Representational Gradients of Musical Information and Evoked Emotions Revealed by CNN

MAX PLANCK INSTITUTE
FOR EMPIRICAL AESTHETICS

Seung-Goo Kim[1], **Tobias Overath**[2], **Daniela Sammler**[1,3]

[1] Research Group Neurocognition of Music and Language, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany
[2] Department of Psychology and Neuroscience, Duke University, Durham, NC, USA
[3] Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Introduction

Background. Recent work from our group suggests that a pretrained audio convolutional neural network (CNN; VGGish) can capture information in real-world music that is relevant to evoked emotions and brain activity in the medial prefrontal cortex (mPFC)[1]. However, we only focused on the relevance of the final layer-space.

Here, we explored the neural encoding of all layer-spaces and their relationships to music-evoked emotions. In particular, we investigated whether the representational gradient of increasing abstraction—from superficial to deep layers of the CNN—bares any resemblance with the well-established functional gradient in the human cerebral cortex—from unimodal sensory to transmodal associative regions[2].

Research Questions

Q1. Would increasingly abstract representations of music in different layers of the CNN be encoded along the axis of the functional gradient[2]?

Q2. How do layer-specific CNN embeddings predict human behavioral ratings of music-evoked emotions?

Methods

Open-access fMRI dataset. openneuro-ds003085[3] (*n* = 37, mean age = 24)

- Imaging: 3-T EPI (multiband = 8x, TR = 1 s, 3-mm isovoxel, whole brain)
- Musical pieces: "happy" (2 min 48 s), "sad-short" (4 min 16 s), "sad-long" (8 min 35 s) in styles of movie soundtracks
- Continuous ratings: "Felt Emotionality" and "Enjoyment" after scanning

Feature extraction from music. High-dimensional embeddings (128–393k dimensions) were extracted from the 24 layers of the VGGish network. For every layer, the first 50 principal components explaining 40–98% of the total variance were used as predictors in linearized encoding models **(Table 1)**.

Encoding models. Independent models were fit for each layer to predict either fMRI timeseries in each voxel or emotional ratings (both with lags of 4, 5, 6 s) as shown in **Figure 1**. Layer-wise profiles of prediction accuracies were inspected in:

- Regions-of-interest (ROIs): superior and middle temporal gyri (STG & MTG), inferior frontal gyrus (IFG), and medial prefrontal cortex (mPFC)[1].
- Principal components (PCs): PCA was applied to the matrix of 24 prediction accuracies x #voxels, to identify topographies of the layer-wise profiles.

Representational gradient mapping. The spatial correspondence between the superficial-to-deep-layer and unimodal-to-transmodal gradients was tested.

- After surface projection, a best (argmax) and a centroid layer were determined in the profiles of prediction accuracies at each vertex **(Figure 5a, b)**, which we call "representational gradients".
- The correspondence with the functional gradient[2] (recreated from the template data included in BrainSpace v0.1.10) **(Figure 5c)** was statistically tested using a geometrical permutation test ("spin-test"[4]), which involves 10,000 random rotations of spherical coordinates of the surface-mapped data.
- As negative control, best and centroid layers from encoding models with negative lags (-6, -5, -4 s) were used **(Figure 5d, e)**.

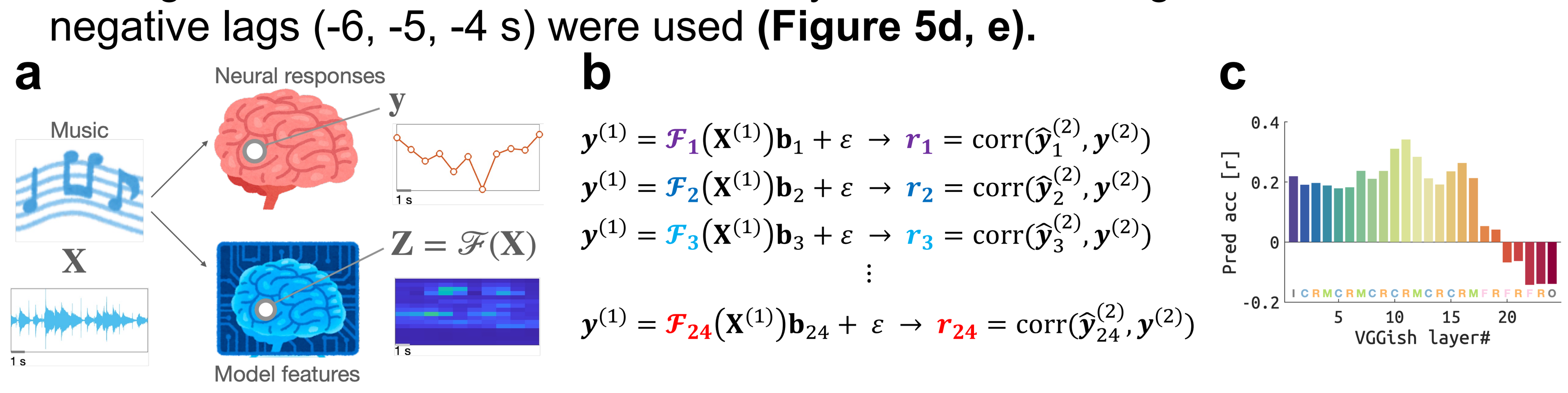

Figure 1. Analysis overview. (a) Linearized encoding analysis[5]. (b) Models for the embeddings ($\mathbf{Z}$) of each layer: $\mathbf{y}$ is either the fMRI time series or the emotional ratings; $\mathbf{X}$ is the Mel-spectrogram of a music sample, $\mathcal{F}_i$ is the truncated embedding function of the *i*–th layer of the VGGish network; superscriptions in parentheses indicate cross-validation partitions [1=training, 2=testing]. (c) Prediction accuracies (Pearson correlation coefficients) result in a layer-wise prediction profiles for each voxel.

Layer#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Type	Input	Conv	RL	MP	Conv	RL	MP	Conv	RL	Conv	RL	MP	Conv	RL	Conv	RL	MP	FC	RL	FC	RL	FC	RL	Output
#Dim	6144	393216	393216	98304	196608	196608	49152	98304	98304	98304	98304	24576	49152	49152	49152	49152	12288	4096	4096	4096	4096	128	128	128
Exp%	80	63.9	63.9	64.6	55.1	53.6	52.8	40.9	40.3	42.9	41.5	48.3	46.5	46.8	43.5	42.9	49.1	71.5	71.8	87.8	87.9	97.6	97.6	97.6

Table 1. Dimensionality of VGGish embeddings per layer. Conv, convolutional; RL, rectified linear unit; MP: max pooling; FC: fully connected. Exp%: explained variance by 50 principal components.

Conclusions

C1. The transformation of the auditory information along the functional gradient may involve an abstraction mechanism similar to what the CNN implements.

C2. Basic and aesthetic emotional experiences may depend on different abstraction levels of the audio signal represented along the functional gradient.

Results

R0. Layer-wise neural encoding of VGGish embeddings

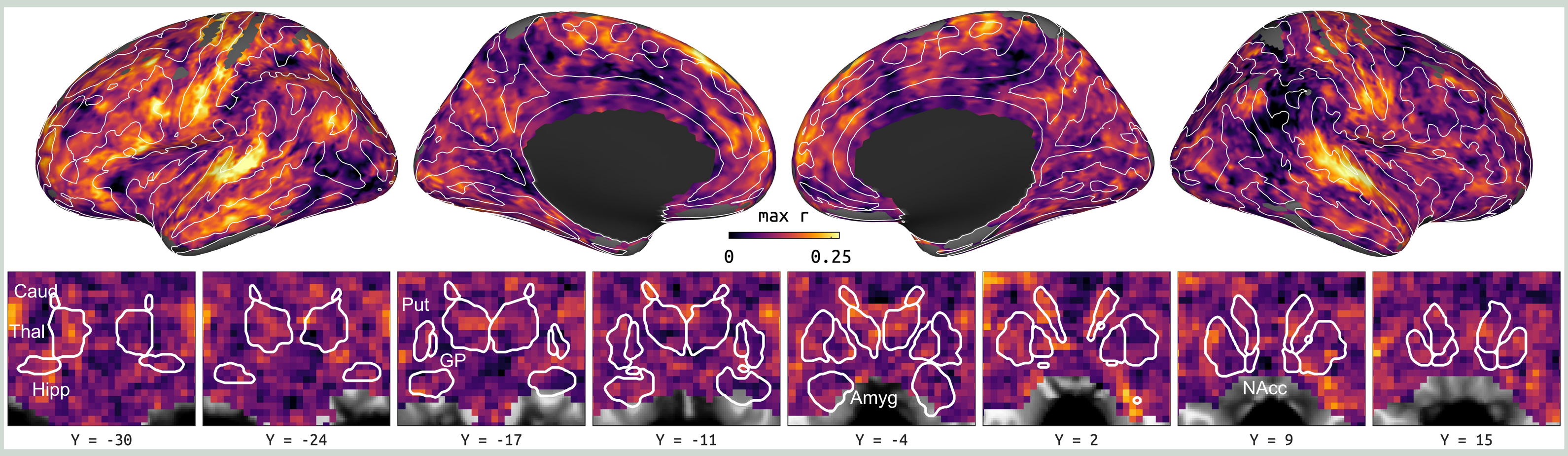

Figure 2. Maximal prediction accuracy across all layers. Hipp, hippocampus; Thal, thalamus; Caud, caudate nucleus; Put, Putamen; GP, Globus pallidum; Amyg, amygdala; NAcc, Nucleus accumbens

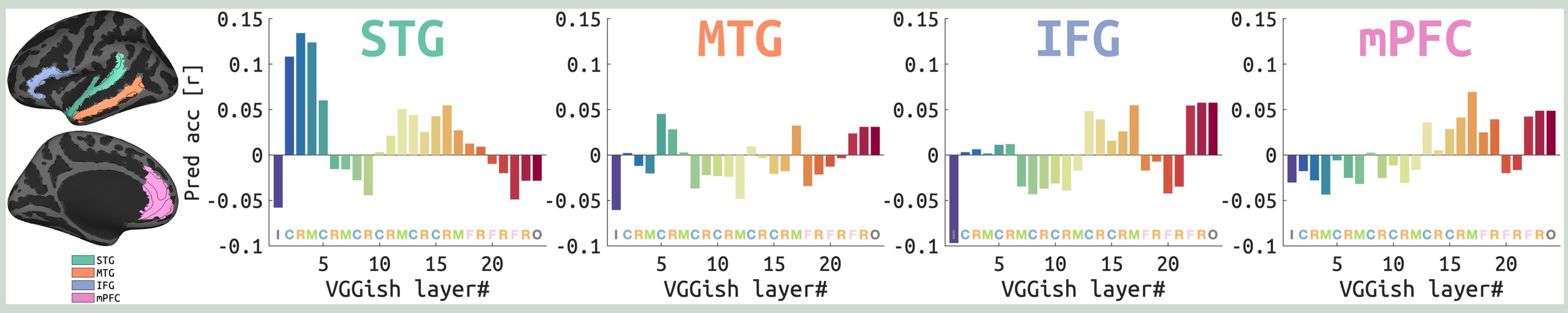

Figure 3. Profiles of prediction accuracies in the four ROIs.

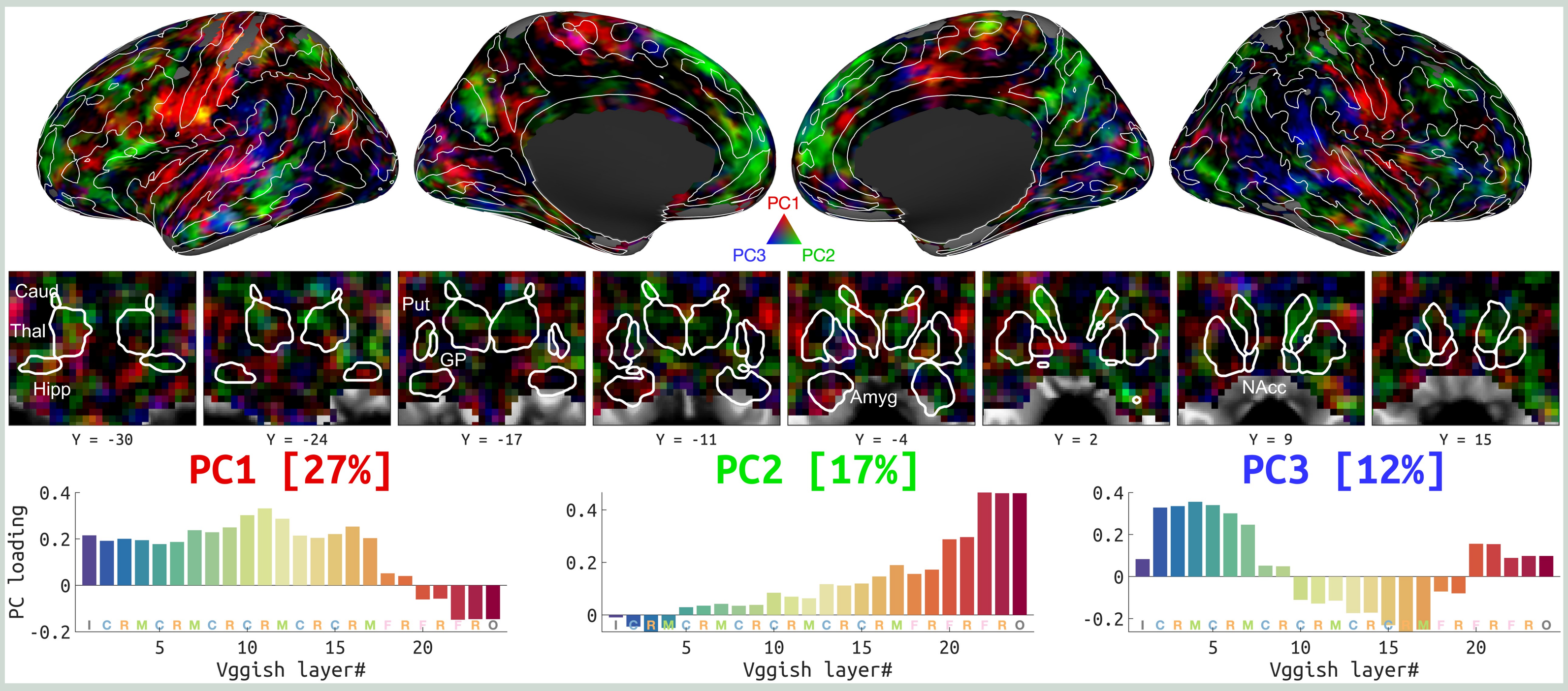

Figure 4. Topographies and profiles of prediction accuracies of the first 3 PCs. (top) RGB values indicate scaled positive PC scores. (bottom) PC loadings highlight main contributions from superficial/middle layers to PC1, from deep layers to PC2, and from superficial layers to PC3. Abbreviations are the same as in Figure 2.

R1. VGGish representational gradient maps onto functional gradient

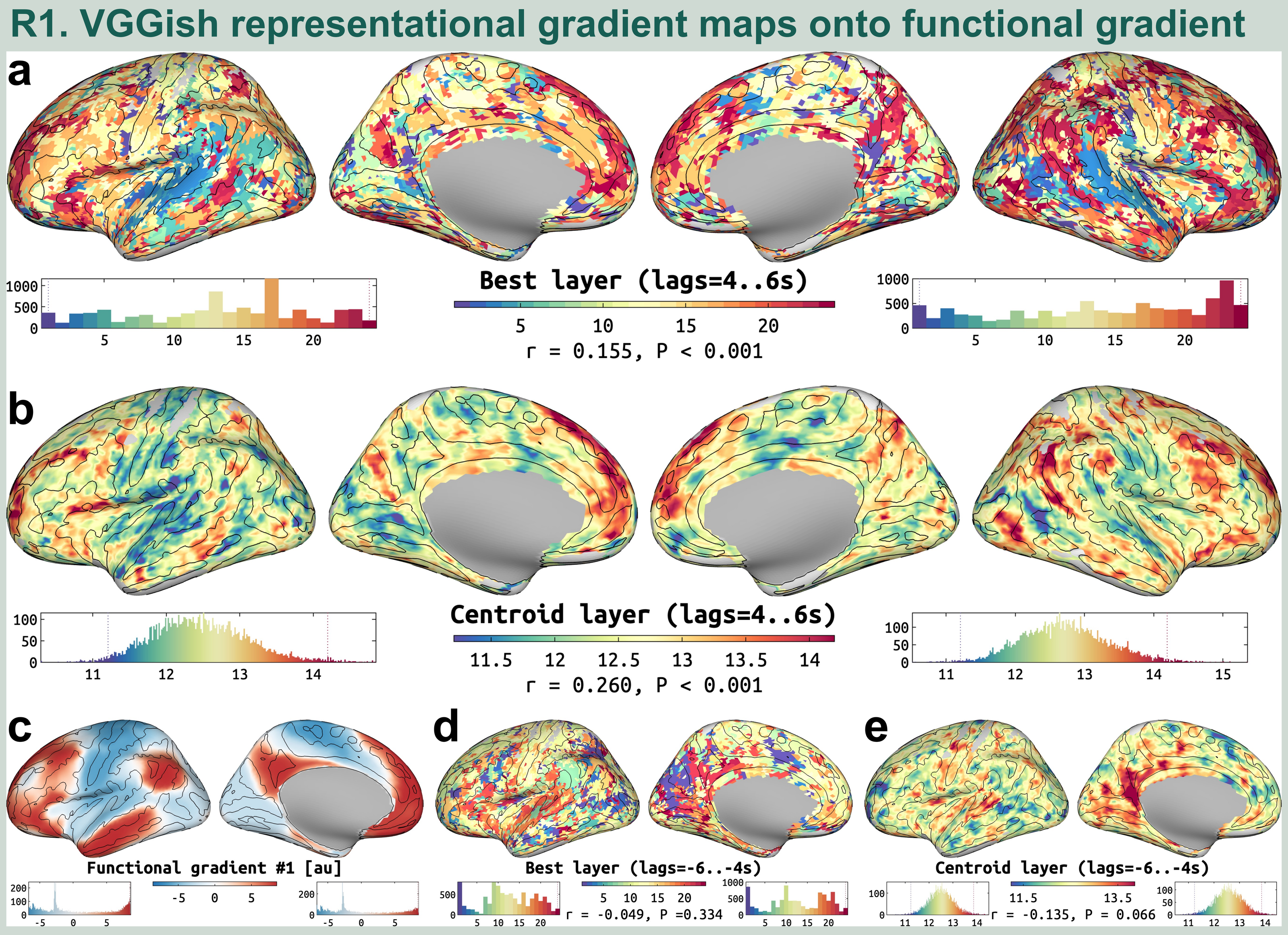

Figure 5. Representational gradient of VGGish layers. (a) Best [i.e., argmax] layers with positive lags. (b) Centroid layers with positive lags. Spin-test results (correlation coefficient and P-value) in (a, b) indicate significant correspondence to (c) the first functional gradient axis[2]. (d) Best and (e) centroid layers from encoding models with negative lags showed no correspondence to the functional gradient.

R2. Distinct encoding patterns of emotional ratings

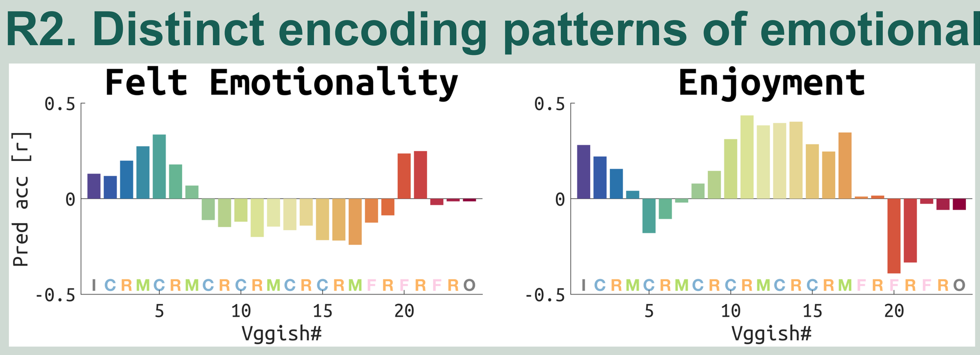

Figure 6. Layer-wise profiles. (left) Felt Emotionality highlighted superficial layers [similarly to PC3]. (right) Enjoyment highlighted middle layers.

Contact: Seung-Goo Kim
seung-goo.kim@ae.mpg.de
https://www.ae.mpg.de/
+49 69 8300 479 851

Acknowledgement: We thank Matthew Sachs for the openneuro data and continuous rating data and irasutoya.com for image sources.

References: **[1]** Kim et al., 2023, doi:10.6084/m9.figshare.24085104 **[2]** Margulies et al., 2016, doi:10.1073/pnas.1608282113 **[3]** Sachs et al., 2020, doi:10.1016/j.neuroimage.2019.116512 **[4]** Alexander-Bloch, 2018, doi:10.1016/j.neuroimage.2018.05.070 **[5]** Kim, 2022, doi:10.3389/fnins.2022.928841