VIBE SORCERY: INTEGRATING EMOTION RECOGNITION WITH GENERATIVE MUSIC FOR PLAYLIST CURATION

Isabel Urrego-Gómez
Queen Mary University
i.urrego@se24.qmul.ac.uk

Simon Colton
Queen Mary University
s.colton@qmul.ac.uk

Iran Roman
Queen Mary University
i.roman@qmul.ac.uk

ABSTRACT

Vibe Sorcery generates emotionally coherent playlists using text-to-audio synthesis. The system creates dynamic musical journeys through Markov-like transitions, with each new track conditioned only on its immediate predecessor. Its three components work sequentially: the Listener predicts moods and genres, the Captioner converts these to text prompts, and Stable Audio synthesizes matching tracks. Evaluations show significantly smoother emotional progression than random sampling (average Arousal Valence-space distance: 0.82 vs. 2.4). This approach demonstrates how language-prompted audio generation can create controlled, adaptive listening experiences.

1. INTRODUCTION AND APPROACH

Vibe Sorcery is a mood-based playlist generator that leverages generative music to create emotionally-cohesive listening experiences. Conventional playlist generators treat songs as clusters, selecting based on shared patterns of the entire song set, such as the same artist, similar tempos or harmonic structures [17, 21, 6, 12, 16, 5]. This assumes that songs in a playlist should feel similar to one another, prioritizing uniformity across songs over the progression between them. We propose a system based on an alternative perspective: modeling playlist generation as a Markov process [14, 15], where each song depends only on the preceding one. In our system, both the genre and the mood of the current track shape the selection of the next song, with mood progression taking precedence and relying solely on the immediately preceding track rather than the entire playlist history. The core hypothesis is that an effective playlist mirrors an emotional journey with dynamic progression. While each transition between songs is determined locally, the overall sequence can still trace a path through widely different emotional states. These changes occur gradually, allowing the playlist to move, for instance, from somber and subdued moods to uplifting and energetic ones [22].

Recent advances in language-prompted generative music, for example Jukebox [9], MusicGen [7], and Stable Audio [10] have enabled high-quality music synthesis from textual prompts. However, the intersection of language-guided music generation and emotion-aware playlist curation remains underexplored. Building on this gap, we propose how emotion recognition models [1, 19] and language-prompted generative music models might converge to power adaptive musical experiences by generating original compositions that respond to emotional cues. We present Vibe Sorcery as a proof of concept for this synthesis, prioritizing the demonstration of feasibility over exhaustive performance optimization. Some parameters in the current system are heuristic choices that have proven effective but have not yet been systematically optimized, which we plan to address in future work. The open-source implementation is available on github [1] .

2. METHOD

Vibe Sorcery generates emotionally-coherent playlists through an iterative, three-component pipeline showed in Figure 1. The process begins with an track chosen by the user (the only point of direct user control over the generative process in the current implementation). This track is analyzed by the **Listener** to extract audio features and predict moods (56 classes) and genres (87 classes) using a multi-label classifier [2] trained on the MTG-Jamendo dataset [4]. Predictions are filtered using manually-tuned activation thresholds (0.07 for moods, 0.1 for genres). If no predictions surpass these thresholds, the top prediction is selected as fallback (see Appendix A). While these values demonstrated functional efficacy in our preliminary experiments, their optimization represents an opportunity for future research. The **Captioner** then converts these predictions into text prompts using predefined templates with placeholder fields, which are dynamically populated using the listener's predictions. For example, "A [**genres**] tune for your [**moods**] moments" and "The perfect [**genres**] soundtrack for your [**moods**] day". The resulting caption is passed to the **Generator** (Stable Audio Open 1.0 [10]), which synthesizes a song conditioned on the prompt. This song becomes the input to the listener in the next iteration.

To evaluate and visualize songs from an emotional perspective, the Arousal-Valence plane (AV plane) [20] is employed (see Appendix B), a two-dimensional model where valence (ranging from negative to positive) represents the pleasantness of a musical experience, while arousal (rang-

© . Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** , "Vibe Sorcery: Integrating Emotion Recognition with Generative Music for Playlist Curation ", in *Proc. of the 26th Int. Society for Music Information Retrieval Conf.,* Daejeon, South Korea, 2025.

[1] github.com/IsitaRex/Vibe-Sorcery

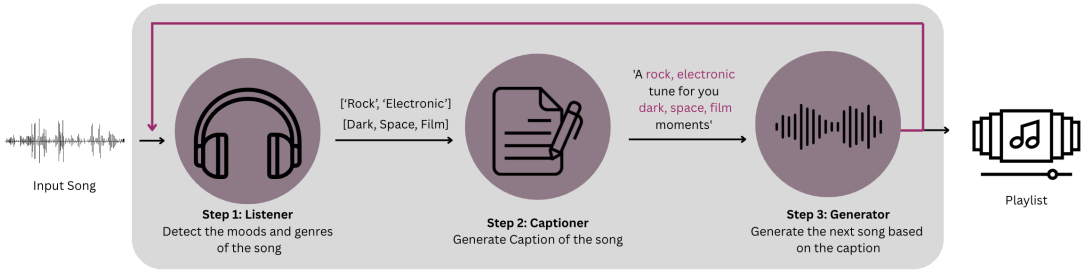

Figure 1. Overview of the Vibe Sorcery system architecture. The pipeline consists of three core components: (1) the **Listener**, which extracts audio features and predicts moods/genres from an input track; (2) the **Captioner**, which converts these predictions into text prompts using grammar templates and synonym substitutions; and (3) the **Generator**, which synthesizes a new track conditioned on the caption using Stable Audio. The process iterates to generate a playlist with emotionally coherent transitions, where each new track depends only on its immediate predecessor.

ing from calm to excited) captures the intensity of the emotion evoked. This framework enables songs to be mapped within an emotional 2D coordinate system. MusiCNN [18, 2] trained on DEAM dataset [24] is used to predict arousal and valence values for each song, ranging from 1 to 9. By plotting these values, songs can be visualized within this emotional space, and the playlist trajectories can be examined. This method provides both quantitative and qualitative insights into the flow of the generated playlists.

3. RESULTS AND DISCUSSION

The system was evaluated against a random playlist baseline. For the baseline, 100 songs were generated by independently sampling random mood and genre combinations and generating its captions with the Captioner. In this case, consecutivity is defined purely by the order of prompt generation. The average Euclidean distance between consecutive songs (measured in AV-plane) was 2.4 for the random approach. In contrast, 100 songs generated with VibeSorcery achieved an average distance of 0.82, demonstrating its ability to produce smoother and more coherent emotional transitions. Figure 2 compares a Vibe Sorcery-generated playlist with a randomly generated playlist. The random playlist exhibits abrupt transitions (e.g., from low to high arousal/valence) with an average distance of 2.09 and a maximum of 3.03 in the AV space. In contrast, Vibe Sorcery produces smoother transitions, achieving significantly lower average (0.37) and maximum (0.77) distances. This demonstrates the system's ability to maintain emotional coherence through mood-driven captioning.

4. CONCLUSIONS AND FUTURE WORK

In the current design, users control the starting point but not the endpoint, with emotionally coherent transitions guaranteed. While this proof-of-concept demonstrates feasibility, we aim to address the following limitations in future work: (1) the template-based captioner could be enhanced with LLMs, (2) key parameters like Listener thresholds need systematic optimization, (3) studying the impact of using different text-to-audio models for the system, (4) validate the alignment of prompts and their generated songs in the AV-plane to match target emotional regions, (5) broaden evaluation beyond random playlists to include comparisons with existing methods, and (6) complement computational measures with human evaluations to confirm listener-perceived coherence. Furthermore, incorporating intensity-aware affective representations into both the Listener and Captioner components could improve emotional modeling precision [26].

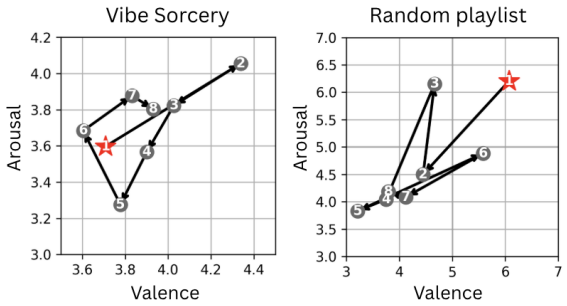

Figure 2. AV-plane of playlists generated by Vibe Sorcery (left) and randomly (right). Each song is a numbered circle, indicating the order in which the songs were generated. The input song is marked by a red star symbol. Arrows illustrate the transitions between tracks.

A deeper challenge for emotion-aware systems lies in the interpretation of music itself. Existing music (such as commercially released tracks) complicates emotional modeling due to listeners' subjective associations. For instance, a song heard at a funeral may forever evoke sadness, regardless of its compositional qualities. Generative music addresses this by circumventing two key issues: the variability of internal emotion perception, and the external contextual biases tied to familiar repertoire. Unlike existing music, which carries unpredictable personal histories, generative compositions lack pre-existing associations. This ensures that evoked emotions align more reliably with the music's intrinsic features, rather than individual memories or cultural context.

5. ETHICAL CONSIDERATIONS

Vibe Sorcery raises several ethical considerations that merit reflection. First, the system relies on pre-trained models that may encode cultural, demographic, or contextual biases, potentially resulting in uneven or misleading predictions across user groups. Moreover, current generative music models lack broad representation across genres and sonorities [28, 23], which may lead to systems that enhance the emotional experience for some listeners while marginalizing others. In the specific case of emotion recognition, the assumption that affective states can be universally inferred from musical features is problematic, as it risks oversimplification and misrepresentation given the subjective, culturally contingent, and context-dependent nature of emotional responses to music [13, 8, 27, 11]. Additionally, the lack of human validation in our evaluation means that we cannot yet assess how listeners experience these playlists emotionally. Ethical use would therefore require careful attention to user consent, transparency, and safeguards harmful mood manipulation.

6. ACKNOWLEDGEMENTS

The authors thank Meta Platforms Inc. for partially supporting this work through an unrestricted gift to Dr. Iran R. Roman's research program, which funded Isabel's attendance at ISMIR 2025. Meta had no role in the design, execution, or interpretation of the study.

APPENDICES

A. LISTENER DETAILS

The first stage of playlist generation involves analyzing an input track to predict its moods and genres. This is achieved by extracting audio embeddings and feeding them into two multi-label classifiers: one for mood and theme prediction, and another for genre prediction. These classifiers, along with their pre-trained weights, were obtained from the TensorFlow audio models library [2].

The embedding model is based on the EfficientNet architecture [25], trained on Discogs metadata [3] to predict 400 music styles. It processes raw audio signals and produces fixed-size embeddings.

For mood modeling, we use a classifier trained on the mood and theme subset of the MTG-Jamendo dataset [4], covering 56 classes. Predictions are filtered using an activation threshold of 0.07: labels exceeding this threshold are retained, with a maximum of four selected. If no label surpasses the threshold, the highest-scoring label is chosen as fallback. Genre prediction follows the same procedure, but with a classifier trained on 87 genre classes from a separate MTG subset and a threshold of 0.1.

B. AROUSAL VALENCE PLANE

The Arousal–Valence (AV) plane, also referred to as the Circumplex Model of Affect, was introduced by James A. Russell [20]. This model proposes that all emotional states can be characterized along two fundamental dimensions: valence, representing the degree of pleasantness, and arousal, denoting the intensity or level of activation.

On the AV plane, the x-axis corresponds to valence, ranging from negative emotions such as sadness, fear, anger, and disgust (left side) to positive emotions such as happiness, excitement, and contentment (right side). The y-axis corresponds to arousal, distinguishing between high-arousal emotions such as excitement, anxiety, anger, and alarm (upper half) and low-arousal emotions such as calmness, relaxation, boredom, and sadness (lower half). By combining these two independent dimensions, emotions can be systematically mapped onto a two-dimensional space. Accordingly, the plane is divided into four quadrants as shown in Fig A1.

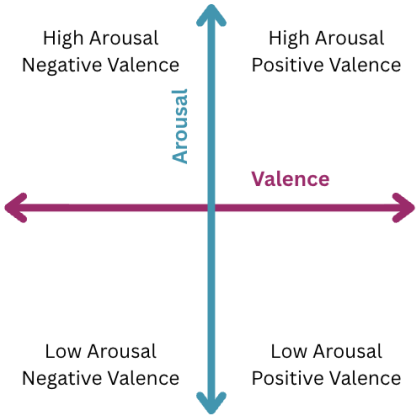

Figure A1. The AV plane maps emotions in a two-dimensional space: valence (x-axis) ranges from unpleasant to pleasant, and arousal (y-axis) from low to high activation. Quadrants correspond to High Arousal–Positive Valence (e.g., excitement), High Arousal–Negative Valence (e.g., fear, anger), Low Arousal–Positive Valence (e.g., calmness, contentment), and Low Arousal–Negative Valence (e.g., sadness, boredom).

REFERENCES

[1] Zakaria Aldeneh and Emily Provost. "You're not you when you're angry: Robust emotion features emerge by recognizing speakers". In: *IEEE Transactions on Affective Computing* 14.2 (2021).

[2] Pablo Alonso-Jiménez et al. "Tensorflow Audio Models in Essentia". In: *International Conference on Acoustics, Speech and Signal Processing* (2020).

[3] Dmitry Bogdanov and Xavier Serra. "Quantifying music trends and facts using editorial metadata from the discogs database". In: *International Society for Music Information Retrieval* (2017).

[4] Dmitry Bogdanov et al. "The MTG-Jamendo Dataset for Automatic Music Tagging". In: *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning* (2019).

[5] Brandon James Carone and Pablo Ripollés. "SoundSignature: What Type of Music do you Like?" In: *IEEE 5th International Symposium on the Internet of Sounds* (2024).

[6] K. Chen et al. "Learning Audio Embeddings with User Listening Data for Content-Based Music Recommendation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing* (2020).

[7] Jade Copet et al. "Simple and controllable music generation". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 47704–47720.

[8] Alan S. Cowen et al. "What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures". In: *Proceedings of the National Academy of Sciences* 117.4 (2020), pp. 1924–1934. DOI: 10.1073/pnas.1910704117.

[9] Prafulla Dhariwal et al. *Jukebox: A Generative Model for Music*. 2020. arXiv: 2005.00341.

[10] Zach Evans et al. "Stable Audio Open". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10888461.

[11] Petri Laukka et al. "Universal and Culture-Specific Factors in the Recognition and Performance of Musical Affect Expressions". In: *Emotion (Washington, D.C.)* 13 (Feb. 2013), pp. 434–449. DOI: 10.1037/a0031388.

[12] Jiaxin Li. "Music recommendation algorithm based on user portrait". In: *IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology*. 2021, pp. 10–13. DOI: 10.1109/CEI52496.2021.9574460.

[13] Marjorie G. Li, Kirk N. Olsen, and William Forde Thompson. "Cross-Cultural Biases of Emotion Perception in Music". In: *Brain Sciences* 15.5 (2025). DOI: 10.3390/brainsci15050477.

[14] Brian McFee and Gert Lanckriet. "The Natural Language of Playlists." In: *Proceedings of the 12th International Society for Music Information Retrieval Conference* (2011), pp. 537–542.

[15] Brian McFee and Gert RG Lanckriet. "Hypergraph models of playlist dialects." In: *Proceedings of the 13th International Society for Music Information Retrieval Conference* (2012), pp. 343–348.

[16] Hope Mogale and M.B Esiefarienrhe. "Optimizing Recommendation Algorithms Using Self-Similarity Matrices for Music Streaming Services". In: *International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems*. 2021, pp. 1–4.

[17] Steffen Pauws and Berry Eggen. "PATS: Realization and user evaluation of an automatic playlist generator". In: *Journal of New Music Research* 32 (2002). DOI: 10.1076/jnmr.32.2.179.16739.

[18] Jordi Pons and Xavier Serra. "musicnn: Pre-trained convolutional neural networks for music audio tagging". In: *arXiv preprint arXiv:1909.06654* (2019).

[19] Mariana Rodríguez Castañeda and I. R. Roman. "The Voice of an Instrument: Analysis of X-vectors for Music Applications". In: *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*. 2024. URL: https://ismir2024program.ismir.net/lbd_499.html.

[20] James A Russell. "A circumplex model of affect." In: *Journal of personality and social psychology* (1980).

[21] Badrul Sarwar et al. *Item-based collaborative filtering recommendation algorithms*. Hong Kong, Hong Kong, 2001. DOI: 10.1145/371920.372071.

[22] Harald Schweiger, Emilia Parada-Cabaleiro, and Markus Schedl. "The Impact of Playlist Characteristics on Coherence in User-Curated Music Playlists". In: *EPJ Data Science* (2025). DOI: 10.1140/epjds/s13688-025-00531-3.

[23] Ahmet Solak et al. "Bias Research on Generated Music from around the World". In: *Proceedings of the AES International Conference on Artificial Intelligence and Machine Learning for Audio (AES AIMLA)*. London, UK, Sept. 2025.

[24] Mohammad Soleymani, Anna Aljanaki, and Y Yang. "DEAM: Mediaeval database for emotional analysis in music". In: *Geneva, Switzerland* (2016).

[25] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[26] A. Wang and I. Roman. "Toward Affective Empathy in AI: Encoding Internal Representations of 'Artificial Pain'". In: *Cognitive Computational Neuroscience* (2025).

[27] Karn N Watcharasupat et al. "Uncertainty Estimation in the Real World: A Study on Music Emotion Recognition". In: *European Conference on Information Retrieval*. Springer. 2025, pp. 218–232.

[28] Gus Xia and Monojit Choudhury. "Music for All: Representational Bias and Cross-Cultural Adaptability of Music Generation Models". In: Association for Computational Linguistics. 2025.