Emotion-relevant Representations of Music Extracted by Convolutional Neural Networks Are Encoded in Medial Prefrontal Cortex

Seung-Goo Kim (a), Tobias Overath (b), Daniela Sammler (c)

(a) Research Group Neurocognition of Music and Language, Max Planck Institute for Empirical Aesthetics, Germany, seung-goo.kim@ae.mpg.de (b) Department of Psychology and Neuroscience, Duke University, USA, t.overath@duke.edu (c) Research Group Neurocognition of Music and Language, Max Planck Institute for Empirical Aesthetics, daniela.sammler@ae.mpg.de

Keywords: Music-evoked emotion, fMRI, Machine Learning, Artificial neural network, Real-world music

Introduction

Music often evokes strong emotions. How auditory information of music is transformed across neural processing levels and how these levels contribute to the emergence of felt emotions is still debated. Recently, the embeddings of convolutional neural networks (CNNs) trained to generate "semantic" labels for sounds were found to be better predictors for the emotions that human raters recognized in music than traditional low-level audio features (Koh & Dubnov, 2021).

In the current work, we compared the predictive performance of CNN embeddings and traditional audio features for felt emotions and neural activity during music listening. Moreover, we explored direct associations between felt emotions and neural activity to shed light on emotional processes beyond auditory processes.

Methods

We used the functional MRI dataset of Sachs et al. (2020) [openneuro-ds003085]. During fMRI data acquisition, 37 participants (mean age = 24) listened to one "happy" and two "sad" instrumental musical pieces (3–8 min). After scanning, they rated the intensity of their instantaneous felt emotions: "emotionality" (how happy or sad) and "enjoyment" in two separate sessions.

- *Audio features*: (1) A broadband (180–7040 Hz) cochleogram envelope was created using NSL tools as a low-level feature, (2) CNN embeddings (final layer activation) from OpenL3 and VGGish were extracted as mid-level features.
- *Audio-to-emotion*: Audio features and ratings were cross-correlated with lags from 0 to 5 s. Maximum correlations were compared across features (envelope, OpenL3, VGGish) with repeated measures ANOVAs.
- *Encoding analyse*s: The fMRI time-series was modeled by ridge regression (Huth et al., 2016) with predictors of either audio features (audio-to-brain) or ratings (emotion-to-brain).

Results

- *Audio-to-emotion*: Changes in CNN embeddings (OpenL3) revealed stronger cross-correlations to changes in emotion ratings than the envelope ($P < 0.0001$).
- *Audio-to-brain*: All audio features were uniquely encoded in dorsomedial prefrontal cortex (dmPFC; $P < 0.031$). Only the envelope, not the CNN feature (VGGish), showed their encoding in auditory cortex ($P < 0.01$).
- *Emotion-to-brain*: Finally, changes in emotionality ratings predicted future fMRI activity in distributed regions including dmPFC and temporal cortices (min $P = 0.001$), while activity in major default mode network nodes (vmPFC and precuneus) predicted future changes in enjoyment ratings (min $P = 0.01$).

Discussion and Conclusion

CNN embeddings carry audio information relevant for emotional responses, beyond low-level audio features. In particular, the CNN encoding confined in mPFC suggests that mid-level representations of music contribute to the emergence of emotions. The differential neural encoding of emotionality and enjoyment may reflect distinct mechanisms of felt emotions and aesthetic judgements (Juslin, 2013).

References

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453-458.

Juslin, P. N. (2013). From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions. *Physics of Life Reviews*, *10*(3), 235-266.

Koh, E., & Dubnov, S. (2021). Comparison and analysis of deep audio embeddings for music emotion recognition. AAAI Workshop on Affective Content Analysis, Virtual.

Sachs, M. E., Habibi, A., Damasio, A., & Kaplan, J. T. (2020). Dynamic intersubject neural synchronization reflects affective responses to sad music. *Neuroimage*, *218*, 116512.