Decoding Musical Perception: Music Stimuli Reconstruction from Brain Activity

Matteo Ciferri
University of Rome, Tor Vergata
Department of Biomedicine and Prevention
matteo.ciferri@students.uniroma2.eu

Matteo Ferrante
University of Rome, Tor Vergata
Department of Biomedicine and Prevention
matteo.ferrante@uniroma2.it

Nicola Toschi
University of Rome, Tor Vergata
Department of Biomedicine and Prevention
A.A. Martinos Center for Biomedical Imaging
Harvard Medical School/MGH, Boston (US)

Abstract

This study explores the feasibility of reconstructing musical stimuli from functional MRI (fMRI) data using generative models. Specifically, we employ *MusicLDM*, a latent diffusion model capable of generating music from text descriptions, in order to decode musical stimuli from fMRI signals. We first identify music-responsive regions in the brain by correlating neural activity with representations derived from the *CLAP* (Contrastive Language-Audio Pretraining) model. We then map the fMRI data from these music-responsive regions to the latent embeddings of *MusicLDM* using regression models, without relying on empirical descriptions of the musical stimuli. To enhance between-subject consistency, we apply functional alignment techniques to align neural data across participants. Our evaluation, based on *Identification Accuracy*, achieves a high correspondence between the reconstructed embeddings and the original musical stimuli in the *MusicLDM* space, with an accuracy of 0.914 ± 0.019, surpassing previous methods. Additionally, a human evaluation experiment showed that participants were able to identify the correct decoded stimulus with an average accuracy of 84.1%, further demonstrating the perceptual similarity between the original and reconstructed music. Future work will aim to improve temporal resolution and investigate applications in music cognition.

1 Introduction

Music exerts a profound influence on the human brain, engaging distinct neural networks that modulate emotions, trigger memory recall, and affect various neurological states [Margulis et al., 2019]. These interactions underscore the importance of scientific investigation into the neural processing of music, particularly in relation to medical applications. For instance, Brain-Computer Music Interfacing (BCMI) [Miranda et al., 2011] offers the potential to tailor therapeutic music interventions to an individual's brain state, with possible implications for the treatment of neurological conditions such as depression and anxiety. Furthermore, BCMI may enable individuals with severe motor disabilities to compose or control music solely through neural activity, offering novel pathways for communication and self-expression. Also, music-based cognitive tasks could enhance cognitive functions such as mental flexibility and creativity [Olszewska et al., 2021].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

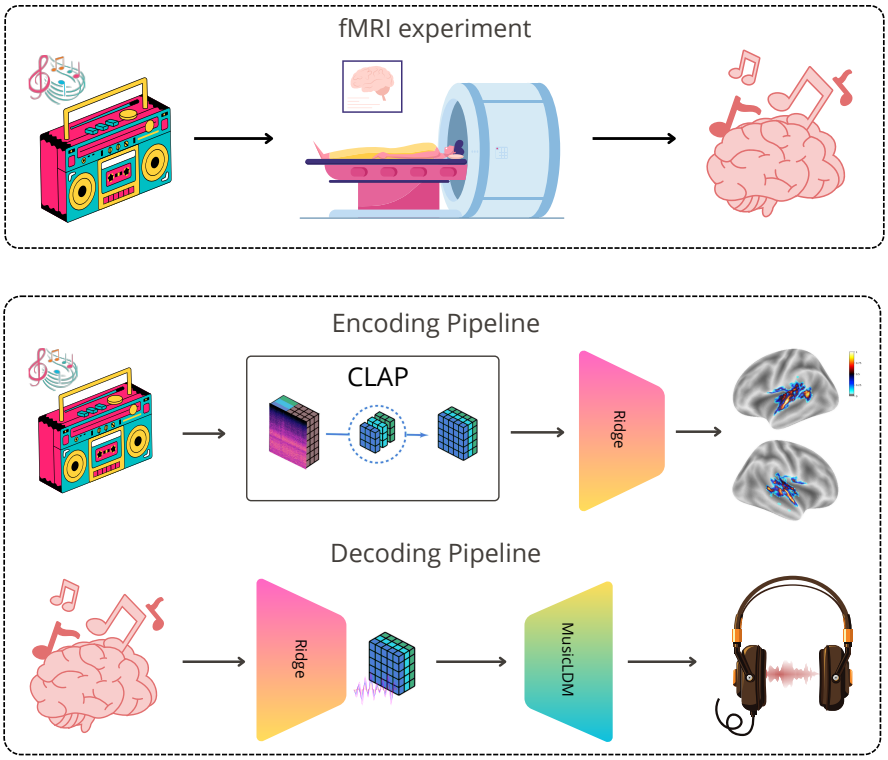

Figure 1: The proposed pipeline is composed of three stages. In the **top** section, participants listened to musical stimuli during the GTZAN-fMRI experiment, with concurrent fMRI recordings capturing their brain activity. The **middle** section involves extracting latent representations of the stimuli using the CLAP model, followed by voxel-wise encoding to correlate brain responses with music, with a correlation threshold identifying music-responsive regions. In the **bottom** section, a regression model predicts MusicLDM embeddings from these regions, which are subsequently used to reconstruct new musical outputs via the MusicLDM decoder.

This study investigates the intricate relationship between neural activity and music, focusing on the feasibility of reconstructing musical stimuli from functional MRI (fMRI) data using generative models. A key challenge is decoding high-frequency musical information (in our case within the 0-8,000 Hz range) given the lower temporal resolution of fMRI data, which is further complicated by regional variations in the brain's Haemodynamic Response Function (HRF). [Denk et al., 2023] similarly tackles generative music decoding using the same fMRI dataset. However, unlike our approach, it employs subject-specific pipelines based on anatomical atlases and proprietary models such as MuLan and MusicLM [Agostinelli et al., 2023, Huang et al., 2022]. Figure 1 provides an overview of our methodology.

2 Related Work

The neural basis of music processing has been explored extensively in classical neuroscience [Raglio et al., 2019]. However, recent advancements in artificial intelligence have enabled more detailed and data-driven analyses of brain responses to musical stimuli [Oota et al., 2023]. Building upon prior research, significant progress has been made in mapping fMRI activity to latent representations of diverse stimuli, including images, video, language, and music, through techniques such as linear mappings and subject-specific models [Ferrante et al., 2023, Scotti et al., 2023, Chen et al., 2023b, Denk et al., 2023]. Bellier et al. [2023] demonstrate that music reconstruction can be performed using both linear and nonlinear approaches to decode the auditory experience using EEG data. The advent of pre-trained models has further facilitated the extraction of latent representations capable of driving retrieval tasks or conditioning generative models. Notably, the development of text-to-music models has made it possible to generate high-fidelity music in a conditional framework, linking

language-based representations with the generation of coherent musical outputs [Agostinelli et al., 2023, Lam et al., 2023, Copet et al., 2024].

Our work extends the approaches of Ferrante et al. [2024] and Denk et al. [2023], advancing previous methods by decoding cross-subject brain activity within a generative framework that operates independently of empirically derived captions.

3 Material and Methods

This section details the dataset and methods employed in this study. The dataset is publicly accessible at `https://openneuro.org/datasets/ds003720/versions/1.0.1`. Code is available at this repository: `https://github.com/neoayanami/fmri-music-gen`.

We used the GTZAN fMRI dataset [Nakai et al., 2023], which consists of fMRI data from five subjects (sub-001 to sub-005) exposed to music stimuli drawn from ten distinct genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Each genre was represented by 54 tracks (i.e. stimuli) sampled at 22.050 kHz. Subjects underwent 18 fMRI runs, of which 12 were used for training and 6 for testing. Each run consisted of 40 music clips, each 15 seconds in duration. The stimuli were RMS-normalized and included a 2-second fade-in and fade-out. During testing, each stimulus was presented four times, and the brain activity (i.e. fMRi signal) was averaged across identical stimuli to enhance the signal-to-noise ratio.

Scanning was performed using a 3.0T MRI scanner with a repetition time (TR) of 1,500 ms, yielding 400 volumes per run. Preprocessing included motion correction, co-registration to Montreal Neurological Institute (MNI) space using T1-weighted anatomical images, detrending, and run-level standardization. Brain activity was time-shifted by 3 TRs (4.5 s) to account for the delayed hemodynamic response, and neural representations were averaged over 10 volumes (15 seconds). The final dataset comprised 540 fMRI-stimulus pairs per subject (480 for training and 60 for testing).

3.1 Encoding Model

Music processing in the brain involves complex, non-linear mechanisms. To capture this, we used the CLAP (Contrastive Language-Audio Pretraining) model [Elizalde et al., 2022], a multimodal neural network employing contrastive learning for audio and text alignment. CLAP extracts audio features using the SWINTransformer [Liu et al., 2021] and log-Mel spectrograms, as well as text features using RoBERTa [Liu et al., 2019], projecting both into a shared latent space. Cosine similarity is commonly used to measure the correspondence between elements of this shared latent space. Further, to identify brain regions most responsive to musical stimuli, we applied a voxel-wise encoding model, mapping CLAP's audio embeddings to fMRI data using Ridge regression with cross-validation (further details in Appendix A.1). Model training incorporated a hyperparameter search for the regularization parameter α (ranging on a logarithmic scale from 10^{-2} to 10^{3}) and we empirically determined a correlation threshold (in a discrete range of values $[0.01, 0.02, 0.05, 0.08, 0.10, 0.15, 0.20]$) to define music-responsive brain regions by choosing the value which maximized identification accuracy (see "Evaluation"). Regression models were trained per voxel, and Pearson correlation coefficients were used to create a voxel-wise correlation map between real and predicted fMRI activity, identifying regions most responsive to musical stimuli.

3.2 Functional Alignment

To mitigate individual variability in brain structure and function, we employed cross-subject data aggregation techniques following Ferrante et al. [2024]. Anatomical alignment is widely used in neuroimaging as it facilitates the direct comparison of localized brain activity across subjects. However, relying solely on the brain's physical structure for alignment and decoding lacks the precision needed for fine-grained tasks due to inherent subject-specific anatomical variability, which may not exactly mirror functional differences. We adopted a Ridge regression framework with cross-validation to regularize and merge voxel-wise fMRI data across subjects, potentially improving model generalization abilities. In detail, the input was brain activity from the subject to be aligned, and the output was the activity aligned to a template subject's space (the target subject was sub-001). 5-fold cross-validation was used, where each fold predicted aligned brain activity using held-out data.

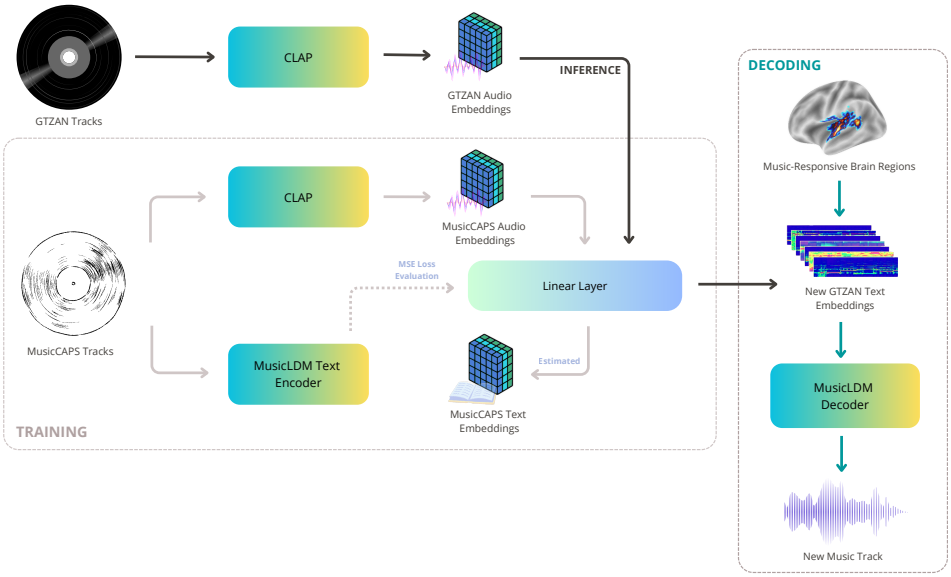

Figure 2: Overview of the music generation framework from fMRI. (1) GTZAN music stimuli processed by the CLAP model to extract audio embeddings, used for inference. (2) MusicCaps stimuli processed by CLAP for audio embeddings and MusicLDM for text representations. (3) A linear layer maps these embeddings into a 512-dimensional latent space. (4) Music-responsive brain regions are used to estimate new GTZAN embeddings via regression. (5) The predicted embeddings are passed to the MusicLDM decoder for music generation.

3.3 Decoding Model

To decode music from brain activity, we used the Music Latent Diffusion Model (MusicLDM) [Chen et al., 2023a], a generative model conditioned on text. MusicLDM integrates CLAP for audio-text contrastive learning, a latent diffusion model for audio generation, and HiFi-GAN [Kong et al., 2020] for audio reconstruction. Since MusicLDM relies on text conditioning, we aligned CLAP's audio embeddings with textual embeddings from the MusicCaps dataset [Agostinelli et al., 2023]. To this end, a linear layer was trained to minimize the mean squared error (MSE) between the two representations, resulting in 512-dimensional embeddings. During inference, we predicted embeddings from the GTZAN dataset and mapped brain activity to MusicLDM's latent space using Ridge regression with 5-fold cross-validation. This established a direct mapping from neural activity to musical representations. The full pipeline is illustrated in Figure 2.

3.4 Evaluation

We evaluated model performance using the identification accuracy metric defined in the Brain2Music framework [Denk et al., 2023], which quantifies the correspondence between predicted and target music embeddings using Pearson correlation. A correct identification occurs when the correlation between a predicted and true embedding is higher than with any other target embedding. We computed a correlation matrix C, where $C_{i,j}$ represents the Pearson correlation between the i-th predicted embedding and the j-th target embedding. The identification accuracy was computed as:

$$id\_acc_i = \frac{1}{n-1} \sum_{j=1}^{n} 1\left[C_{i,i} > C_{i,j}\right]$$

where $1[\cdot]$ is the indicator function. The overall accuracy was averaged across all predictions. This metric ensures the robustness of the model in discriminating between embeddings, which is crucial for applications requiring high precision.

As a qualitative evaluation, we developed a human metric to assess the perceived similarity between reconstructed and original music stimuli. 10 participants listened to pairs of stimuli from the test set and were asked to identify which of the two stimuli was the correct decoded version (additional

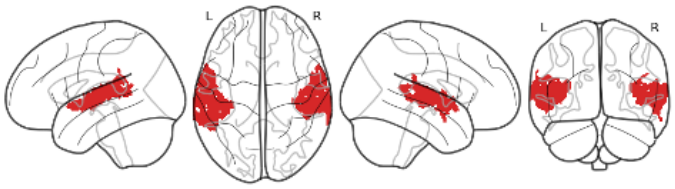

Figure 3: Brain regions corresponding to music-responsive areas, identified using a correlation-based threshold to differentiate predicted from real fMRI activity.

information is provided in Appendix A.3). Subsequently, we calculated the percentage of correct identifications and averaged across subjects.

4 Results

A Pearson correlation threshold of 0.05 was selected based on its optimal performance in terms of Identification Accuracy. This threshold identified 3,433 voxels involved in music processing. The spatial distribution of these voxels, primarily located in lateral and temporal brain regions, is depicted in Figure 3. The identified brain regions in the temporal lobes and lateral frontal areas—including the bilateral superior temporal gyri and inferior frontal gyri—are plausible music-responsive regions involved in auditory processing and complex aspects of music cognition, aligning with established research linking music perception to both auditory and higher-order cognitive regions.

Table 1 presents the performance of the proposed method, which utilizes functional alignment. The model achieved an Identification Accuracy of 0.914 ± 0.019, outperforming several baseline methods.

Table 1: Comparison of Test Identification Accuracy

Embedding	**Test Identification Accuracy**
SoundStream-avg	0.674 ± 0.016
w2v-BERT-avg	0.837 ± 0.005
$\text{MuLan}_{\text{text}}$	0.817 ± 0.014
$\text{MuLan}_{\text{music}}$	0.876 ± 0.015
Our - Functional Alignment	$\mathbf{0.914 \pm 0.019}$

In the human metric experiment, the participants' average accuracy was 84.1% with respect to the chance level (50%), indicating that the human ear is able to detect the correlation between the original musical stimulus and the brain-generated counterpart. Samples of the generated tracks compared to the original stimuli are available at: `https://musicdecod.my.canva.site/decoding-musical-perception`.

5 Discussion and Conclusion

This study demonstrates the feasibility of decoding music from neural activity across multiple subjects with high accuracy, utilizing advanced computational methods and neural alignment techniques. These results contribute to a deeper understanding of cognitive music processing and have implications for potential applications, such as therapeutic interventions and brain-computer interfaces. Our findings are consistent with previous research, indicating that genres with distinct structural features, such as classical and jazz, are more robustly represented in the brain. In contrast, closely related genres like rock and blues are more challenging to differentiate, suggesting the need for more refined modelling approaches. The inherent noise in fMRI signals, along with their subsampled nature, limits the precision and fidelity of the reconstructed music. The coarse temporal resolution of fMRI (1.5 seconds) constrains the ability to decode rhythmic components accurately, and the extended scanning duration required may reduce the practicality of these methods in real-world applications. In the future, comparative studies of brain-reconstructed music among participants with different levels of musical expertise or from diverse cultural backgrounds could yield valuable insights.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text, 2023.

L. Bellier, A. Llorens, D. Marciano, A. Gunduz, G. Schalk, P. Brunner, et al. Music can be reconstructed from human auditory cortex activity using nonlinear decoding models. *PLoS Biology*, 21(8):e3002176, 2023. doi: 10.1371/journal.pbio.3002176. URL `https://doi.org/10.1371/journal.pbio.3002176`.

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies, 2023a.

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity, 2023b.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation, 2024.

Timo I. Denk, Yu Takagi, Takuya Matsuyama, Andrea Agostinelli, Tomoya Nakai, Christian Frank, and Shinji Nishimoto. Brain2music: Reconstructing music from human brain activity, 2023.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision, 2022.

Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, and Nicola Toschi. Multimodal decoding of human brain activity into images and text. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023. URL `https://openreview.net/forum?id=rGCabZfV3d`.

Matteo Ferrante, Matteo Ciferri, and Nicola Toschi. Rb – rhythm and brain: Cross-subject decoding of music from human brain activity, 2024.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. Mulan: A joint embedding of music audio and natural language, 2022.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, 2020.

Max W. Y. Lam, Qiao Tian, Tang Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, Jitong Chen, Yuping Wang, and Yuxuan Wang. Efficient neural music generation, 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

Elizabeth Hellmuth Margulis, Patrick C. M. Wong, Rhimmon Simchy-Gross, and J. Devin McAuley. What the music said: narrative listening across cultures. *Palgrave Communications*, 5(1):146, Nov 2019. ISSN 2055-1045. doi: 10.1057/s41599-019-0363-1. URL `https://doi.org/10.1057/s41599-019-0363-1`.

Eduardo Miranda, NMT-F Wilson, Ramaswamy Palaniappan, Joel Eaton, and Wendy Magee. Brain-computer music interfacing (bcmi) from basic research to the real world of special needs. *Music and Medicine*, 3:134–140, 07 2011. doi: 10.1177/1943862111399290.

Tomoya Nakai, Naoko Koide-Majima, and Shinji Nishimoto. "music genre fmri dataset", 2023.

Alicja M. Olszewska, Maciej Gaca, Aleksandra M. Herman, Katarzyna Jednoróg, and Artur Marchewka. How musical training shapes the adult brain: Predispositions and neuroplasticity. *Frontiers in Neuroscience*, 15, 2021. ISSN 1662-453X. doi: 10.3389/fnins.2021.630829. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2021.630829.

Subba Reddy Oota, Manish Gupta, Raju S. Bapi, Gael Jobard, Frederic Alexandre, and Xavier Hinaut. Deep neural networks and brain alignment: Brain encoding and decoding (survey), 2023.

Alfredo Raglio, Enrico Oddone, Lara Morotti, Yasmin Khreiwesh, Chiara Zuddas, Jessica Brusinelli, Chiara Imbriani, and Marcello Imbriani. Music in the workplace: A narrative literature review of intervention studies. *Journal of Complementary & Integrative Medicine*, pages /j/jcim.ahead–of–print/jcim–2017–0046/jcim–2017–0046.xml, October 2019. ISSN 1553-3840. doi: 10.1515/jcim-2017-0046.

Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors, 2023.

A Supplemental Material

A.1 Encoding Model Details

Formally, for each subject i, we estimate whole-brain encoding weights β to predict brain activity z using the audio latent representations h as inputs. Specifically, we train a model $\hat{z}_i = h\beta_i$ for each subject. To optimize the weights β, we perform nested cross-validation on the training set, minimizing the loss function $\mathcal{L} = |z_i^{tr} - h^{tr}\beta_i|^2 + \alpha|\beta_i|^2$. In each fold, we predict the held-out data (20% of the training set not used to train that specific model). When predicting on the entire training set, we compute the voxelwise correlation between the predicted and real brain activity, $corr(\hat{z_i^{tr}}, z_i^{tr})$, selecting only voxels that exceed the correlation threshold for further analysis.

A.2 Similarity in Music Latent Diffusion Space

We also computed the cosine similarity matrix between the real and predicted MusicLDM embeddings using the following formula: $cosine\_sim_{i,j} = \frac{\mathbf{r}_i \cdot \mathbf{p}_j}{\|\mathbf{r}_i\| \|\mathbf{p}_j\|}$ where $\mathbf{r}_i$ and $\mathbf{p}_j$ are the normalized real and predicted embeddings, respectively. Each score represents the degree of alignment between these embeddings, which is critical for evaluating the model's performance in producing embeddings consistent with the target data. The computed similarity matrix is visualized as a heatmap (Figure 4), where each cell represents the cosine similarity between a real target embedding (row) and a predicted embedding (column). The matrix illustrates that the predicted latent representations of the stimuli are well-aligned with the real ones (along the diagonal) and, in some cases, also between representations of closely related genres, such as rock, reggae, and blues.

A.3 Human Evaluation

We used *Streamlit*, a Python framework, to develop a simple web interface for participants. *Streamlit* is an open-source library for turning data scripts into web applications without requiring the implementation of a custom front end. The user experience consisted of listening to three distinct audio stimuli: the original stimulus, the brain-decoded version of the target stimulus, and a randomly selected brain-decoded track from the test set. In each trial, the randomly selected song was chosen without regard to genre, which permitted participants to encounter pairs from closely related genres (e.g. 'hip-hop' vs. 'pop' or even 'hip-hop' vs. 'hip-hop') as well as contrasting genres (e.g. 'hip-hop' vs. 'jazz'). Participants were asked to choose the track they perceived as most similar to the original stimulus before proceeding to the next musical stimulus. The application interface is shown in Figure 5.

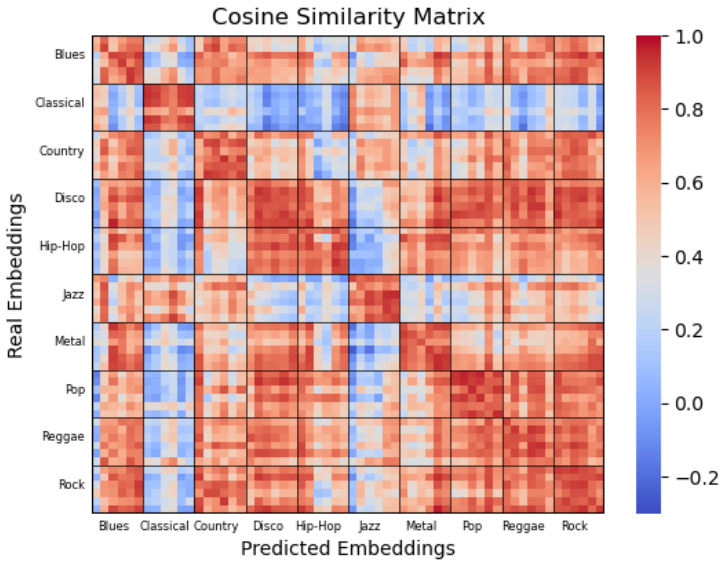

Figure 4: Heatmap of cosine similarity scores between real (rows) and predicted (columns) GTZAN embeddings. The colour scale ranges from -0.2 (blue) to 1.0 (red), with red indicating higher similarity. The concentration of red along the diagonal indicates a strong match between predicted embeddings and their corresponding real embeddings.

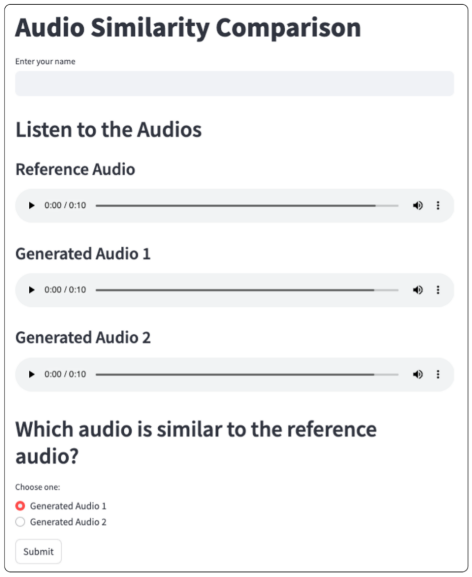

Figure 5: Reprentative image of the *Streamlit* interface used by participants during the human metric experiment.