Predicting Musical Chills From Autoencoder Latent Space Representations

Arun Asthagiri[1], Psyche Loui[1]

[1]Northeastern University

Code here!

MIND Lab

I. Introduction

- Listening to music can induce physiological changes, which align in time with peak moments of pleasure (Salimpoor et al., 2009) and engage distinct dopaminergic responses in the mesolimbic reward system (Salimpoor et al., 2011).
- During naturalistic listening, inter-subject physiology synchronizes during musical transitions and formal boundaries (Czepiel et al., 2021) while chills have been linked to moments of expectation and surprise (Huron, Margulis, 2012).
- It is unclear how continuous interactions between low-level acoustic features (e.g. spectral and temporal changes) and high-level musical structure (e.g. musical form) contribute to perceptual and physiological responses.
- Latent spaces of neural audio encoders contain compressed information from raw audio, introducing a mechanism for extracting salient musical information from raw acoustic data.

Do latent space representations of neural audio encoders contain perceptually meaningful musical information that can predict physiological responses?

II. Methods

Setup and Materials:

- Northeastern undergraduates (n=46) listened to self-selected chill-inducing, neutral, and researcher-selected music while physiology was measured
- Wearable physiological sensors developed by NeuLog and NeuroScouting LLC were used to record skin conductance (GSR), heart rate and heart rate variability
- Pre-screening surveys assessed musical reward sensitivity (eBMRQ), absorption (AIMS), sophistication (Gold-MSI), and physical anhedonia (PAS)

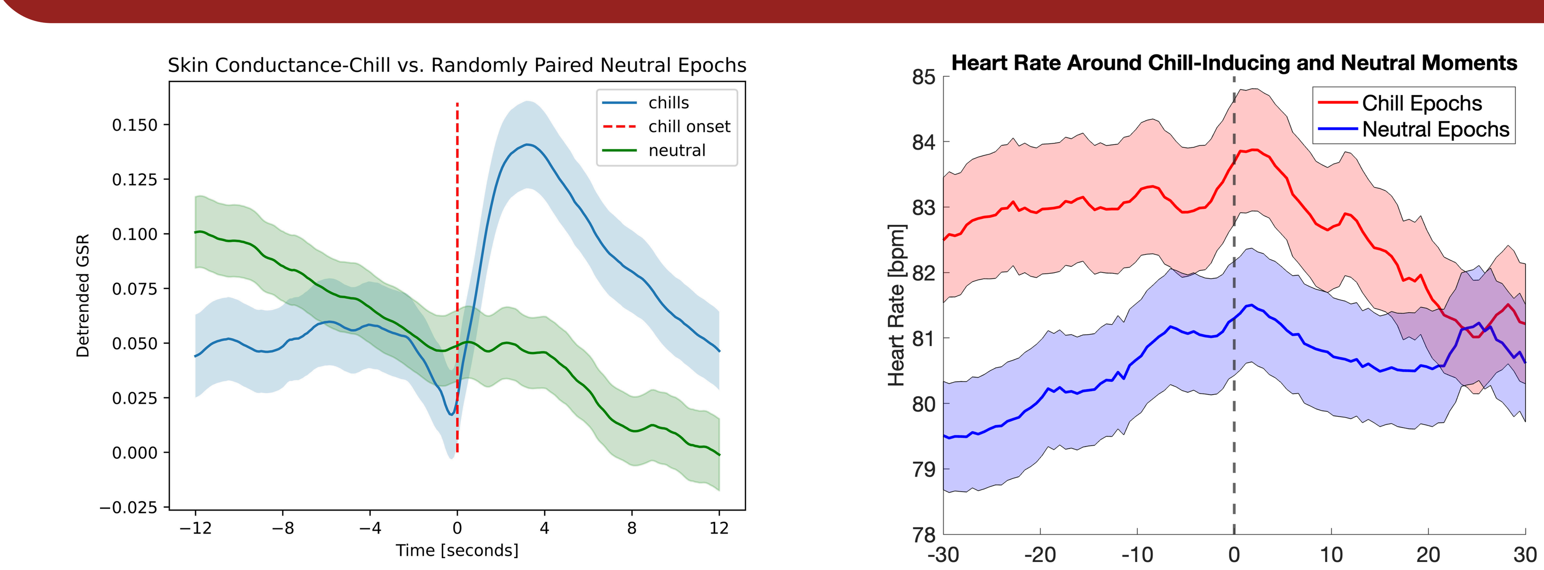
Listening:

- Participants listened to **seven 90-second excerpts**, order pseudo-randomized
  - 2 self-reported neutral songs
  - 3 self-reported chill-inducing songs
  - 2 randomized from chill-inducing stimuli from Sachs et al., 2016
- Instances of musical chills were self-reported in real-time
- Post-listening questionnaire assessed engagement, familiarity, pleasure, enjoyment, arousal, valence, thrill, and surprise on a 7-point Likert scale

Data Processing:

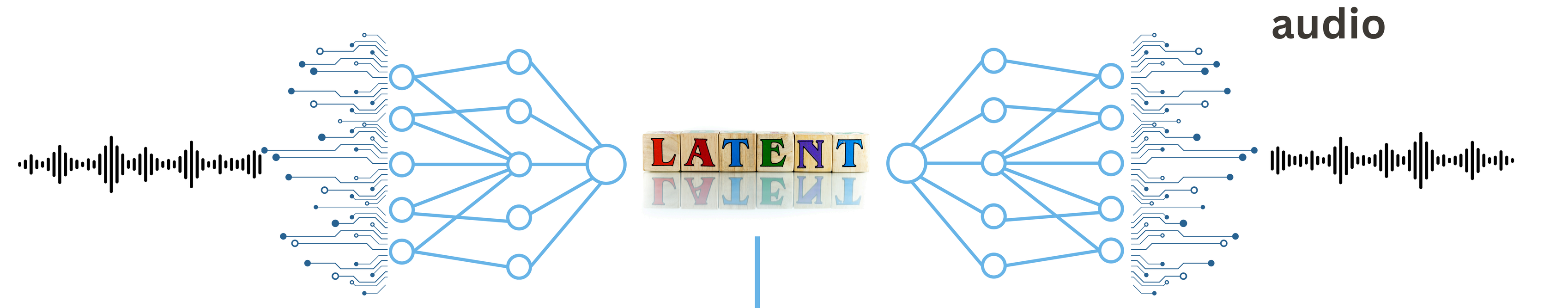
- Skin conductance was normalized by trial and linearly detrended. Heart rate was computed from a weighted average of inter-beat intervals over 10 second windows. Systolic peaks were identified using the ppg-beats Matlab toolbox.
- Data was epoched around self-reported chill-inducing moments.

III. Initial Physiological Results

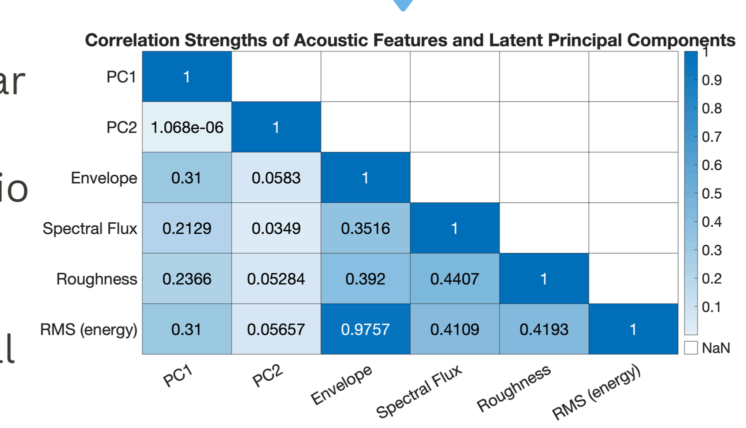


Heart rate and skin conductance increase around self-reported chill-inducing moments, demonstrating measurable physiological effects of music listening.

IV. Adapting Audio Autoencoders for MIR

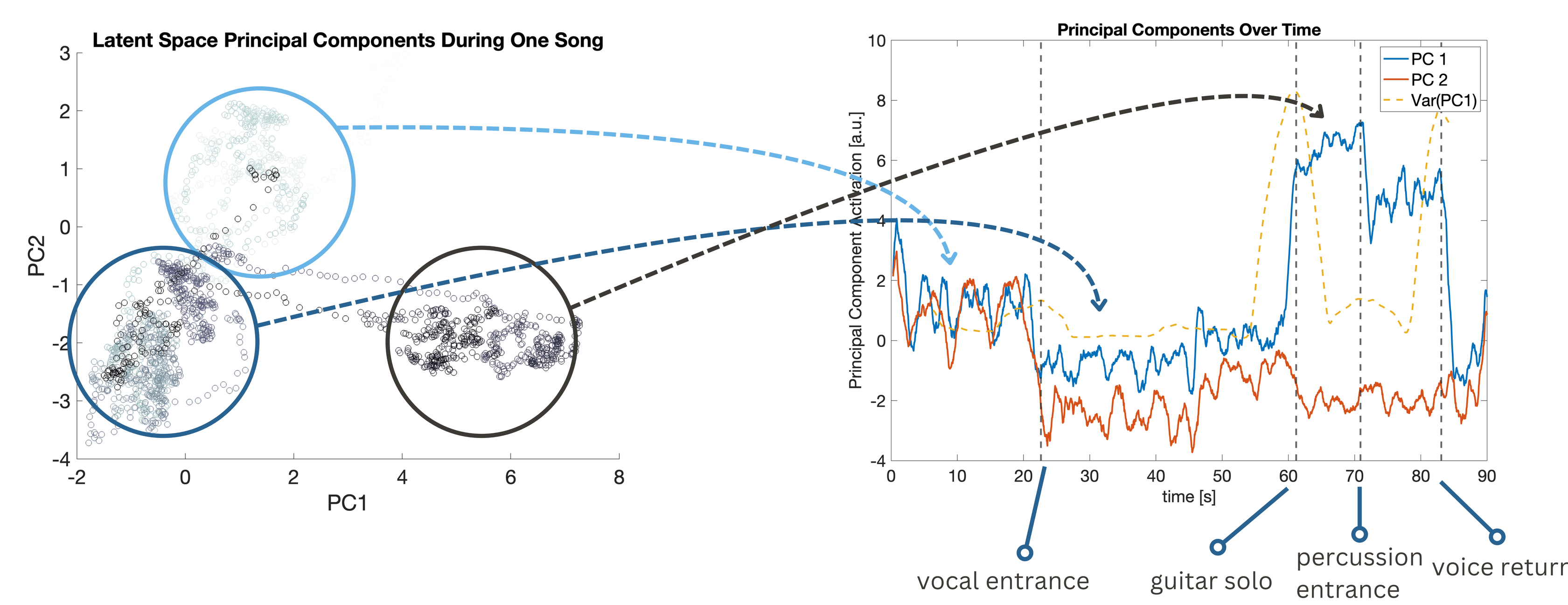


Descript's neural audio encoder (Kumar et al., 2023) compresses raw audio using a deep convolutional neural network with residual vector quantization.



The latent space encodes the most relevant acoustic information to reconstruct the original audio. The first two principal components from this space are extracted within-song and correlated with other musical features.

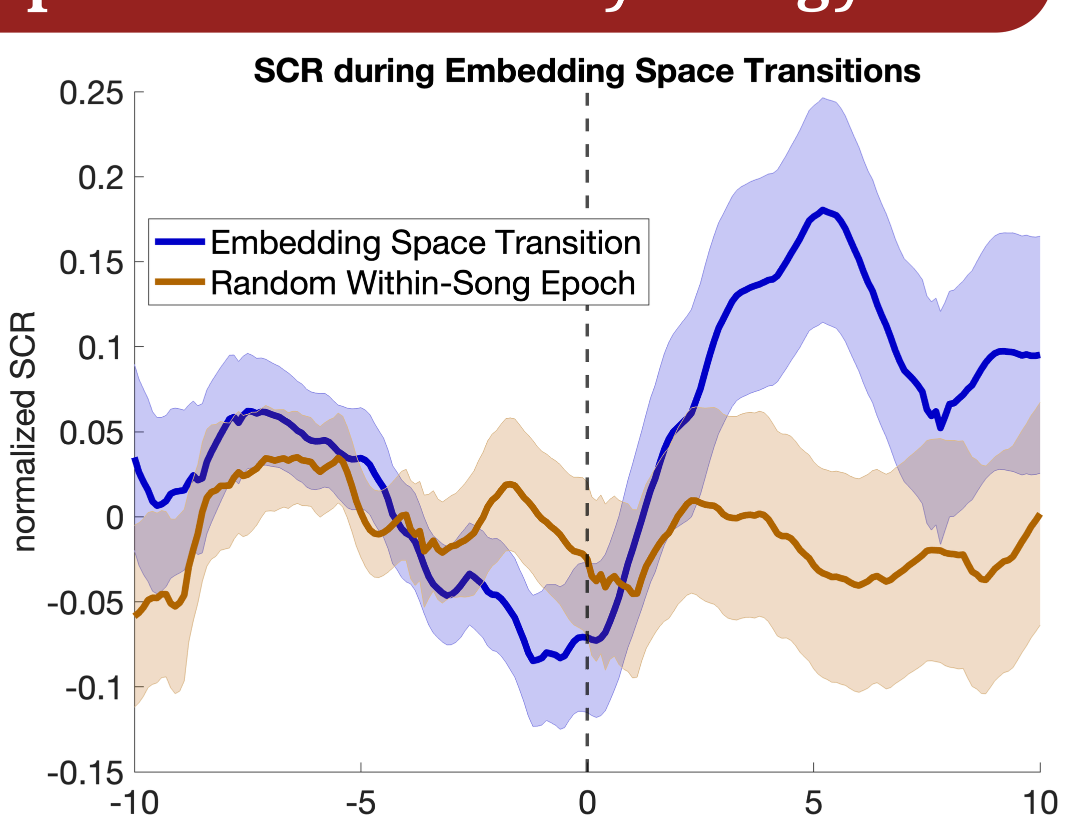
V. Structural Clustering in Latent Space



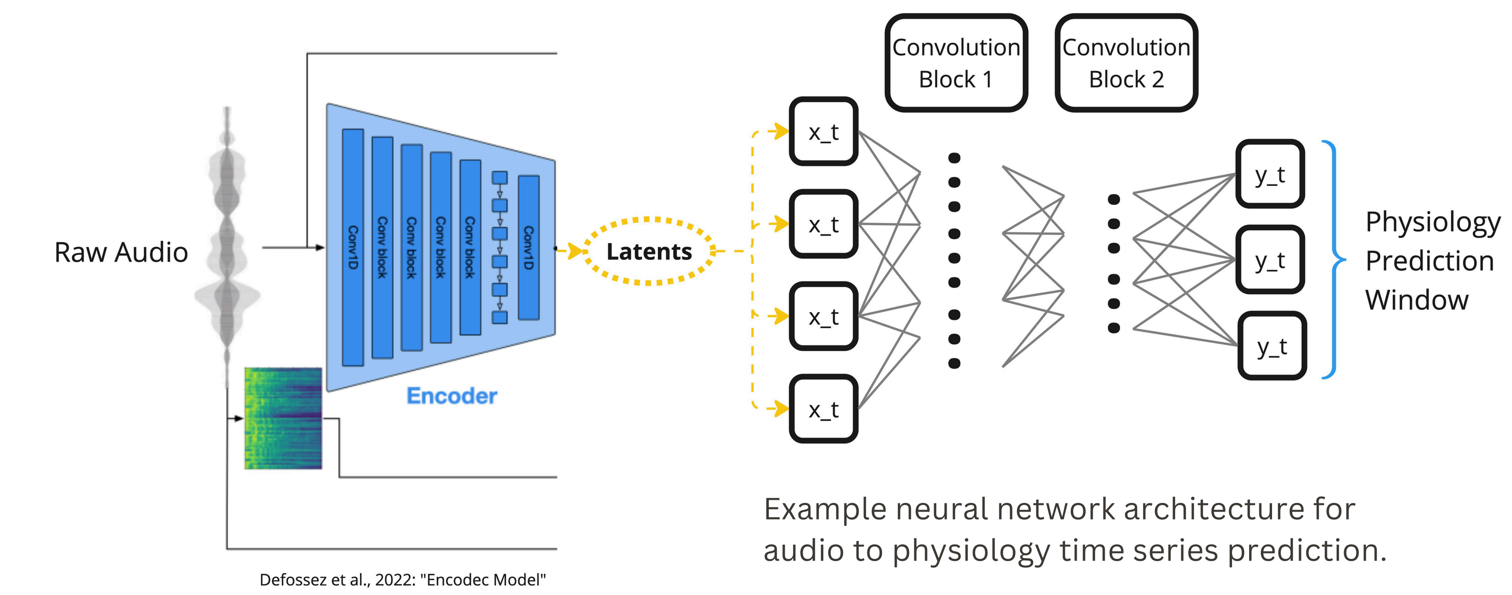
- Embeddings were averaged using a sliding window of 100 frames and hop size of 4 frames, resulting in a temporal resolution of 21.5 Hz.
- Acoustically related sections in the song appear as clusters in the first two principal components of the latent space.
- Transitions between clusters can be identified as peaks in the variance of the first principal component over time.

VI. Leaps in Latent Space Linked to Physiology

- Figure: skin conductance response epoched around transitions in the latent space (blue) versus randomly selected epochs within song (orange) with baseline subtraction (t<0).
- **A marked increase in skin conductance follows these transitional moments in the latent space**, suggesting that the latent space can be used to index moments of physiological change.



VII. Deep Learning Approach for Arousal Models



Example neural network architecture for audio to physiology time series prediction.

Audio encoders have been utilized within deep music generation models to embed and manipulate acoustic representations (Copet et al., 2024). We propose an approach for modeling continuous physiology from sequences of raw audio representations.

Model design and training:

- A 2-layer convolutional neural network was used to predict continuous changes in phasic skin conductance (filtered and normalized) from the embedding time series (12.8 sec observation window, 3.2 sec prediction window, 4 sec hop size).
- The model was trained on song data from 30 participants and evaluated on separate data from 10 participants.
- Reconstructions of the true rescaled GSR yielded a median Pearson's correlation of .397 over songs with $p < .001$ using Fisher's combined test.

Limitations:

- Lack of sensitivity to individual differences in physiological dynamics.
- Danger of overfitting and learning spurious music-physiology interactions.

VIII. Discussion

- This is the first project to our knowledge that uses **audio autoencoder latent spaces as a tool to examine musical structure** related to **intense physiological moments** during music listening (Harrison, Loui, 2014).
- The latent space of Descript's Audio Codec model correlates in time with low-level auditory features including amplitude envelope, spectral flux and roughness.
- Additionally, the latent space provides insight into high-level structural features. Cluster boundaries within the latent space reflect transitions in musical form.
  - This is a promising approach to **extracting perceptually relevant musical features** that aggregate over a distribution of low-level auditory features.
- We show that **prominent transitions in the latent space**, defined by peaks in the variance of the first principal component, **are associated with increases in skin conductance** and further demonstrate potential for capturing intense moments of pleasure with continuous predictions using deep-learning.

References

1. Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A. (2024). Simple and Controllable Music Generation. arXiv.
2. Czepiel, A., Fink, L. K., Fink, L. T., Wald-Fuhrmann, M., Tröndle, M., & Merrill, J. (2021). Synchrony in the periphery: Inter-subject correlation of physiological responses during live music concerts. Scientific Reports.
3. Défossez, A., Copet, J., Synnaeve, G., & Adi, Y. (2022). High Fidelity Neural Audio Compression. arXiv.
4. Harrison, L., & Loui, P. (2014). Thrills, chills, frissons, and skin orgasms: Toward an integrative model of transcendent psychophysiological experiences in music. Frontiers in Psychology.
5. Huron, D., & Margulis, E. H. (2010). Musical Expectancy and Thrills. In P. N. Juslin (Ed.), Handbook of Music and Emotion: Theory, Research, Applications. Oxford University Press.
6. Kumar, R., Seetharaman, P., Luebs, A., Kumar, I., & Kumar, K. (2023). High-Fidelity Audio Compression with Improved RVQGAN. arXiv.
7. Sachs, M. E., Ellis, R. J., Schlaug, G., & Loui, P. (2016). Brain connectivity reflects human aesthetic responses to music. Social Cognitive and Affective Neuroscience.
8. Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A., & Zatorre, R. J. (2011). Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. Nature Neuroscience.
9. Salimpoor, V. N., Benovoy, M., Longo, G., Cooperstock, J. R., & Zatorre, R. J. (2009). The Rewarding Aspects of Music Listening Are Related to Degree of Emotional Arousal. PLoS ONE.

Acknowledgements

Support from NeuroScouting LLC; Connie Chong, Tanvi Das, Spencer Shao, Russell Zingler