Echoes of Humanity: Exploring the Perceived Humanness of AI Music

Flavio Figueiredo G. Martinelli H. Sousa P. Rodrigues F. Pedrosa L. N. Ferreira

NEURAL INFORMATION PROCESSING SYSTEMS

UFMG

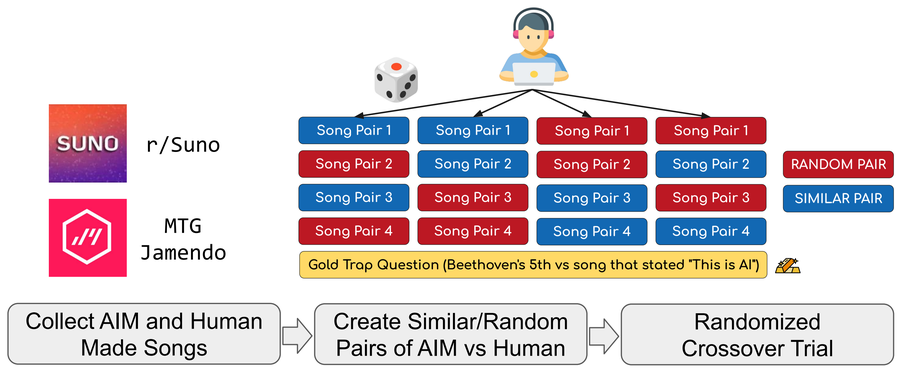
Can you distinguish AI music from Human-made music? What cues do you employ?

Motivating Hypothesis

Human listeners rely on contextually grounded cues (e.g., as repetitive structure or synthetic-sounding vocals) that help discern whether a piece of music is AIM or human-made.

- We present a Turing-test like study to understand how humans perceive AI-music (AIM) and Human-made songs;
- Different from other studies, we do focus on personal tastes (e.g., liking a song over the author);
- Our main focus is on how listeners perceive;
- More importantly, we employ a novel *in-the-wild* dataset of AIM songs (not controlled nor generated by authors).
- We employ a Randomized Crossover Controlled Trial (RCCT) that allows for **causal interpretation.**

The Humanness Perception Study



https://hpc.uai.science/

Try it !

- Similar pairs were had the same genre and belonged to the top quartile (25%) similarity (CLAP embedding cosine).
- No participant evaluated the same song more than once. Song order was randomized within every pair.

Dataset

1. **AIM Songs (Reddit):** – Novel dataset of real *in-the-wild* AIM songs
   - Crawled from Reddit r/Suno (July 2023–2025);
   - Genre classified with Essentia's open source model (same labels as Human songs below).
2. **Human Made Songs (Jamendo):** – Real recordings from real artists (non-commercial)
   - Predates the rise of prompt based AI Music
   - Comes from the public MTG-Jamendo dataset; Contains genre-data.

Participant Pool

1. Volunteer Pool
   - Seeded from social-media of the Music and Computer Science Departments;
   - Also featured on local news outlets;
   - Portuguese speaking (mostly).
2. Crowd-worker Pool (Prolific)
   - Focused on English speaking countries;
   - 100 participants hired in total;
   - Payed 2 GBP (this value was based on the median study time of the volunteer pool and Prolific's recommended wage).

In total 653 participants logged on to our study. However, we only analyzed results for those that (1) did not know any song; (2) got the gold-sanity trap question right. We were left with 337 participants.

When Participants Differentiate (Quantitative)

For random pairs, participants are tied to random chance (50%) when inferring AIM songs.

For similar song pairs, participants are better than random chance ($obs = 66\%^{***}$).

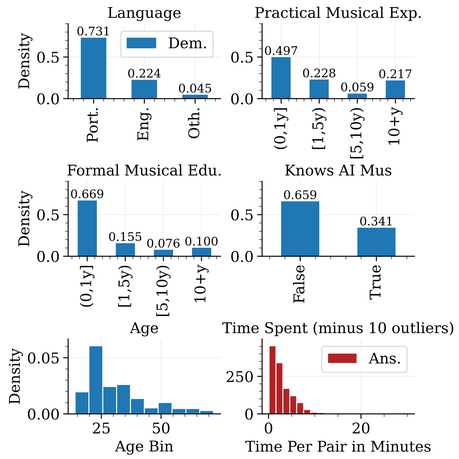


Figure 1. Demographic and Answer Variables

	Estimate	Pr(>|z|)	Sig.
Intercept	-24.26	0.9968	
↑ Similar Pair	0.61	0.0999	*
Choice: Song A or B	22.29	0.9971	
Choice: Both Songs	0.82	0.9999	
Choice: Neither Song	4.94	0.9994	
↑ $\log_{10}$(TimeSpent+1)	0.49	0.0611	*
Lang. Port.	-0.07	0.7621	
Prac. Exp. 1 to 5 y	0.42	0.1157	
↑ Prac. Exp. 5 to 10 y	0.92	0.0995	*
↑ Prac. Exp. Over 10 y	1.25	0.0009	***
Formal Edu. 1 to 5 y	-0.22	0.4803	
↓ Formal Edu. 5 to 10 y	-1.30	0.0086	***
Formal Edu. Over 10 y	-0.82	0.0614	*
↑ Knowledge on AIM	0.89	0.00005	***
↓ Participants' Age	-0.03	0.0009	***

Table 1. Covariates $^{*}p < .1$, $^{**}< .05$, $^{***}< .01$

How Participants Differentiate (Quantitative and Qualitative – Mixed Methods)

Methodology: Ground Theory Based Coding of Participants Free-Text Feedback.

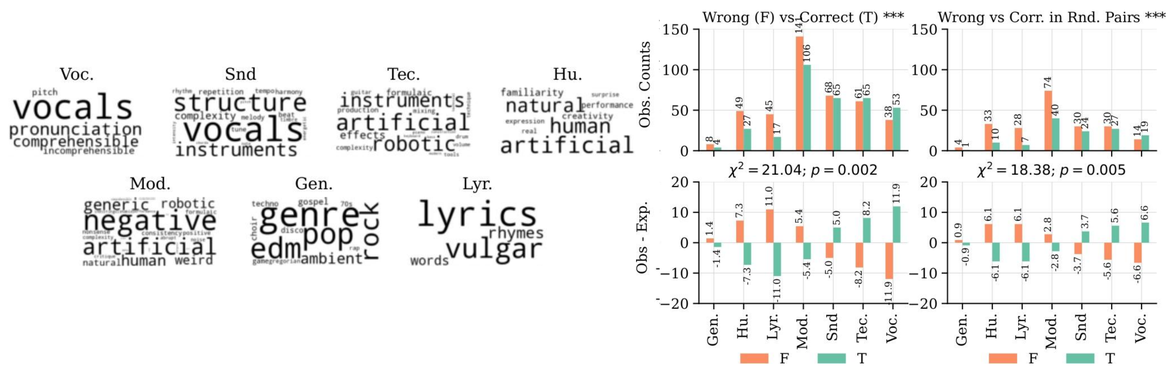


Figure 2. Topics and Classes and the Observed Topic Frequencies vs. Differences Towards the Expected. $^{***}p < .01$

Key result: *Sound, technical, and vocal* cues, help discern whether a piece of music is AIM (see plots on right).

Conclusions

- We characterized both *when* listeners can detect AIM and *how* they do it;
- Our results may be used to improve models. Particularly, the **vocal** style of AIM songs is a significant giveaway;
- Future work extends beyond WEIRD-domain and other media.