A real-time multimodal system for music preference decoding combining EEG and acoustic features

Thomas S. Binns [1*,2], Shinichi Furuya [2], Vincent K.M. Cheung [2]

1. Bernstein Center for Computational Neuroscience Berlin, Germany
2. Sony Computer Science Laboratories, Inc., Tokyo, Japan

* Work completed during internship at Sony Computer Science Laboratories, Tokyo

INTRODUCTION

- EEG recordings of brain activity offer objective measures of emotion, potentially improving music recommendation systems.
- However, previous attempts to decode emotions and preference for music from EEG have:
  1. Shown limited success.
  2. Used music familiar to users, not representing the typical use case of recommendation for novel songs.
  3. Rarely demonstrated real-time implementations.
- We present a multimodal system combining acoustic and neural features for preference decoding of novel songs.
- We demonstrate the system's feasibility in a real-time scenario, controlling song presentation based on predicted preferences.

METHODS

Music stimuli and data collection

- 300 anime and VOCALOID songs were randomly chosen from Kiite [1] and cut from the beginning of the introduction sections.
- EEG was collected from an active 64-channel system, streamed at 200 Hz, and referenced to the common average.
- Audio data was streamed as a single channel at 16,000 Hz.

Training phase

- Users (n = 3) listened to the first 6 s from 150 songs and rated how much they wanted to listen (1-9), collecting ratings into 3 classes: 1) 1-3, wanted to skip; 2) 4-6, neutral; 3) 7-9, wanted to listen.
- EEG and audio data was divided into 3 s segments with 1 s overlap.

Decoding pipelines

- **EEG:** filter bank common spatial pattern filters [2] were fit from 4-40 Hz in 4 Hz intervals (Fig. 1A); classification used an SVM with an RBF kernel, scaled gamma, and balanced class weights.
- **Audio:** features were extracted using pre-trained musicnn models [3] and PCA; classification used an SVM with an RBF kernel, scaled gamma, and balanced class weights.
- **Ensemble:** logistic regression classifier with balanced class weights.
- Optimal hyperparameters were identified for the cohort in a grid search using a 5-fold stratified, group cross-validation:
  - **EEG:** 9 components per frequency band; SMOTE feature oversampling (5 neighbours) [4]; C of 10.
  - **Audio:** mean and max pool features of musicnn's 'MSD big' model; PCA components explaining 80% variance of features; no oversampling; C of 10.
  - **Stacking:** SMOTE oversampling (3 neighbours); C of 0.1; elastic net regularisation (0.15 L1 ratio); intercept fitted.

Real-time decoding phase

- Users (n = 2) listened to the first 10 s from 75 other songs (Fig. 1B).
- EEG and audio data was preprocessed in real-time and fed through the decoding pipelines to produce preference predictions.
- Trials ended when: 1) the system predicted the user "wanted to skip"; 2) the user skipped; or 3) the system predicted the user "wanted to listen", and the user did not skip.
- For a baseline performance estimate, users (n = 1) listened to the final 75 songs with the preference predictions randomly generated.

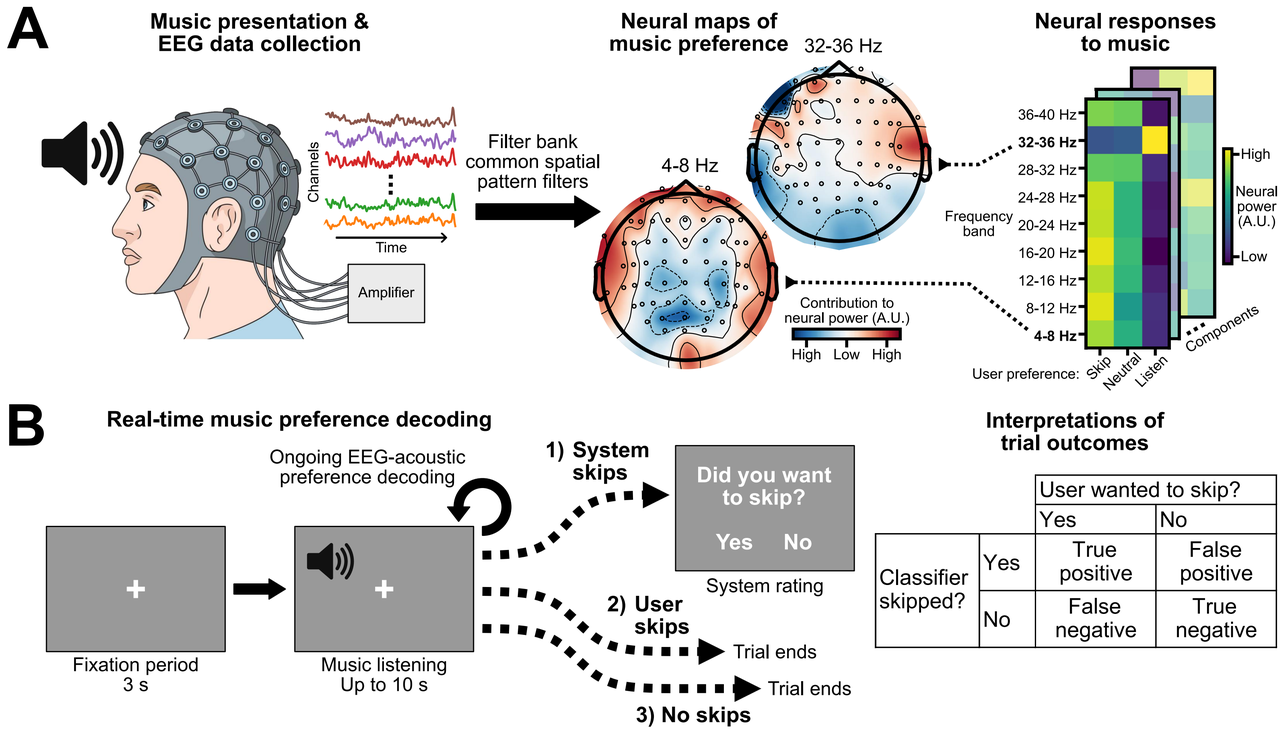

Figure 1: Neural representations of music and real-time preference decoding. (A) Example spatial maps of the neural contributions to preference extracted from filter bank common spatial pattern filters fit to the EEG data during music listening, alongside examples of the associated neural features used for preference decoding. **(B)** Schematic representation of real-time preference decoding, the possible trial outcomes, and their interpretations.

RESULTS

Optimised prediction accuracies

- Hyperparameters with the highest mean balanced accuracy on a held-out test set of 15% of trials across users in the 3-class problem were:

EEG 37.3 ± 0.9%
Audio 42.8 ± 1.5%
→ **Stacked** 46.6 ± 1.6% (mean ± SD)

Real-time system performance – *using unoptimised hyperparameters*

- The system achieved a precision of 55.6 ± 9.1% and recall of 94.1 ± 8.3%, superior to the random system's 40% accuracy and 44% recall (Table 1).
- Latencies from data acquisition to final predictions (based on the preceding 3 s of data) were on average 80.5 ± 7.6 ms, with 43.5 ± 3.8 ms for the EEG pipeline, and 32.9 ± 4.1 ms for the audio pipeline.

Real system		User wanted to skip?	
		Yes	No
Classifier skipped?	Yes	47 ± 4%	38 ± 11%
	No	3 ± 4%	12 ± 11%

Random system		User wanted to skip?	
		Yes	No
Classifier skipped?	Yes	24 ± 0%	35 ± 0%
	No	29 ± 0%	12 ± 0%

Table 1: Real-time preference decoding outcomes. Trial outcomes for the real system with preference predictions based on EEG and audio data (left; n_{users} = 2), and the baseline comparison system with preference predictions randomly generated (right; n_{users} = 1). Mean ± SD shown.

DISCUSSION

- Acoustic features offered better decoding performance than EEG features, but combining acoustic and EEG features offered optimal performance, in line with state-of-the-art approaches [5].
- This performance was additionally demonstrated for songs not previously heard by users, better reflecting recommendation system use cases.
- Real-time feasibility of the system was demonstrated, with 3 s of data being preprocessed and predictions generated in only ~80 ms.
- Nevertheless, there are many opportunities for future improvement:
  1. Incorporate the optimised hyperparameters.
  2. Incorporate other physiological (e.g., neural connectivity [6], cardiac and respiratory activity [7]) and acoustic (e.g., EnCodec [8]) features.
  3. Tweak the system's sensitivity by requiring multiple "skip" predictions to skip songs (and tune the prediction output rate).
- Future work will collect data from a larger cohort to better evaluate and continually develop the system.

References: [1] Tsukuda *et al.* (2021). Kiite Cafe: a web service for getting together virtually to listen to music. *Proc ISMIR*. [2] Ang *et al.* (2008). Filter bank common spatial pattern (FBCSP) in brain-computer interface. *Proc IEEE IJCNN*; [3] Pons & Serra (Pre-print). musicnn: Pre-trained convolutional neural networks for music audio tagging. *arXiv*; [4] Chawla *et al.* (2002). SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*; [5] Calcagno *et al.* (2025). EEG-Music Emotion Recognition: Challenge Overview. *Proc IEEE ICASSP*; [6] Nolte *et al.* (2004). Identifying true brain interaction from EEG data using the imaginary part of coherency. *Clin Neurophysiol*; [7] Chaturvedi *et al.* (2021). Music mood and human emotion recognition based on physiological signals: a systematic review. *Multimed Syst*; [8] Défossez *et al.* (Pre-print). High Fidelity Neural Audio Compression. *arXiv*.

Sony CSL

bccn berlin

CONTACT
t.s.binns@outlook.com
cheung@csl.sony.co.jp
GitHub: /tsbinns
LinkedIn: /tsbinns

Thomas S. Binns