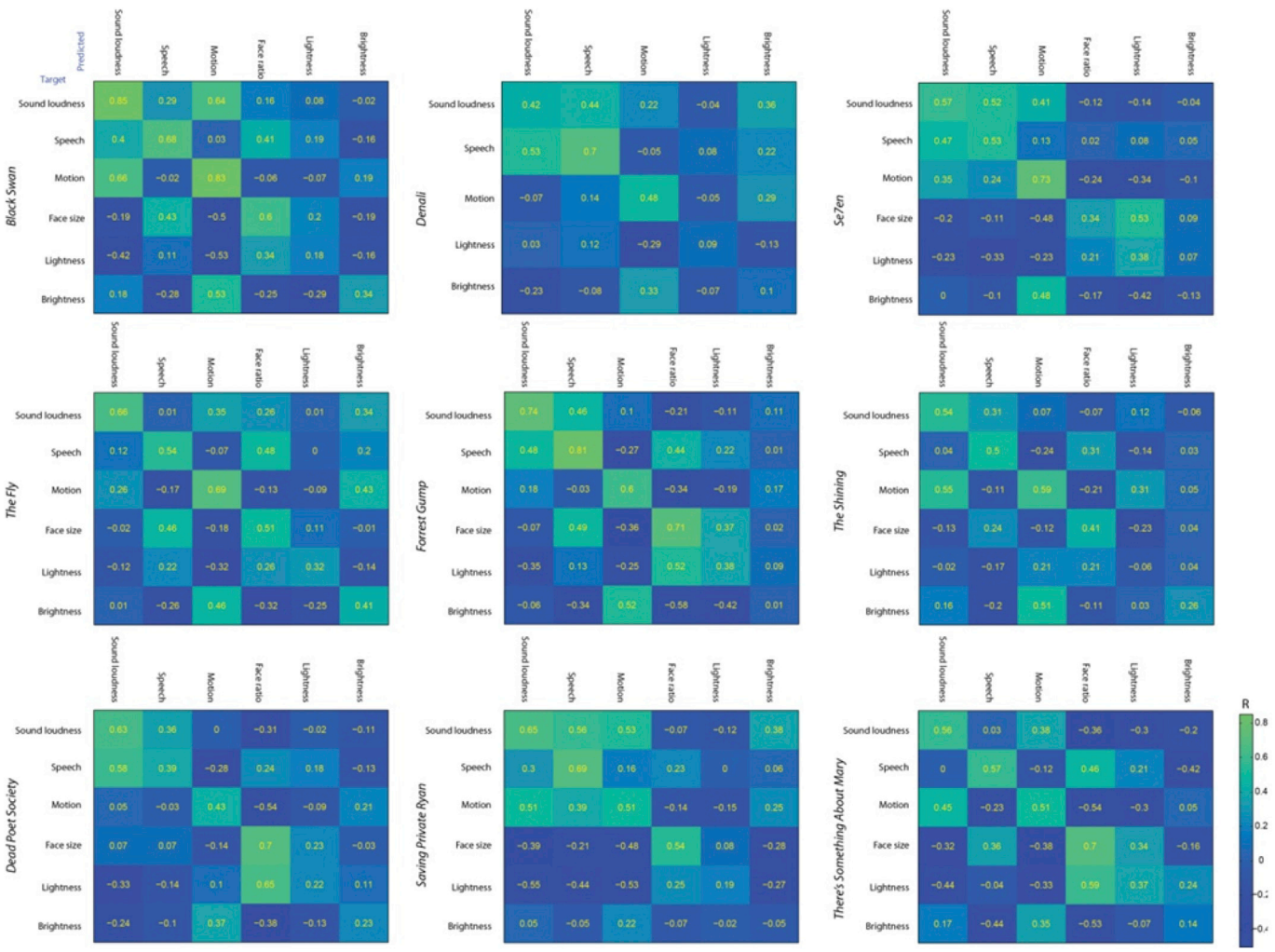

Fig. 7. Correlations across predictors and descriptors. The correlations between the average predictor for each of the features and the target descriptors are presented for each of the movies. The value in each cell indicates Pearson's correlation coefficient for the prediction of the column target descriptor by the row predictor. Note that in the case of the movie Denali, no face ratio descriptor was extracted.

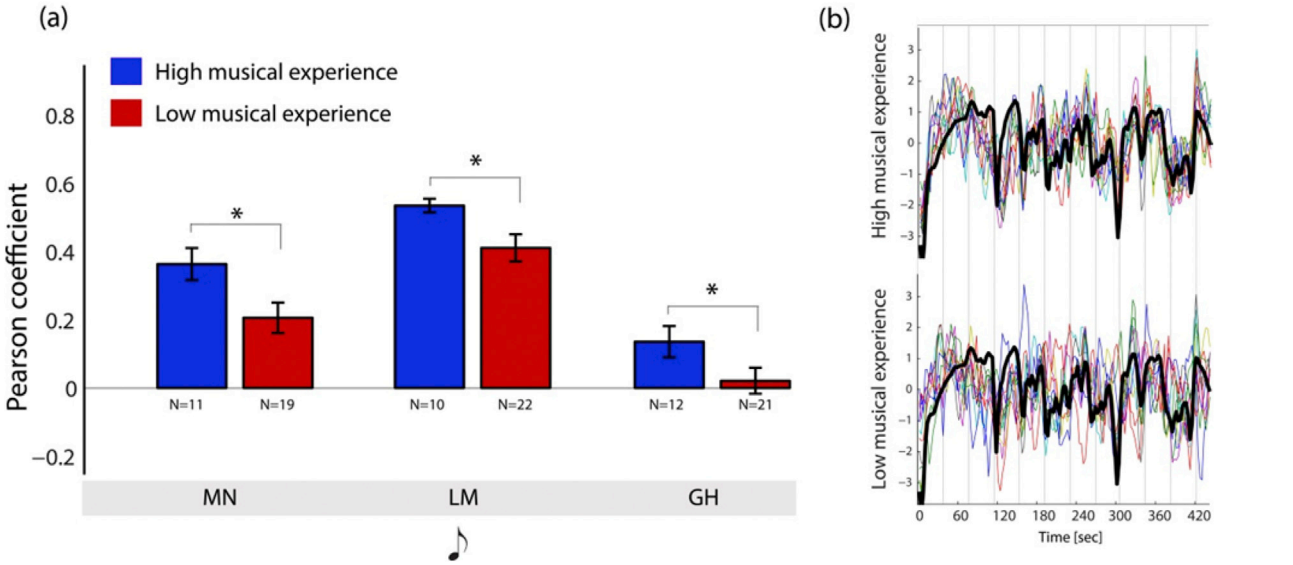

Fig. 8. Reconstruction of loudness time-courses in three musical pieces based on the general loudness model and the fMRI data. (a) Reconstruction accuracy is compared for individuals with high and low musical experience in terms of the average Pearson correlation coefficient (and the standard error) between the individual predicted descriptor and the observed loudness descriptor. (b) Individual predicted descriptors (colored curves) and the observed loudness descriptor (solid black curve) after Z-scoring for professional musicians and non-musicians in LM. Descriptors of 11 out of 21 subjects with low experience were randomly selected so that the visualization will include an even number of curves for both groups.

Table 2
Details on the model maps for the examined audiovisual features.

Region label	Mean weight	Max weight	Cluster size	X	Y	Z
Sound loudness						
Positive weights						
R transverse temporal gyrus/ Heschl's convolutions, superior temporal gyrus	4	8	272	51	−13	7
L transverse temporal gyrus/ Heschl's convolutions, superior temporal gyrus	4.08	7.14	235	−48	−19	7
Posterior cingulate cortex	2.65	3.07	11	−3	−40	7
L medial geniculate nucleus (thalamus)	2.31	2.71	8	−12	−28	−2
Negative weights						
R posterior insula, parietal operculum	−3.02	−4.57	93	36	−19	19
L posterior insula	−3.11	−4.91	55	−36	−19	16
R superior temporal sulcus	−2.66	−3.35	37	48	−16	−5
R superior temporal gyrus	−2.46	−2.81	18	51	−43	13
R superior frontal gyrus	−2.43	−3.07	16	6	35	46
Medial frontal gyrus	−2.25	−2.11	8	3	−13	49
Speech						
Positive weights						
R superior temporal gyrus, middle temporal gyrus	3.96	7.47	296	63	−10	4
L superior temporal gyrus	3.98	8	242	−60	−22	4
L middle temporal gyrus	2.95	3.55	23	−51	−43	10
L fusiform gyrus, inferior occipital gyrus	2.79	4.52	99	−39	−79	−11
L precentral gyrus	2.94	4.16	36	−48	−7	52
R precentral gyrus	2.27	2.35	6	51	−1	43
L medial occipitotemporal gyrus	2.85	3.59	20	−15	−67	−11
R V2v	2.85	3.47	22	3	−70	4
L V1v, V2v	2.79	3.3	7	−6	−67	−2
L intraparietal sulcus	2.57	3	13	−24	−55	43
R declive	2.27	2.66	12	9	−67	−20
R postcentral gyrus	2.26	2.56	7	54	−28	40
R red nucleus/pons	2	2.19	8	6	−25	−20
Negative weights						
L parahippocampal gyrus (PHC1)	−3.25	−3.9	12	−27	−49	−5
R middle temporal gyrus	−2.93	−3.65	12	66	−16	−8
L middle temporal gyrus	−2.28	−2. 5	8	−42	−61	1
L transverse temporal gyrus	−3	−3.25	7	−33	−28	13
R occipitotemporal gyrus	−2.79	−2.94	7	21	−55	−8
R superior temporal gyrus	−2.66	−2.86	8	63	−31	16
R postcentral gyrus	−2.37	−2.84	23	39	−28	58
R angular gyrus	−2.44	−2.82	8	30	−79	4
R precentral gyrus	−2.21	−2.56	13	39	8	31
L posterior superior fissure (cerebellum)	−2.44	−2.53	8	−18	−76	−23
L superior frontal gyrus	−1.98	−2.1	8	−27	44	31
Motion						
Positive weights						
Lingual gyrus (R+L V1v & V2v, R V3v, R VO2)	5.42	8	223	9	−76	−8
R posterior cingulate sulcus	4.75	7.42	41	12	−22	42
L posterior cingulate sulcus	3.89	5.82	27	−12	−22	40
R cuneus (dorsal), V3B, V6	4.77	7.33	142	18	−85	22
L cuneus (dorsal), V3B, V6	4.04	5.72	104	−18	−85	19
L medial temporal (MT) area/V5	4.77	6.13	30	−40	−66	4
Precuneus	4.96	3.92	37	9	−49	52
R posterior insula	3.62	4.35	24	36	−34	19
L posterior insula	3.93	4.35	8	−39	−34	19
Negative weights						
Cuneus (V1d, V2d)	−4.26	−6.28	92	−3	−82	16
R lateral occipital complex, inferior occipital gyrus	−4.63	−6.16	126	30	−85	−11
L inferior occipital gyrus	−3.89	−4.66	32	−30	−85	−17
Brightness						
Negative weights						
L pulvinar	−7.21	−7.97	6	−15	−28	19
Cuneus	−5.3	−6.45	7	−6	−76	13
Lightness						
Positive weights						
L inferior occipital gyrus, fusiform gyrus	6.05	8	45	−39	−79	−11
R inferior occipital gyrus	5.70	6.86	11	39	−79	−11
						Negative weights
L middle occipital gyrus	−4.95	−5.90	8	−39	−70	7
Face ratio						
Positive weights						
L inferior temporal sulcus	7.23	8	7	−39	−76	−8
R fusiform face area, inferior temporal sulcus	5.73	7.2	47	39	−76	−11
R precentral gyrus	5.68	6.24	6	48	−4	46
Negative weights						
R parahippocampal gyrus (PHC1, PHC2), fusiform gyrus (VO2)	−8.49	−13.97	117	27	−49	−5
L parahippocampal gyrus (PHC1, PHC2)	−8.38	−13.36	85	−27	−49	−5
R transverse occipital sulcus	−6.17	−7.48	36	33	−76	10

case the observed patterns in our data is congruent with previous literature, indicating the reliability of the method: the centripetal or bifocal spatiotemporal pattern identified in the right MTG/STG in the speech map is in line with a prominent theoretical model of the cortical organization of speech processing (Hickok and Poeppel, 2007). This model suggests that while the more dorsal aspect of the STG, including the parietal operculum, is part of a dorsal stream implicated in auditory motor integration, a more anterior section of the STG facilitates spectrotemporal analysis of the auditory input. The right MTG/STG cluster in our speech model is split in a way, which is consistent with this bifocal anatomical segregation.

While audiovisual decoding was highly reliable in four different cases, lightness and brightness intensities were not reconstructed

Table 3
Details on the forward model maps for the examined audiovisual features after correction for multiple targets.

Region label	Mean T	Max T	Cluster size	X	Y	Z
Sound level						
R transverse temporal gyrus/Heschl's convolutions, superior temporal gyrus	4.38	7.61	158	51	−13	7
L transverse temporal gyrus/Heschl's convolutions, superior temporal gyrus	4.64	7.62	158	−39	−31	10
V1, lingual gyrus	4.64	4.44	98	18	−67	−8
R posterior insula	−3.74	−4.88	51	36	−19	19
L posterior insula	−3.39	−3.92	9	−36	−19	13
L posterior insula	−3.15	−3.32	6	48	−4	−11
R inferior semi lunar lobule (cerebellum)	−3.47	−3.66	6	18	−73	−38
Speech						
R superior temporal gyrus, middle temporal gyrus	4.02	6.37	167	63	−7	1
L superior temporal gyrus, middle temporal gyrus	4.33	6.55	158	−60	−25	4
L medial occipitotemporal gyrus	3.39	3.98	11	−18	−97	−11
L fusiform gyrus	3.33	3.80	10	−33	−46	−20
L superior frontal gyrus	−3.64	−4.34	6	12	65	7
L transverse temporal gyrus	−3.26	−3.71	7	−36	−28	10
Motion						
R lingual gyrus	4.01	5.71	88	18	−67	−8
L lingual gyrus	3.17	3.47	9	−12	−73	−5
R cuneus (dorsal), V3B, V6	3.56	4.62	79	18	−79	22
L cuneus (dorsal), V3B, V6	3.81	5.61	70	−18	−88	19
R posterior cingulate sulcus	4.03	5.35	15	12	−22	40
L posterior cingulate sulcus	3.97	5.29	11	−12	−22	40
R medial temporal (MT) area/V5	3.61	4.46	9	39	−61	4
L medial temporal (MT) area/V5	3.50	4.38	19	−39	−67	4
R parahippocampal gyrus	3.44	4.06	11	18	−43	−8
Vermis (cerebellum)	3.40	3.22	6	0	−70	−32
Cuneus (V1d, V2d)	−3.90	−5.60	65	−3	−79	16
R lateral occipital complex, inferior occipital gyrus	−3.35	−3.85	13	27	−85	−14
L middle occipital gyrus	−3.25	−3.53	13	−30	−79	7

reliably. This intriguing null finding may be explained by the fundamental non-linearity of luminance perception in the visual system (Fiorentini, 2004; Gilchrist et al., 1999) as manifested by functions such as edge enhancement and dynamic range adjustments. It is possible that future implementation of non-linear kernels will yield better results. Interestingly, while linear models failed to predict lightness intensity, the linear sound intensity model did yield fairly reliable accurate predictions. This difference points to an important dissimilarity between these modalities: while the visual input is nonlinearly normalized, the auditory processing linearly maintains key properties that may support functions such as distance estimation.

A key finding in our study is the increased veridicality of loudness reconstruction among individuals with high musical experience, which points to the sensitivity of our method to functionally meaningful intergroup differences. It should be noted that these findings are preliminary in terms of their significance to the research of music perception. For example, it is yet to be tested whether similar intersubject differences can be found when reconstructing other auditory features such as pitch and tempo. Another open question is whether this difference is indeed related to acquired skills or alternatively to biological predispositions. However, regardless of the specific interpretation given to the findings, they suggest that the accuracy level of the feature reconstruction from the individual brain may comprise a valuable neuropsychological measure and a supplementation to existing neural decoding tools.

4.1. Caveats, limitations and potentials

The accuracy and reliability of the decoding of some of the audiovisual features examined here is a noteworthy achievement considering the heterogeneity of the data and the multiple error sources. Such outcomes depend on numerous parameters including the accurate annotation of the audiovisual features, sufficiently linear mapping between the stimulus magnitude, the neural activation and the BOLD signal, a proper allocation of the subject's attention to the movies, inter-subject similarity in the brain's reaction to the content, valid fMRI data acquisition, reliable preprocessing procedures, accurate registration of the functional and anatomical data, and a solid standardization of the brain images. Given this set of methodological challenges, the fairly reliable decoding not only provides evidence for the robustness of the MVPA method applied here, but also validates the set of brain imaging tools and standardization procedures that were applied to acquire and process the data.

The robustness of our decoding approach is indeed limited by any of the above-mentioned parameters, but they may also be further optimized to increase the accuracy and replicability of the results. Thus, for instance, the co-registration of the functional data may be improved using cortex-based alignment (Frost and Goebel, 2012) and inter-subject cortical alignment (Frost and Goebel, 2013; Sabuncu et al., 2010). Hyper-alignment across subjects may also be performed based on one common movie (or its part) shown to all subjects (Chen et al., 2015; Guntupalli et al., 2016; Haxby et al., 2011). In addition, temporal and spatial filtering parameters may be fitted as part of the optimization procedure considering their added value to the prediction.

While our decision to use uncontrolled and highly variable cinematic stimuli provided an appropriately challenging context for validating the decoding generalizability, this choice also entails the methodological risk

Table 4
Spatiotemporal patterns in the predictive models.

Location	Model	Valence	Pattern	R	P
R insula	loudness	negative weights	centrifugal	0.44	$Q_{FDR}<0.005$
R insula	loudness	negative weights	axial (posterior-anterior)	0.42	$Q_{FDR}<0.005$
R insula	loudness	negative weights	coronal (superior-inferior)	0.57	$Q_{FDR}<0.0001$
R insula	loudness	negative weights	sagittal (medial-lateral)	0.39	$Q_{FDR}<0.01$
L posterior insula	loudness	negative weights	coronal (inferior- superior)	0.43	$Q_{FDR}<0.05$
R primary auditory cortex	loudness	positive weights	sagittal (medial-lateral)	0.27	$Q_{FDR}<0.0005$
R Superior and middle temporal gyri	speech	positive weights	centripetal	0.19	$Q_{FDR}<0.05$
R Superior and middle temporal gyri	speech	positive weights	axial (posterior-anterior)	0.18	$Q_{FDR}<0.05$
R Superior and middle temporal gyri	speech	positive weights	coronal (inferior- superior)	0.32	$Q_{FDR}<0.0001$
L medial occipitotemporal gyrus	speech	positive weights	axial (posterior-anterior)	0.66	$Q_{FDR}<0.05$
R middle temporal gyri	speech	negative weights	axial (posterior-anterior)	0.83	$Q_{FDR}<0.05$
L precentral gyrus	speech	positive weights	coronal (superior- inferior)	0.68	$Q_{FDR}<0.0005$
R lateral occipital complex	motion	positive weights	sagittal (medial-lateral)	0.51	$Q_{FDR}<0.05$
L lateral occipital complex	motion	positive weights	sagittal (medial-lateral)	0.32	$Q_{FDR}<0.01$
R parahippocampal gyrus	face ratio	negative weights	sagittal (lateral -medial)	0.29	$Q_{FDR}<0.05$

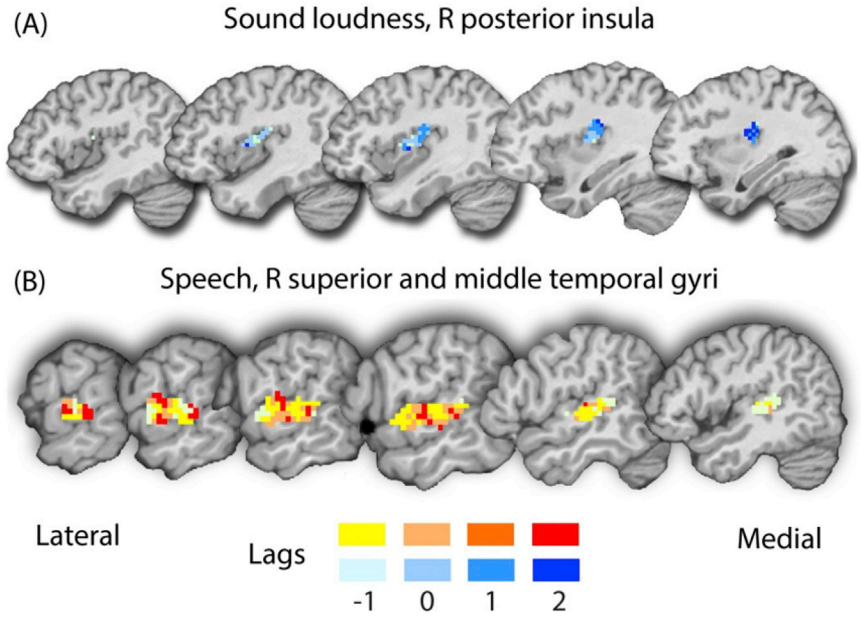

Fig. 9. Centrifugal (a) and centripetal (b) spatiotemporal patterns in optimized models for audiovisual features. The colors represent the optimized temporal shift for the voxel's time series in relation to the descriptor (after it was convolved with HRF) so that −1 indicates that the BOLD series was shifted backward in one time point relative to the descriptor, and 1 indicates a single time point shift in the other direction. Yellow to red colors indicate positive weights, while pale blue to blue colors indicate negative weights.

of modeling spurious and accidental correlations between features. Although we selected varied audiovisual materials with diverse structures of cross-feature correlations (see Fig. S6) to reduce such confounds, our stimuli may still include such correlations that result from common cinematic stylistic strategies (e.g., simultaneous intensification of motion and loudness). In this case, our models might be less generalizable to content other than directed movies. Therefore, future research should further test the validity of our models using controlled data sets to assess their bias to accidental correlations. It should also make use of non-directed movies to better account cross-feature correlations that are not resulting from cinematic style (see Adolphs et al., 2016 for a discussion on the benefits of naturalistic designs).

Furthermore, while in principle GCV-KRR can be used to generate generalizable brain models for any annotated continuous mental and perceptual feature, the current proof-of-concept study focuses on the decoding of coarse and salient features. It does not warrant the successful decoding of more refined features (e.g., subtypes of motion or sound) whose investigation may have higher value in specific scientific contexts. The validity and productivity of our decoding method in these cases has yet to be proven.

Finally, validated decoding models may be employed in various neuroscientific and clinical contexts. For example, naturalistic stimuli and GCV-KRR may be combined to produce multi-feature fMRI localizer, which will simultaneously demarcate brain correlates of multiple processes. Moreover, given the high reproducibility of the reconstructed time series across subjects and the evidence on the sensitivity of this method to inter-individual factors, GCV-KRR could be used to generate population norms for patterns of reaction to specific stimuli. The extent to which a specific feature can be reliably decoded from one's neuroimaging data is a potential neuroscientific measure or even a biomarker that may be used for clinical ends. Current models may be used, for instance, to characterize and study perception during sedation and minimal conscious states, hearing loss, verbal deficiencies, and face perception.

5. Conclusions

MVPA based on the combination of GCV-KRR and temporal optimization via simulated annealing produces a fairly accurate, reliable, and generalizable reconstruction of four coarse audiovisual features when testing cinematic content: loudness, speech, motion intensity, and face ratio. This method also produces interpretable spatiotemporal brain models, which are congruent with the literature but also suggest additional insights. Based on these findings and possible future optimizations, we believe that this approach offers a reliable and valuable tool for scrutinizing dynamic neural processes in both scientific and diagnostic contexts.

Funding

This work was supported by Human Enhancement and Learning (HEaL) research programme of Maastricht University, by the BRAIN-TRAIN consortium under the EU FP7 Health Cooperation Work Program (602186), and Bial Grants for Scientific Research 299/14.

Conflicts of interest

The authors declare no conflict of interests.

Acknowledgements:

We thank Talma Hendler for valuable conceptual and material help, Yoav Benjamini for statistical consultancy and Halen Baker for methodological assistance.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2017.09.032.

References

Adolphs, R., Nummenmaa, L., Todorov, A., Haxby, J.V., 2016. Data-driven approaches in the investigation of social perception. Phil Trans. R. Soc. B 371, 20150367. https://doi.org/10.1098/rstb.2015.0367.

Aronofsky, D., 2010. Black Swan.

Baldassano, C., Fei-Fei, L., Beck, D.M., 2016. Pinpointing the peripheral bias in neural scene-processing networks during natural viewing. J. Vis. 16 https://doi.org/10.1167/16.2.9, 9–9.

Barnich, O., Droogenbroeck, M.V., 2011. ViBe: a universal background subtraction algorithm for video sequences. IEEE Trans. Image Process 20, 1709–1724. https://doi.org/10.1109/TIP.2010.2101613.
Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 1165–1188.
Bertalmío, M., Caselles, V., Provenzi, E., 2009. Issues about Retinex theory and contrast enhancement. Int. J. Comput. Vis. 83, 101–119. https://doi.org/10.1007/s11263-009-0221-5.
Bigler, E.D., Mortensen, S., Neeley, E.S., Ozonoff, S., Krasny, L., Johnson, M., Lu, J., Provencal, S.L., McMahon, W., Lainhart, J.E., 2007. Superior temporal gyrus, language function, and autism. Dev. Neuropsychol. 31, 217–238. https://doi.org/10.1080/87565640701190841.
Bishop, C., 2007. Pattern Recognition and Machine Learning. Springer, New York.
Bishop, L., Bailes, F., Dean, R.T., 2013. Musical expertise and the ability to imagine loudness. PLOS ONE 8, e56052. https://doi.org/10.1371/journal.pone.0056052.
Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., Carpenter, T.A., Brammer, M., 2001. Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. Hum. Brain Mapp. 12, 61–78. https://doi.org/10.1002/1097-0193(200102)12:2<61::AID-HBM1004>3.0.CO;2-W.
Cardin, V., Smith, A.T., 2011. Sensitivity of human visual cortical area V6 to stereoscopic depth gradients associated with self-motion. J. Neurophysiol. 106, 1240–1249. https://doi.org/10.1152/jn.01120.2010.
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, e13812. https://doi.org/10.1371/journal.pone.0013812.
Chen, M., Han, J., Hu, X., Jiang, X., Guo, L., Liu, T., 2013. Survey of encoding and decoding of visual stimulus via FMRI: an image analysis perspective. Brain Imaging Behav. 8, 7–23. https://doi.org/10.1007/s11682-013-9238-z.
Chen, P.-H.C., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., Ramadge, P.J., 2015. A reduced-dimension fMRI shared response model. In: Advances in Neural Information Processing Systems, pp. 460–468.
Chu, C., Ni, Y., Tan, G., Saunders, C.J., Ashburner, J., 2011. Kernel regression for fMRI pattern prediction. NeuroImage, Multivar. Decoding Brain Read. 56, 662–673. https://doi.org/10.1016/j.neuroimage.2010.03.058.
Cohen, J.R., Asarnow, R.F., Sabb, F.W., Bilder, R.M., Bookheimer, S.Y., Knowlton, B.J., Poldrack, R.A., 2011. Decoding continuous variables from neuroimaging data: basic and clinical applications. Front. Neurosci. 5 https://doi.org/10.3389/fnins.2011.00075.
Cohen, J.R., Asarnow, R.F., Sabb, F.W., Bilder, R.M., Bookheimer, S.Y., Knowlton, B.J., Poldrack, R.A., 2010. Decoding developmental differences and individual variability in response inhibition through predictive analyses across individuals. Front. Hum. Neurosci. 4 (47) https://doi.org/10.3389/fnhum.2010.00047.
Columbus, C., 1998. Stepmom.
Connolly, A.C., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y.O., Wu, Y.-C., Abdi, H., Haxby, J.V., 2012. The representation of biological classes in the human brain. J. Neurosci. 32, 2608–2618. https://doi.org/10.1523/JNEUROSCI.5547-11.2012.
Conroy, B.R., Singer, B.D., Guntupalli, J.S., Ramadge, P.J., Haxby, J.V., 2013. Inter-subject alignment of human cortical anatomy using functional connectivity. NeuroImage 81, 400–411. https://doi.org/10.1016/j.neuroimage.2013.05.009.
Cronenberg, D., 1986. The Fly.
Daubechies, I., et al., 1992. Ten Lectures on Wavelets. SIAM.
Dupont, P., Bruyn, B.D., Vandenberghe, R., Rosier, A.M., Michiels, J., Marchal, G., Mortelmans, L., Orban, G.A., 1997. The kinetic occipital region in human visual cortex. Cereb. Cortex 7, 283–292. https://doi.org/10.1093/cercor/7.3.283.
Farrelly, P., Farrelly, B., 1998. There is Something About Mary.
Fiorentini, A., 2004. Brightness and color. In: Chalupa, L.M., Werner, J.S. (Eds.), The Visual Neurosciences. MIT Press, pp. 881–891.
Fincher, D., 1995. Se7en.
Fischer, E., Bülthoff, H.H., Logothetis, N.K., Bartels, A., 2012. Visual motion responses in the posterior cingulate sulcus: a comparison to V5/MT and MST. Cereb. Cortex N. Y. N. 1991 (22), 865–876. https://doi.org/10.1093/cercor/bhr154.
Fletcher, H., Munson, W.A., 1933. Loudness, its definition, measurement and calculation*. Bell Syst. Tech. J. 12, 377–430. https://doi.org/10.1002/j.1538-7305.1933.tb00403.x.
Formisano, E., Goebel, R., 2003. Tracking cognitive processes with functional MRI mental chronometry. Curr. Opin. Neurobiol. 13, 174–181. https://doi.org/10.1016/S0959-4388(03)00044-8.
Frost, M.A., Goebel, R., 2013. Functionally informed cortex based alignment: an integrated approach for whole-cortex macro-anatomical and ROI-based functional alignment. NeuroImage 83, 1002–1010. https://doi.org/10.1016/j.neuroimage.2013.07.056.
Frost, M.A., Goebel, R., 2012. Measuring structural–functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment. NeuroImage 59, 1369–1381. https://doi.org/10.1016/j.neuroimage.2011.08.035.
Fujiwara, Y., Miyawaki, Y., Kamitani, Y., 2009. Estimating image bases for visual image reconstruction from human brain activity. In: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22. Curran Associates, Inc, pp. 576–584.
Gilchrist, A., Kossyfidis, C., Bonato, F., Agostini, T., Cataliotti, J., Li, X., Spehar, B., Annan, V., Economou, E., 1999. An anchoring theory of lightness perception. Psychol. Rev. 106, 795–834.
Golub, G.H., Heath, M., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21, 215–223. https://doi.org/10.1080/00401706.1979.10489751.
Guntupalli, J.S., Hanke, M., Halchenko, Y.O., Connolly, A.C., Ramadge, P.J., Haxby, J.V., 2016. A model of representational spaces in human cortex. Cereb. Cortex 26, 2919–2934. https://doi.org/10.1093/cercor/bhw068.
Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject synchronization of cortical activity during natural vision. Science 303, 1634–1640. https://doi.org/10.1126/science.1089506.
Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 87, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067.
Haxby, J.V., 2012. Multivariate pattern analysis of fMRI: the early beginnings, 20 YEARS OF fMRI20 YEARS OF fMRI 62 NeuroImage 852–855. https://doi.org/10.1016/j.neuroimage.2012.03.016.
Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430. https://doi.org/10.1126/science.1063736.
Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., Ramadge, P.J., 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72, 404–416. https://doi.org/10.1016/j.neuron.2011.08.026.
Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., Passingham, R.E., 2007. Reading hidden intentions in the human brain. Curr. Biol. CB 17, 323–328. https://doi.org/10.1016/j.cub.2006.11.072.
Heller, R., Golland, Y., Malach, R., Benjamini, Y., 2007. Conjunction group analysis: an alternative to mixed/random effect analysis. NeuroImage 37, 1178–1185. https://doi.org/10.1016/j.neuroimage.2007.05.051.
Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402. https://doi.org/10.1038/nrn2113.
Horikawa, T., Tamaki, M., Miyawaki, Y., Kamitani, Y., 2013. Neural decoding of visual imagery during sleep. Science 340, 639–642. https://doi.org/10.1126/science.1234330.
Howard, R.J., Brammer, M., Wright, I., Woodruff, P.W., Bullmore, E.T., Zeki, S., 1996. A direct demonstration of functional specialization within motion-related visual and auditory cortex of the human brain. Curr. Biol. CB 6, 1015–1019.
Hudspeth, A.J., Jessell, T.M., Kandel, E.R., Schwartz, J.H., Siegelbaum, S.A., 2013. Principles of Neural Science.
Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224. https://doi.org/10.1016/j.neuron.2012.10.014.
Juslin, P.N., Laukka, P., 2003. Communication of emotions in vocal expression and music performance: different channels, same code? Psychol. Bull. 129, 770–814. https://doi.org/10.1037/0033-2909.129.5.770.
Kahnt, T., Heinzle, J., Park, S.Q., Haynes, J.-D., 2011. Decoding different roles for vmPFC and dlPFC in multi-attribute decision making. NeuroImage, Multivar. Decoding Brain Read. 56, 709–715. https://doi.org/10.1016/j.neuroimage.2010.05.058.
Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., others, 1983. Optimization by simulated annealing. Science 220, 671–680.
Kohler, P.J., Fogelson, S.V., Reavis, E.A., Meng, M., Guntupalli, J.S., Hanke, M., Halchenko, Y.O., Connolly, A.C., Haxby, J.V., Tse, P.U., 2013. Pattern classification precedes region-average hemodynamic response in early visual cortex. NeuroImage 78, 249–260. https://doi.org/10.1016/j.neuroimage.2013.04.019.
Kubrick, S., 1980. The Shining.
Land, E.H., McCann, J.J., 1971. Lightness and Retinex theory. JOSA 61, 1–11. https://doi.org/10.1364/JOSA.61.000001.
Larsson, J., Heeger, D.J., 2006. Two retinotopic visual areas in human lateral occipital cortex. J. Neurosci. 26, 13128–13142. https://doi.org/10.1523/JNEUROSCI.1657-06.2006.
Larsson, J., Heeger, D.J., Landy, M.S., 2010. Orientation selectivity of motion-boundary responses in human visual cortex. J. Neurophysiol. 104, 2940–2950. https://doi.org/10.1152/jn.00400.2010.
Lehky, S.R., Tanaka, K., 2016. Neural representation for object recognition in inferotemporal cortex. Curr. Opin. Neurobiol., Neurobiol. cognitive Behav. 37, 23–35. https://doi.org/10.1016/j.conb.2015.12.001.
Limare, N., Petro, A.B., Sbert, C., Morel, J.-M., 2011. Retinex Poisson Equation: a Model for Color Perception. Image Process. Line 1. https://doi.org/10.5201/ipol.2011.lmps_rpe.
Manners, K., 1996. The X-Files, the Episode "Home".
Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M., Morito, Y., Tanabe, H.C., Sadato, N., Kamitani, Y., 2008. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron 60, 915–929. https://doi.org/10.1016/j.neuron.2008.11.004.
Mograbi, A., 2006. Avenge But One of My Two Eyes.
Morel, J.M., Petro, A.B., Sbert, C., 2010. A PDE formalization of Retinex theory. IEEE Trans. Image Process 19, 2825–2837. https://doi.org/10.1109/TIP.2010.2049239.
Murray, M.M., Thelen, A., Thut, G., Romei, V., Martuzzi, R., Matusz, P.J., 2016. The multisensory function of the human primary visual cortex. Neuropsychologia, special issue: functional selectivity in perceptual and cognitive systems. A Tribute Shlomo Bentin 1946–2012 (83), 161–169. https://doi.org/10.1016/j.neuropsychologia.2015.08.011.
Nakata, H., 2005. The Ring Two.
Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. 21, 1641–1646. https://doi.org/10.1016/j.cub.2011.08.031.
Pakula, A.J., 1982. Sophie's Choice.

Panagiotakis, C., Tziritas, G., 2005. A speech/music discriminator based on RMS and zero-crossings. IEEE Trans. Multimed. 7, 155–166. https://doi.org/10.1109/TMM.2004.840604.

Pikrakis, A., Theodoridis, S., 2014. Speech-music discrimination: a deep learning perspective. In: 2014 22nd European Signal Processing Conference (EUSIPCO). Presented at the 2014 22nd European Signal Processing Conference (EUSIPCO), pp. 616–620.

Pitzalis, S., Fattori, P., Galletti, C., 2015. The human cortical areas V6 and V6A. Vis. Neurosci. 32 (E007), 15. https://doi.org/10.1017/S0952523815000048.

Poldrack, R.A., Halchenko, Y.O., Hanson, S.J., 2009. Decoding the large-scale structure of brain function by classifying mental states across individuals. Psychol. Sci. 20, 1364–1372. https://doi.org/10.1111/j.1467-9280.2009.02460.x.

Rammsayer, T., Altenmüller, E., 2006. Temporal information processing in musicians and nonmusicians. Music Percept. Interdiscip. J. 24, 37–48. https://doi.org/10.1525/mp.2006.24.1.37.

Raz, G., Jacob, Y., Gonen, T., Winetraub, Y., Soreq, E., Flash, T., Hendler, T., 2013. Cry for her or cry with her: context-dependent dissociation of two modes of cinematic empathy reflected in network cohesion dynamics. Soc. Cogn. Affect. Neurosci. https://doi.org/10.1093/scan/nst052.

Raz, G., Touroutoglou, A., Wilson-Mendenhall, C., Gilam, G., Lin, T., Gonen, T., Jacob, Y., Atzil, S., Admon, R., Bleich-Cohen, M., Maron-Katz, A., Hendler, T., Barrett, L.F., 2016. Functional connectivity dynamics during film viewing reveal common networks for different emotional experiences. Cogn. Affect. Behav. Neurosci. 1–15. https://doi.org/10.3758/s13415-016-0425-4.

Raz, G., Winetraub, Y., Jacob, Y., Kinreich, S., Maron-Katz, A., Shaham, G., Podlipsky, I., Gilam, G., Soreq, E., Hendler, T., 2012. Portraying emotions at their unfolding: a multilayered approach for probing dynamics of neural networks. NeuroImage 60, 1448–1461. https://doi.org/10.1016/j.neuroimage.2011.12.084.

Rottenberg, J., Ray, R.R., Gross, J.J., 2007. Emotion elicitation using films. Handb. Emot. Elicitation Assess. 9–28.

Sabuncu, M.R., Singer, B.D., Conroy, B., Bryan, R.E., Ramadge, P.J., Haxby, J.V., 2010. Function-based intersubject alignment of human cortical anatomy. Cereb. Cortex 20, 130–140. https://doi.org/10.1093/cercor/bhp085.

Schaefer, A., Nils, F., Sanchez, X., Philippot, P., 2010. Assessing the effectiveness of a large database of emotion-eliciting films: a new tool for emotion researchers. Cogn. Emot. 24, 1153–1172.

Schoenmakers, S., Barth, M., Heskes, T., van Gerven, M., 2013. Linear reconstruction of perceived images from human brain activity. NeuroImage 83, 951–961. https://doi.org/10.1016/j.neuroimage.2013.07.043.

Shinkareva, S.V., Mason, R.A., Malave, V.L., Wang, W., Mitchell, T.M., Just, M.A., 2008. Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. PLOS ONE 3, e1394. https://doi.org/10.1371/journal.pone.0001394.

Silbert, L.J., Honey, C.J., Simony, E., Poeppel, D., Hasson, U., 2014. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. Proc. Natl. Acad. Sci. 111, E4687–E4696. https://doi.org/10.1073/pnas.1323812111.

Singer, N., Jacoby, N., Lin, T., Raz, G., Shpigelman, L., Gilam, G., Granot, R.Y., Hendler, T., 2016. Common modulation of limbic network activation underlies musical emotions as they unfold. NeuroImage 141, 517–529. https://doi.org/10.1016/j.neuroimage.2016.07.002.

Spielberg, S., 1998. Saving Private Ryan.

Thomas, P., 1997. Alaska's Wild Denali.

Valente, G., Castellanos, A.L., Vanacore, G., Formisano, E., 2014. Multivariate linear regression of high-dimensional fMRI data with multiple target variables. Hum. Brain Mapp. 35, 2163–2177. https://doi.org/10.1002/hbm.22318.

Valente, G., De Martino, F., Esposito, F., Goebel, R., Formisano, E., 2011. Predicting subject-driven actions and sensory experience in a virtual world with Relevance Vector Machine Regression of fMRI data. NeuroImage. Multivar. Decoding Brain Read. 56, 651–661. https://doi.org/10.1016/j.neuroimage.2010.09.062.

van Gerven, M.A.J., de Lange, F.P., Heskes, T., 2010. Neural decoding with hierarchical generative models. Neural Comput. 22, 3127–3142. https://doi.org/10.1162/NECO_a_00047.

Vanduffel, W., Zhu, Q., 2015. Topographic Layout of Monkey Extrastriate Visual Cortex.

Vinberg, J., Grill-Spector, K., 2008. Representation of shapes, edges, and surfaces across multiple cues in the human visual cortex. J. Neurophysiol. 99, 1380–1393. https://doi.org/10.1152/jn.01223.2007.

Wall, M.B., Smith, A.T., 2008. The representation of egomotion in the human brain. Curr. Biol. 18, 191–194. https://doi.org/10.1016/j.cub.2007.12.053.

Weir, P., 1989. Dead Poet Society.

Weisstein, E.W., 2017. Least Squares Fitting–logarithmic – from Wolfram MathWorld [WWW Document]. URL. http://mathworld.wolfram.com/LeastSquaresFittingLogarithmic.html (Accessed 07 14 2017).

Zemeckis, R., 1994. Forrest Gump.

Zendel, B.R., Alain, C., 2012. Musicians experience less age-related decline in central auditory processing. Psychol. Aging 27, 410–417. https://doi.org/10.1037/a0024816.

Zhu, X., Ramanan, D., 2012. Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886. https://doi.org/10.1109/CVPR.2012.6248014.