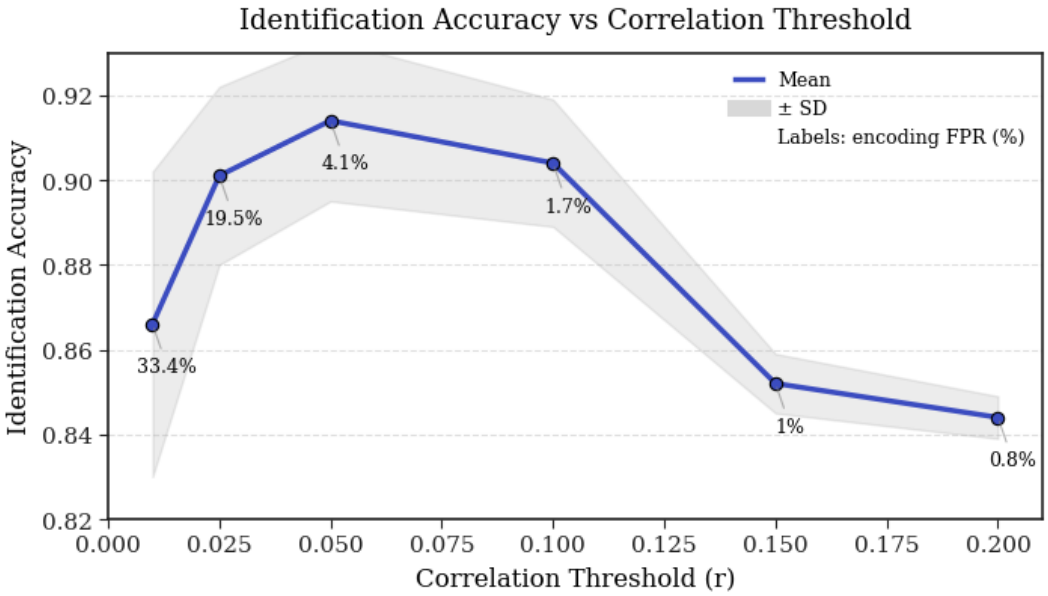

Figure 5. Identification accuracy as a function of the voxel-selection correlation threshold r (mean $\pm$ SD across subjects, shaded band). Text labels indicate the estimated encoding FPR (%) obtained from a permutation-based null (100 shuffles of the audio–fMRI correspondence). The choice r=0.05 balances the accuracy plateau with a low expected false-positive rate.

A Supplementary Materials

A.1 Voxel Selection Procedure

To select informative voxels we computed voxel-wise Pearson correlations between the predicted stimulus features and the corresponding fMRI on the training set. We then thresholded these correlations at candidate values $[0.01, 0.02, 0.05, 0.10, 0.15, 0.20]$.

To estimate the expected false-positive rate under the null hypothesis, we built an empirical null distribution by permuting the audio–fMRI correspondence (100 permutations). For each permutation we re-fit the same encoding model (RidgeCV with the identical set of regularization values) on the shuffled inputs and recomputed voxel-wise correlations. For a given threshold $\hat{r}$, the proportion of null voxels provides an estimate of the false positive rate for that cutoff.

We summarized (i) test identification accuracy as a function of r (mean $\pm$ SD) and (ii) the estimated false positive rate at the same r values (labels in Fig. 5). We selected r=0.05 as a trade-off point (accuracy is the peak, estimated FPR $\approx$4%), and used this fixed threshold for all subjects and subsequent analyses.

A.2 Generative Prior Details

For each fMRI-derived latent embedding, we generated 100 candidate audio samples using the MusicLDM diffusion pipeline. Different random seeds were employed to sample diverse realizations from the generative prior while preserving the same conditioning embedding. All generation parameters were kept constant across experiments to ensure reproducibility and comparability. Specifically, we used a guidance scale of 2.0 to apply moderate classifier-free guidance, balancing sample fidelity and diversity, and performed 100 denoising inference steps. A negative prompt (*"Low quality"*) was applied to steer the diffusion process away from degraded or artifact-prone outputs.

A.3 Human Listener Study

Ten adult participants (6 males, 4 females; age range 25–35 years) took part in the evaluation. Half of them were active researchers, and two were trained musicians with experience in audio synthesis. All participants reported normal hearing and gave informed consent. Listening took place individually in a quiet environment using high-fidelity headphones. Each trial presented three 15-s audio clips: (i) the original musical stimulus, (ii) the brain-decoded reconstruction corresponding to that stimulus, and (iii) a randomly selected brain-decoded track. The order of presentation was randomized. Participants were asked to choose which of the two decoded clips (ii) or (iii) sounded more similar to the original (i). The application interface is shown in Figure 6.

The human-listener study did not involve any clinical intervention or the collection of sensitive data. Participants were healthy adults and voluntarily took part in the experiment, which consisted solely of listening to musical excerpts and selecting

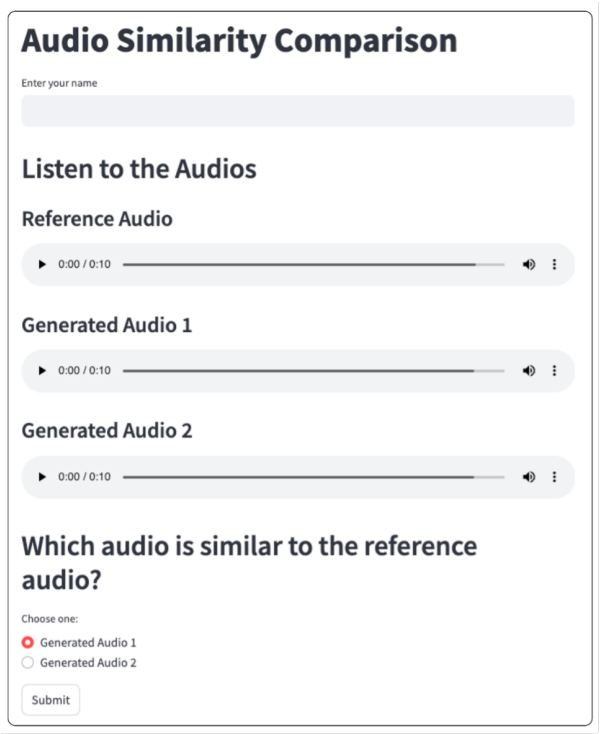

Figure 6. Representative image of the *Streamlit* interface used by participants during the human metric experiment.

which reconstruction best matched the original. The procedure posed no risk to participants and therefore did not require formal ethics committee approval. All participants gave their verbal agreement before taking part in the task.