TOWARD CI-ACCESSIBLE MUSIC AI: EMOTIONAL PERCEPTION AND REPRESENTATION OF AI-GENERATED MUSIC FOR COCHLEAR IMPLANT USERS

Anonymous Authors
Anonymous Affiliation
anonymous@ksmi.kr

ABSTRACT

The perception of diverse emotions in music, driven by multiple acoustic cues, plays a central role in human life. However, for Deaf and Hard-of-Hearing (DHH) populations, including cochlear implant (CI) users, degraded acoustic cues can limit access to musical emotion. Motivated by recent advances in generative AI, we investigate its potential to improve music accessibility for these populations. We conducted a behavioral experiment in which CI and normal-hearing (NH) groups rated perceived emotions of human-composed and AI-generated music. Moreover, under the same stimuli (human vs. AI) and listening conditions (NH vs. CI), we analyzed foundation model representations to examine how musical emotions are encoded and align with human responses. Behaviorally, emotion categories were less separable for AI-generated music than for human-composed music. A similar reduction was observed in the CI group compared to NH group, resulting in the lowest separability in the CI–AI condition. Foundation models showed reduced prediction accuracy for ratings in the CI group, and exhibited similar patterns of emotion separability as observed in the behavioral experiment. These results suggest that current AI-generated music fails to fully capture emotional structure, particularly under degraded listening conditions, highlighting the need for improved accessibility for diverse auditory populations.

1. INTRODUCTION

The perception of diverse emotions in music, driven by multiple acoustic cues, plays a central role in human life, supporting emotional regulation, well-being, and social communication [1–5]. It arises from the perception of acoustic cues such as pitch, timbre, and rhythm, making their perception essential for experiencing emotion in music [6–9]. However, this process is significantly challenged in Deaf and Hard-of-Hearing (DHH) populations, such as cochlear implant (CI) users. Due to the limited spectral resolution of CI devices, many acoustic cues are degraded or lost [10, 11], resulting in music that is perceived as distorted and different from prior experiences. As a consequence, post-lingually deafened CI users often report reduced musical experience and a diminished quality of life, while still expressing a strong desire to re-engage with music [11–14]. To address this gap in music accessibility for DHH populations, we turn to recent advances in generative AI, which enable the creation of music conditioned on user-defined attributes, offering new possibilities for more personalized musical experiences. [15–17]. However, it remains unclear how AI-generated music is perceived by CI users and whether its emotional structure is preserved under degraded listening conditions. In this context, we investigate how musical emotion is perceived and represented across both human listeners and AI systems. Specifically, we compare emotional perception across listener groups (normal-hearing (NH) vs. CI) and sources (human vs. AI), and further examine whether similar patterns emerge in music foundation models.

2. METHODS

2.1 Stimuli

To investigate a diverse range of emotions perceived in music, we selected seven emotion categories (*calm, cheerful, dreamy, energetic, in love, sad, tense*) based on the taxonomy proposed by H. Lee et al. [18]. For the human-composed music, excerpts were drawn from the dataset introduced by Cowen et al. [19], and a pilot study was used to select 35 excerpts (5 per emotion category) that clearly expressed the target emotions and had low familiarity. For the AI-generated music, 35 excerpts were generated using Suno v5. Each excerpt was created using a simple text prompt of the form "{emotion category} music," with the instrumental-only option enabled, and a 5-second segment was randomly selected from each piece. For the foundation model inputs, we used both original and CI-simulated audio. The CI-simulated audio was processed using an 8-band noise vocoder to approximate CI hearing in NH listeners, following a CI-simulation algorithm [20].

2.2 Procedure

We adopt a two-part approach combining human behavioral experiments and analyses of foundation model representations. In the behavioral study, 44 participants (22 NH, 22 CI) evaluated 70 musical excerpts using continuous valence and arousal ratings [21] in an online experiment. The ratings were analyzed using linear mixed-effects models across groups (NH vs. CI) and sources (Human vs. AI).

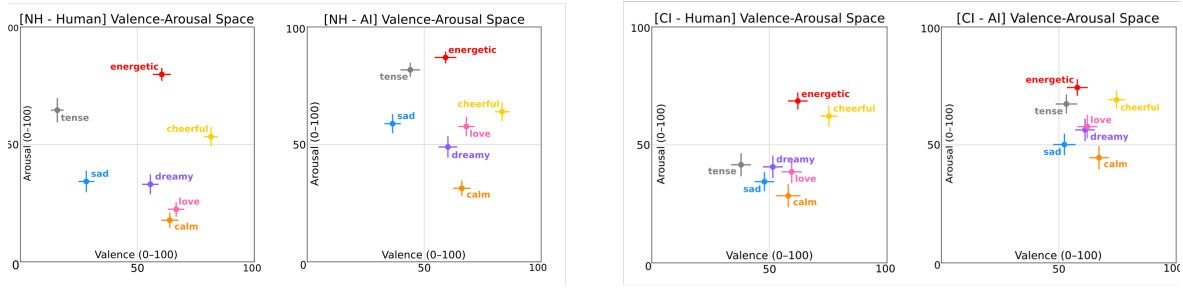

Figure 1. Valence–arousal (VA) ratings for each emotion category, comparing human-composed and AI-generated music in NH (left) and CI (right) groups.

		NH (Original)		CI (CI-sim)	
Approach		Human	AI	Human	AI
Human experiment		43.82	35.49	25.86	21.89
Foundation Model	MuLan	0.651	0.694	0.080	0.079
	CLAP	0.471	0.341	0.394	0.301
	MERT	0.198	0.153	0.104	0.059

Table 1. Emotion separability (mean between-emotion distance) in valence–arousal space (human experiment) and embedding space (foundation model).

In addition, emotion separability was quantified using a distance-based metric, computed as the between-category Euclidean distances in the VA space.

In our analysis of foundation models, we selected representative audio and music foundation models—MuLan, CLAP, and MERT [22–24]—and analyzed how their learned representations align with human emotional perception. Specifically, we evaluated (1) the alignment between embedding features and human valence–arousal (VA) ratings using ridge regression with cross-validation, and (2) the separability of emotion categories in embedding space using between-category cosine distances.

3. RESULTS

3.1 Human Behavioral Experiment

First, in the NH group, AI-generated music increased perceived arousal across emotions and selectively shifted negative emotions toward more positive valence, resulting in reduced emotion separability (**Figure 1**, **Table 1**). Second, within human-composed music, CI users showed more neutral overall ratings compared to NH listeners, while still perceiving negative emotions as less negative, which was accompanied by a reduction in emotion separability. Third, in the CI group, AI-generated music led to a general shift toward the high-valence, high-arousal region (first-quadrant), with emotional categories becoming even less distinguishable, indicating a further reduction in emotion separability. This pattern suggests a compounded effect of AI generation bias and auditory degradation.

3.2 Foundation Model Analysis

First, in predicting human valence–arousal (VA) ratings, models showed higher accuracy for NH group than for CI group overall. Within CI conditions, arousal ratings were relatively better preserved than valence, suggesting that arousal-related acoustic cues are more robust to spectral degradation and remain relatively intact in CI-simulated music. Among the models, CLAP exhibited the strongest alignment overall, including under CI conditions, consistent with the benefits of cross-modal pretraining. Second, embedding geometry analysis revealed patterns consistent with behavioral VA results, with at least two models showing overall reduced emotion separability for AI-generated and CI-simulated conditions, and the greatest reduction in the CI–AI condition (**Table 1**).

4. DISCUSSION

Across both behavioral and representation-level analyses, we observed compressed emotional representations, with AI-generated music reducing emotion separability even for NH listeners. This suggests that music generated from simple prompts (e.g., "{emotion category} music") exhibits reduced emotional separability compared to human-composed music. This limitation is further pronounced under degraded listening conditions: CI users showed additional compression, with the strongest effect in the CI–AI condition, a pattern also observed in foundation model embeddings. Together, these findings indicate that emotional limitations of AI-generated music are amplified under CI-related degradation, leading to the greatest reduction in emotional perception.

Furthermore, VA prediction results from foundation models suggest that current models capture emotional structure under NH conditions but fail to fully preserve it under CI conditions, highlighting limitations beyond NH listeners and the need for perception-aware, accessibility-driven approaches. Future work may explore mitigating the limited emotional range of AI-generated music through more diverse prompting strategies. Additionally, incorporating CI user data and emphasizing robust acoustic cues may further improve emotional expressivity. Notably, the strong performance of CLAP suggests that training on diverse audio inputs, including CI-simulated audio, can enhance robustness across listening conditions. Overall, these findings support the development of CI-accessible music AI systems beyond NH-centered paradigms.

5. AI USAGE STATEMENT

AI tools were used for language editing, including translation from Korean to English, grammar correction, and assistance with LaTeX formatting.

6. ETHICS STATEMENT

All participants provided informed consent prior to participation in the behavioral experiment, and the study was approved by an Institutional Review Board (IRB).

7. References

[1] T. DeNora, *Music in everyday life*. Cambridge university press, 2000.

[2] R. MacDonald, G. Kreutz, and L. Mitchell, *Music, health, and wellbeing*. Oxford University Press, 2013.

[3] S. Saarikallio and J. Erkkilä, "The role of music in adolescents' mood regulation," *Psychology of music*, vol. 35, no. 1, pp. 88–109, 2007.

[4] J. Schulkin and G. B. Raglan, "The evolution of music and human social capability," *Frontiers in neuroscience*, vol. 8, p. 292, 2014.

[5] B. Tarr, J. Launay, and R. I. Dunbar, "Music and social bonding:"self-other" merging and neurohormonal mechanisms," *Frontiers in psychology*, vol. 5, p. 1096, 2014.

[6] A. D. Patel, *Music, language, and the brain*. Oxford university press, 2010.

[7] P. N. Juslin and D. Västfjäll, "Emotional responses to music: The need to consider underlying mechanisms," *Behavioral and brain sciences*, vol. 31, no. 5, pp. 559–575, 2008.

[8] P. N. Juslin and J. Sloboda, *Handbook of music and emotion: Theory, research, applications*. Oxford University Press, 2011.

[9] P. N. Juslin, "From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions," *Physics of life reviews*, vol. 10, no. 3, pp. 235–266, 2013.

[10] H. J. McDermott, "Music perception with cochlear implants: a review," *Trends in amplification*, vol. 8, no. 2, pp. 49–82, 2004.

[11] V. Looi, K. Gfeller, and V. D. Driscoll, "Music appreciation and training for cochlear implant recipients: a review," in *Seminars in hearing*, vol. 33, no. 04. Thieme Medical Publishers, 2012, pp. 307–334.

[12] M. Moran, A. Rousset, and V. Looi, "Music appreciation and music listening in prelingual and postlingually deaf adult cochlear implant recipients," *International journal of audiology*, vol. 55, no. sup2, pp. S57–S63, 2016.

[13] K. Gfeller, J. Oleson, J. F. Knutson, P. Breheny, V. Driscoll, and C. Olszewski, "Multivariate predictors of music perception and appraisal by adult cochlear implant users," *Journal of the American Academy of Audiology*, vol. 19, no. 02, pp. 120–134, 2008.

[14] A. Lehmann, C. J. Limb, and J. Marozeau, "Music and cochlear implants: Recent developments and continued challenges," *Frontiers in Neuroscience*, vol. 15, p. 736772, 2021.

[15] J.-P. Briot, G. Hadjeres, and F.-D. Pachet, "Deep learning techniques for music generation–a survey," *arXiv preprint arXiv:1709.01620*, 2017.

[16] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer," *arXiv preprint arXiv:1809.04281*, 2018.

[17] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[18] H. Lee, F. Hoeger, M. Schoenwiesner, M. Park, and N. Jacoby, "Cross-cultural mood perception in pop songs and its alignment with mood detection algorithms," *arXiv preprint arXiv:2108.00768*, 2021.

[19] A. S. Cowen, X. Fang, D. Sauter, and D. Keltner, "What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures," *Proceedings of the National Academy of Sciences*, vol. 117, no. 4, pp. 1924–1934, 2020.

[20] M. Cousineau, L. Demany, B. Meyer, and D. Pressnitzer, "What breaks a melody: perceiving f0 and intensity sequences with a cochlear implant," *Hearing Research*, vol. 269, no. 1-2, pp. 34–41, 2010.

[21] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[22] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. Ellis, "Mulan: A joint embedding of music audio and natural language," *arXiv preprint arXiv:2208.12415*, 2022.

[23] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[24] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," *arXiv preprint arXiv:2306.00107*, 2023.