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
t.s.binns@outlook.com, furuya@csl.sony.co.jp, cheung@csl.sony.co.jp

ABSTRACT

A recent focus in the development of music recommendation systems is the incorporation of physiological signals. Among this, the possibility of using non-invasive, electroencephalography-based neural activity is of great interest. In this preliminary work, we sought to predict the preference of individuals for previously unheard music through a combination of acoustic and neural features. We developed a real-time system for preference decoding which was used to skip songs with ∼80 ms latency according to users' desires. The results suggest that music recommendation systems could supplement acoustic features with neural activity for characterising an individual's music preferences in real time, with options to incorporate further acoustic and physiological information for improved system accuracy.

1. INTRODUCTION

Music recommendation systems can incorporate acoustic features [1], algorithms [2], and emotion recognition [3] for improved user experience. Among these, the possibility of identifying users' preference for songs with physiological measures has received particular interest, offering an objective measure of emotion and preference recognition [4]. Physiological correlates of emotions have been studied through various means, including cardiac activity, respiratory rate, galvanic skin response, and neural activity. Electroencephalography (EEG) is one such technique for recording neural activity non-invasively. Given its non-invasive nature and the ongoing development of portable, dry electrode systems requiring minimal setup time [5], using EEG to decode emotions and preference for music recommendation systems has received much interest. However, previous attempts to decode emotions and preference from music with EEG alone have shown minimal success [6–9] due to a limited spatial resolution, challenging signal-to-noise ratio, and perhaps a limited encoding of preference in cortical activity. Furthermore, some previous works have used songs familiar to the user [6], with familiarity confounding the neural representations of emotions and preferences, and not representing a typical use case for music recommendation in response to novel songs. Finally, the feasibility of these decoding approaches for real-time implementations have been rarely demonstrated, a critical requirement for a music recommendation system. In this proof-of-concept work, we present a multimodal system combining acoustic and oscillatory neural features for preference decoding of previously unheard songs. We demonstrate the feasibility of this approach in a real-time setting where the system controlled song presentation based on the users' predicted preferences.

2. METHODS

2.1 Music stimuli and data collection

300 anime and VOCALOID songs were randomly chosen from the Kiite platform [10] [1] and the beginning of the introduction sections manually estimated. EEG data was collected from an active 64-channel system (Brain Products) streamed at 200 Hz and referenced to the common average. Audio data was streamed as a single channel at 16,000 Hz.

2.2 Experiment paradigm

2.2.1 Training phase

Participants (n = 3) listened to the first 6 s of the introduction of 150 songs from the pool. After each song, participants rated how much they wanted to listen to the song from 1 (wanted to skip) to 9 (wanted to continue listening). Ratings were collected into 3 classes: wanted to skip (1-3); neutral (4-6); and wanted to listen (7-9).

Afterwards, the streamed EEG and audio data was aligned, and noisy EEG channels identified visually were excluded from the subsequent analyses. EEG and audio data was epoched to the 6 s of music listening, with noisy EEG epochs dropped automatically [11]. The remaining epochs were divided into 3 s segments with 1 s overlap.

[1] https://radar.kiite.jp/

Late-Breaking / Demo Session Extended Abstract, ISMIR 2025 Conference

2.2.2 *Decoding pipelines*

For the EEG data, filter bank common spatial pattern filters [12] were fit from 4-40 Hz in 4 Hz intervals. Classification used a support vector machine (SVM) with a radial basis function (RBF) kernel and scaled gamma.

Compressed audio features were extracted using musicnn [13] and decomposed using principal component analysis. Classification used an SVM with an RBF kernel and scaled gamma.

Probabilities for all training samples were generated from the base EEG and audio pipelines and used to train a logistic regression stacking classifier.

2.2.3 *Real-time decoding phase*

Participants (n = 2) listened to the first 10 s of the introduction of up to 75 songs from the remaining pool. In real time, the EEG and audio data was preprocessed, epoched into 3 s segments, and fed through the respective pipelines before a final preference prediction was generated.

Trials ended in 3 ways: 1) the system predicted the user wanted to skip the song and it was cut short; 2) the user chose to skip the song and it was cut short; or 3) the system predicted the user wanted to listen to the song, the user did not skip the song, and the full duration was played. For scenario 1, the user selected whether the system was correct in skipping the song. In this way, the precision and recall of the system could be assessed.

As a baseline performance estimate, participants (n = 1) also listened to the remaining 75 songs with the preference prediction being randomly generated. Variants of the system were used, with predictions output every 500 or 1,000 ms, and requiring 1 or 3 of the last 5 predictions to be 'skip' to end the song.

3. RESULTS

3.1 Optimised prediction accuracies

Grid searches optimised hyperparameters for the EEG, music, and stacking pipelines using a stratified 5-fold grouped cross-validation. Hyperparameters with the highest balanced accuracy averaged across all participants were chosen. On a held out test set of 15% of trials, the balanced accuracies for the 3-class problem were: EEG - 37.3 $\pm$ 0.9% (mean $\pm$ standard deviation); music - 42.8 $\pm$ 1.5%; stacking - 46.6 $\pm$ 1.6%.

3.2 Real-time system performance

The real-time decoding system was run using decoding pipelines with unoptimised hyperparameters. In these preliminary tests, the system achieved an average precision of 55.6 $\pm$ 9.1%, with a recall of 94.1 $\pm$ 8.3%. In contrast, the baseline random system achieved an accuracy of 40.0% and a recall of 44% in the one participant tested (Table 1).

Latencies from data acquisition to the final prediction (based on the preceding 3 s of data) were in total an average of 80.5 $\pm$ 7.6 ms, with 43.5 $\pm$ 3.8 ms for the EEG pipeline and 32.9 $\pm$ 4.1 ms for the music pipeline.

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

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

Table 1. Real-time decoding trial outcomes for the real system (top; n = 2) and the baseline random system (bottom; n = 1). Mean $\pm$ standard deviation shown.

4. DISCUSSION

4.1 EEG augments acoustics for preference decoding

In line with previous works, we found that acoustic features offered better decoding performance than EEG. However, combining both modalities offered the best performance, with the observed accuracies comparable with recent state-of-the-art approaches [7, 8]. As an added benefit, this performance was demonstrated for songs not previously heard by the user, which avoided introducing confounds of stimulus familiarity, better reflecting the use case of a music recommendation system.

Furthermore, the real-time feasibility of the system for controlling song presentation was demonstrated, with 3 s of acoustic and neural data being preprocessed, and predictions generated, in only $\sim$80 ms.

4.2 Opportunities for future improvement

While the system's performance is comparable to that of recent works, there are nonetheless improvements to be made. The optimised hyperparameters, and other acoustic and physiological features, could be incorporated to further improve system accuracy. Tools such as EnCodec [14] offer additional representations of acoustic information, while features such as oscillatory coupling [15] provide further readouts of neural responses to music. Furthermore, other physiological signals like cardiac activity and respiratory rate [4, 16] could be incorporated into the multimodal ensemble. Crucially, increasing the scale of the system whilst maintaining its real-time compatibility is feasible given the current low latencies of data processing.

Additionally, the performance of the system is not limited to the raw accuracies of the decoding pipelines. For instance, its sensitivity can be tweaked by requiring multiple 'skip' predictions to end a song, which can also be augmented by the rate at which predictions are output.

5. CONCLUSION

We demonstrated a real-time-compatible, multimodal approach for music preference decoding combining acoustic and neural features. Future work will collect data from a larger cohort of participants to better evaluate and continually develop the system.

6. REFERENCES

[1] D. Cheng, T. Joachims, and D. R. Turnbull, "Exploring Acoustic Similarity for Novel Music Recommendation," in *21st International Society for Music Information Retrieval Conference*, 2020, pp. 583–589.

[2] A. van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *26th Advances in Neural Information Processing Systems Conference*, 2013, pp. 1–9.

[3] S. Yousefian Jazi, M. Kaedi, and A. Fatemi, "An emotion-aware music recommender system: bridging the user's interaction and music recommendation," *Multimedia Tools and Applications*, vol. 80, no. 9, p. 13559–13574, 2021.

[4] V. Chaturvedi, A. B. Kaur, V. Varshney, A. Garg, G. S. Chhabra, and M. Kumar, "Music mood and human emotion recognition based on physiological signals: a systematic review," *Multimedia Systems*, vol. 28, no. 1, pp. 21–44, 2022.

[5] J. W. Kam, S. Griffin, A. Shen, S. Patel, H. Hinrichs, H.-J. Heinze, L. Y. Deouell, and R. T. Knight, "Systematic comparison between a wireless EEG system with dry electrodes and a wired EEG system with wet electrodes," *NeuroImage*, vol. 184, pp. 119–129, 2019.

[6] S. Calcagno, S. Carnemolla, I. Kavasidis, S. Palazzo, D. Giordano, and C. Spampinato, "EEG-Music Emotion Recognition: Challenge Overview," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–3.

[7] P. Paukner, M. Ripoll, D. Sabir, D. O. Erdoğan, L. Sacchetto, and K. Diepold, "Classifying Music-Induced Emotion Using Multi-Modal Ensembles of EEG and Audio Feature Models," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.

[8] S. Huang, Z. Jin, D. Li, J. Han, and X. Tao, "Multimodal Fusion for EEG Emotion Recognition in Music with a Multi-Task Learning Framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–2.

[9] H. Vedder, F. Stano, and M. Knierim, "Classification of Music Preferences Using EEG Data in Machine Learning Models," in *Gesellschaft für Informatik e.V. Mensch und Computer*, 2024, pp. 1–4.

[10] T. Kosetsu, I. Keisuke, H. Masahiro, and G. Mastaka, "Kiite Cafe: A web service for getting together virtually to listen to music," in *22nd International Society for Music Information Retrieval Conference*, 2021, pp. 1–8.

[11] M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo, and A. Gramfort, "Autoreject: Automated artifact rejection for MEG and EEG data," *NeuroImage*, vol. 159, pp. 417–429, 2017.

[12] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *IEEE International Joint Conference on Neural Networks*, 2008, pp. 2390–2397.

[13] J. Pons and X. Serra, "musicnn: Pre-trained convolutional neural networks for music audio tagging," *arXiv preprint arXiv:1909.06654*, 2019.

[14] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," *arXiv preprint arXiv:2210.13438*, 2022.

[15] G. Nolte, O. Bai, L. Wheaton, Z. Mari, S. Vorbach, and M. Hallett, "Identifying true brain interaction from EEG data using the imaginary part of coherency," *Clinical Neurophysiology*, vol. 115, no. 10, pp. 2292–2307, 2004.

[16] S. Sakamoto, V. K. M. Cheung, and S. Furuya, "Rapidly predicting music artistic expression preference from heart rate and respiration rate," in *Extended Abstracts for the Late-Breaking Demo Session of the 23rd International Society for Music Information Retrieval Conference*, 2022, pp. 1–3.