# song_sent_scores: Computational Design for Charting Dynamic Emotion in Songs with a Multimodal Circumplex Framework

Frederico Gonçalves Pedrosa[1]

[1]Federal University of Minas Gerais, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, Brazil

**Abstract**

Music profoundly impacts human emotion, yet computationally modeling its dynamic affective qualities remains challenging. This pictorial introduces `song_sent_scores`, a novel open-source toolkit (R/Python) that operationalizes Russell's Circumplex Model of Affect to represent and visualize the dynamic interplay of Valence (pleasantness/unpleasantness) and Arousal (energy/calmness) in songs. The toolkit derives these dimensions multimodally from both the audio signal (via a CLAP model) and lyrical content (transcribed by ASR and analyzed with an NLI-based model). We visually demonstrate `song_sent_scores`' capabilities through: (1) a small-scale experiment comparing its affective classifications against human ratings, revealing varied agreement across songs (e.g., r = 0.79 for 'Negro Drama', r = -0.70 for 'My Baby Just Cares for Me'); and (2) a case study analyzing 'Negro Drama' via its circumplex trajectory and an autoregressive affective network. This computational design approach offers a richer, continuous representation of musical affect, providing a novel lens for designers, music therapists, and researchers to explore music's emotional architecture.

**Keywords**

Music, Artificial Intelligence, Psychometrics, Computational Design, Circumplex Model

## 1. Introduction

Music is a powerful medium for conveying and evoking emotion, yet computationally capturing its nuanced affective dynamics remains a significant challenge [1]. This pictorial introduces `song_sent_scores` [2], a novel open-source toolkit (available in R and Python) that operationalizes Russell's Circumplex Model of Affect [3] to represent and visualize the dynamic interplay of Valence (pleasantness/unpleasantness) and Arousal (energy/calmness) in songs, derived multimodally from both audio signals and lyrical content.

The `song_sent_scores` function leverages state-of-the-art learning models for its core analysis including 1) audio analysis with Contrastive Language-Audio Pretraining (CLAP) model [4] performs zero-shot classification on audio segments to infer Valence and Arousal directly from sonic characteristics; and 2) song lyrics, either provided or transcribed using an Automatic Speech Recognition (ASR) model (e.g., OpenAI's Whisper [5]), are analyzed for Valence and Arousal using a Natural Language Inference (NLI) based zero-shot text classification model [6]. The pipeline was inspired by Tomasevic and colleagues [7].

The toolkit processes song segments, yielding Valence and Arousal scores for both audio and lyrical modalities. These scores are then mapped onto a 2D circumplex plot, allowing for a dynamic visualization of the song's affective journey over time. This method moves beyond static categorization, offering a richer, continuous representation that can reveal subtle emotional shifts and the congruence between musical sound and lyrical meaning.

This pictorial will visually demonstrate `song_sent_scores`' capabilities. We begin by presenting a small-scale experiment to assess the correspondence between the tool's affective classifications and human ratings through statistical analysis. Building on these findings, we then showcase a detailed case study: the analysis of a song's temporal emotional development, visualized on the circumplex, and its affective trajectory explored via dynamic network analysis. We aim to showcase how this computational tool can enhance our understanding and discussion of dynamic musical affect, regarding Artificial Intelligence (AI) classification and computational design.

## 2. Related Work

The computational analysis and understanding of emotion in music is an intense research area within music informatics and affective computing [8, 1]. Researchers have explored both discrete emotion models (categorizing music into labels like happy, sad, angry) and dimensional approaches [9]. Among dimensional models, Russell's circumplex model of affect [3], which posits Valence and Arousal as core affective dimensions, is widely adopted due to its intuitive appeal, extensive empirical support (having garnered over 14,000 citations according to Semantic Scholar as of 2025), and applicability in computational settings [10, 11].

Recent advancements in zero-shot learning (ZSL) have opened new avenues for music emotion recognition. Contrastive Language-Audio Pretraining (CLAP) models, for instance, learn joint representations of audio and text, enabling the classification of audio based on natural language descriptions of target states [4]. Similarly, ZSL for text classification using Natural Language Inference (NLI) techniques allows for affective analysis of lyrics without task-specific training data [6]. Complementing these, some recent works in Zero-Shot Audio Classification (ZSAC) also explore leveraging Large Language Models to generate rich sound attribute descriptions to enhance classification [12], highlighting the growing synergy between audio processing and advanced language understanding.

Although many systems focus on static emotion classification of songs or segments, the temporal dynamics of affect in music are crucial for a deeper understanding, as music itself is a temporal art form [13]. Our work with `song_sent_scores` contributes to this area by not only providing multimodal circumplex scores, but also enabling the visualization of dynamic affective trajectories. The modeling of the interrelations between these affective streams is inspired by approaches such as dynamic Exploratory Graph

*Hybrid Workshop on Computational Design and Computer-aided Creativity (CDCC @ ICCC'25), June 23, 2025, Campinas, Brazil Online*

fredericopedrosa@musica.ufmg.br (F. G. Pedrosa)

https://github.com/FredPedrosa/ (F. G. Pedrosa)

0000-0002-0682-0734 (F. G. Pedrosa)

Analysis (EGA) [14, 7], which has been used in other domains to understand emotion dynamics. Furthermore, by comparing our toolkit's output with human perceptual ratings, we align with the critical need for robust evaluation methodologies in this field. While datasets for AI-driven affective music *generation* are emerging (e.g., SunoCaps by Civit and colleagues [15]), our focus is on providing an analytical tool for existing musical pieces.

## 3. The `song_sent_scores` Toolkit Architecture

Figure 1 illustrates the architecture and workflow of the `song_sent_scores` toolkit. The process begins with an audio file as the primary input and involves parallel pathways for audio and lyrical analysis to derive multimodal affective scores.

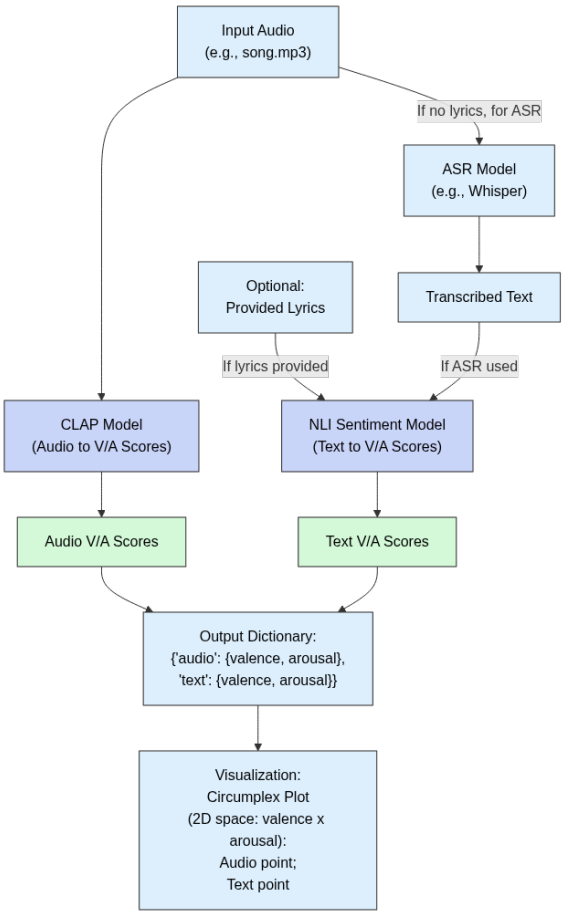


**Figure 1:** System architecture of the `song_sent_scores` toolkit, detailing the audio and text processing pipelines for deriving Valence and Arousal scores, and their subsequent visualization on a circumplex plot.

The core components are as follows:

1. **Input Audio:** The toolkit takes a song audio file (e.g., MP3, WAV) as input. This audio is used for both direct sonic analysis and, if lyrics are not provided, for ASR.
2. **Audio Affective Analysis (CLAP):** The raw audio signal is processed by a Contrastive Language-Audio Pretraining (CLAP) model. This model generates embeddings of the audio segments and performs zero-shot classification against predefined textual descriptions of affective states (e.g., 'audio with positive valence', 'audio with high arousal') to infer Valence and Arousal scores directly from the sonic characteristics. This results in the *Audio V/A Scores*.
3. **Lyrical Content Processing:**
   - **Optional Provided Lyrics:** Users can directly supply the song lyrics.
   - **Automatic Speech Recognition (ASR):** If lyrics are not provided, the input audio is fed into an ASR model (e.g., OpenAI's Whisper) to generate *Transcribed Text*.
4. **Text Affective Analysis (NLI):** The obtained lyrical content (either provided or transcribed) is then analyzed by a Natural Language Inference (NLI) based zero-shot text classification model. This model assesses the interaction between lyrics and phrases representing different Valence and Arousal states to produce the *Text V/A Scores*.
5. **Output Aggregation:** The derived *Audio V/A Scores* and *Text V/A Scores* are compiled into an output dictionary, typically structured with separate entries for 'audio' and 'text' modalities, each containing their respective valence and arousal values.
6. **Visualization:** Finally, these scores are mapped as coordinates onto a 2D circumplex plot. This plot visually represents the song segment's position in the Valence-Arousal space, with distinct points for the audio and text modalities, allowing for an intuitive understanding of the song's multimodal affective profile.

This pipeline allows `song_sent_scores` to offer a comprehensive, multimodal assessment of a song's perceived emotional character, moving beyond unimodal analysis or simplistic discrete emotion labels.

## 4. Methodology

### 4.1. Song Selection and Segmentation

Four songs were selected to ensure linguistic and genre variability: 'My Baby Just Cares for Me' composed and interpreted by Nina Simone (Jazz, English lyrics), 'Negro Drama' by the RAP group Racionais MCs (Brazilian Hip Hop, Portuguese lyrics), 'O Mundo é um Moinho' by the composer Cartola (Samba, Portuguese lyrics), and 'Territory' by Sepultura band (Thrash Metal, English lyrics). From each song, two distinct approximately 22-second excerpts (e.g., 'My Baby Just Cares for me' segment 1: 3-25s) were chosen for analysis, aiming to capture potentially different affective states.

### 4.2. Human Affective Ratings

Thirteen human judges (mean age = 45.92, SD = 14.13; 69.2% female; high academic background with 61.5% holding Master's or PhD degrees; 76.9% with at least one year of music study/practice) rated each of the 8 song excerpts. After listening to each excerpt, judges provided scores on 7-point Likert-type scales for perceived Valence (1=Very Negative/Unpleasant to 7=Very Positive/Pleasant) and Arousal (1=Very Low/Calm to 7=Very High/Agitated), separately for

the musical audio and the lyrical content (transcriptions were provided for lyrical assessment). Inter-rater reliability for these human ratings was assessed using the Intraclass Correlation Coefficient (ICC).

### 4.3. Computational Affective Scoring with `song_sent_scores`

The same 8 song excerpts were processed using the `song_sent_scores` toolkit. The function was configured to use `laion/clap-htsat-unfused` as the CLAP model, `openai/whisper-large-v3` as the ASR model, and `joeddav/xlm-roberta-large-xnli` as the NLI model. For each excerpt, `song_sent_scores` generated Valence and Arousal scores (ranging from -1 to +1) for both the audio (a_v, a_a) and the transcribed lyrical content (t_v, t_a).

### 4.4. Data Analysis

Human ratings were averaged across the 13 judges for each of the 32 items (4 songs x 2 excerpts/song x 2 modalities/excerpt x 2 dimensions/modality). These average human scores were then rescaled from their original 1-7 range to a -1 to +1 range to match the scale of the AI's scores. Pearson's correlation coefficient (r) was used to assess the correspondence between the rescaled average human ratings and the AI's scores, both overall and for specific subsets per song.

### 4.5. Case Study: Dynamic Analysis of 'Negro Drama'

To illustrate the toolkit's capability for dynamic analysis, the full song 'Negro Drama' was segmented into consecutive 15-second chunks. `song_sent_scores` was applied to each chunk to derive time-series data for Valence and Arousal from both audio and lyrics. These time-series were visualized to show the song's affective trajectory. Additionally, a autoregressive network model of these four affective streams (audio valence, audio arousal, text valence, text arousal) across the chunks was estimated using a Graphical Vector Autoregression (GVAR) approach with the `psychonetrics` [16] package in R, and visualized using `qgraph` [17].

## 5. Results

### 5.1. Inter-Rater Reliability of Human Ratings

The inter-rater reliability for human evaluations was calculated using the Intraclass Correlation Coefficient (ICC). The average ratings of the 13 judges demonstrated excellent reliability, with an ICC(2,k) of 0.83 (95% CI [0.73, 0.90]) for absolute agreement and an ICC(3,k) of 0.86 (95% CI [0.77, 0.92]) for consistency. This indicates that the aggregated human scores provide a robust benchmark for comparison.

### 5.2. `song_sent_scores` Example Output: Circumplex Plot

Figure 2 shows an example output from `song_sent_scores` for a segment of 'Negro Drama'. The

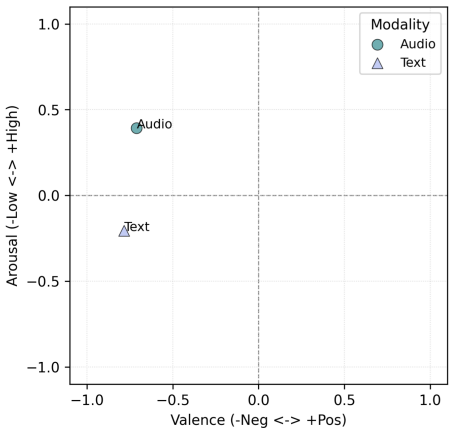


**Figure 2:** Example circumplex plot from `song_sent_scores` for a segment of 'Negro Drama'. The plot maps estimated Valence (X-axis) and Arousal (Y-axis) for audio (circle) and text (triangle). For this segment, audio shows negative valence and moderate-to-high arousal; text shows strong negative valence and lower arousal.

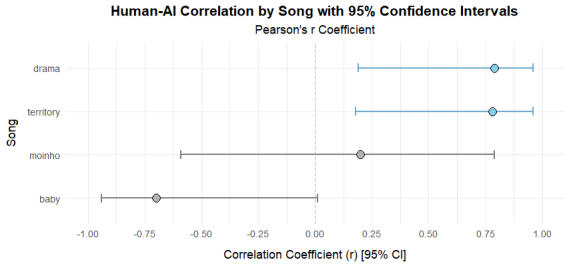


**Figure 3:** Pearson's r correlation coefficients (and 95% CIs) between average human ratings and `song_sent_scores` outputs for each song (N=8 items/song). Blue dots/bars indicate $p < 0.05$.

circumplex plot visually maps the estimated Valence and Arousal for the audio (circle marker) and the transcribed text (triangle marker) within the 2D affective space. For this specific segment, the audio was rated as having negative valence and moderate-to-high arousal, while the text was perceived as having strong negative valence and lower arousal.

### 5.3. Correlation between Human Ratings and `song_sent_scores`

Figure 3 presents the Pearson's r correlation coefficients between the rescaled average human ratings and the `song_sent_scores` outputs for each of the four songs, across all 8 items per song (2 excerpts x 2 modalities x 2 dimensions). The 95% confidence intervals for r are also shown. Strong and statistically significant positive correlations were observed for 'Negro Drama' (r = 0.79, p = 0.019) and 'Territory' (r = 0.78, p = 0.021), indicating good agreement between human perception and the AI for these songs. For 'O Mundo é um Moinho', the correlation was weak and not statistically significant (r = 0.20, p = 0.64), suggesting poor agreement. Notably, for 'My Baby Just Cares for Me', a strong negative correlation was found (r = -0.70, p = 0.054), suggesting that the AI's assessment tended to be opposite to human perception for this song. This warrants further investigation into the specific characteristics of this song or its processing by the models.

#### 5.3.1. Temporal Dynamics

Figures 4 and 5 illustrate the dynamic affective scores (Valence and Arousal, respectively, for both audio and text) for 'Negro Drama' when analyzed in consecutive 15-second chunks across the entire song. These visualizations allow for tracking the evolving affective arc of the song, highlighting moments of convergence or divergence between the emotional tone of the music and the lyrics across both fundamental affective dimensions.

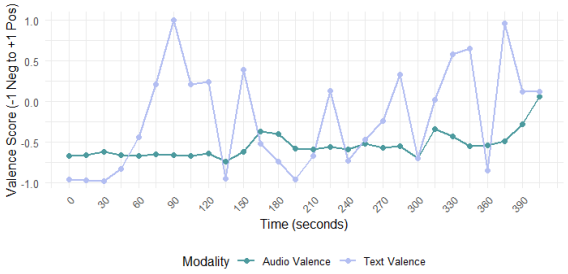


**Figure 4:** Dynamic Valence scores (audio and text) for 'Negro Drama' analyzed in 15-second chunks.

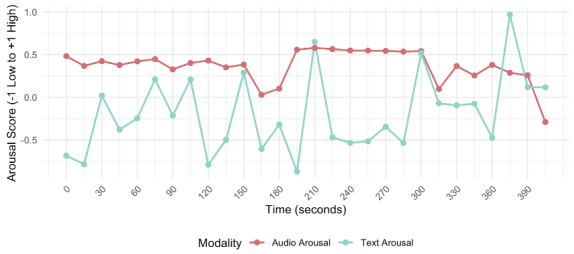


**Figure 5:** Dynamic Arousal scores (audio and text) for 'Negro Drama' analyzed in 15-second chunks.

Observing the Valence dynamics (Figure 4), for instance, around the 90-second mark, text valence becomes highly positive while audio valence remains negative. Issues with ASR transcription (e.g., transcribing 'Music' for instrumental segments) can also be identified, such as at the 75s and 105s marks, where text scores revert to a default pattern.

Similarly, the Arousal dynamics (Figure 5) reveal distinct trajectories for the audio and lyrical modalities. The audio arousal for 'Negro Drama' generally maintains a moderately high level throughout most of the song, consistent with its energetic hip-hop style, though with notable peaks and valleys reflecting structural changes in the music. Text arousal, on the other hand, shows more pronounced fluctuations. For example, a significant peak in text arousal is observed around the 300-second mark, corresponding to a climactic lyrical moment when MC Mano Brown begins to rhyme. The content of his lyrics intensifies the overall atmosphere.

#### 5.3.2. Temporal Influence Network (VAR Coefficients)

Figure 6 displays the temporal influence network estimated from the four affective time-series (audio valence, audio arousal, text valence, text arousal) of "Negro Drama". This network was derived from a Vector Autoregressive model of order 1 (VAR(1)) using the `psychonetrics` package [16]

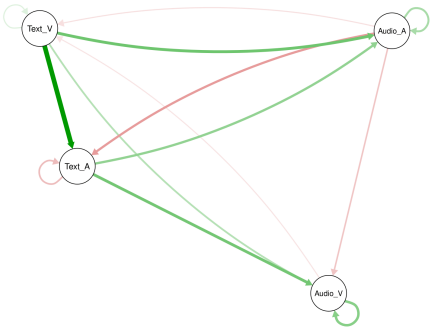


**Figure 6:** Temporal influence network for 'Nego Drama' derived from VAR(1) model coefficients (`beta` matrix). Nodes represent Audio Valence (Audio_V), Audio Arousal (Audio_A), Text Valence (Text_V), and Text Arousal (Text_A). Directed edges indicate the influence of a variable at time $t-1$ on another variable at time $t$. Green edges: positive influence; red: negative influence. Thicker edges: stronger influence. Loops: autoregressive effects (self-influence from $t-1$ to $t$).

and visualized with `qgraph` [17]. The directed edges represent how the state of one affective stream in a given 15-second chunk (time $t-1$) predicts the state of another (or the same) stream in the subsequent chunk (time $t$).

The network reveals the predictive dynamics between modalities and affective dimensions. For instance, one of the strongest predictive paths shows that higher **Text Arousal** at $t-1$ strongly predicts more positive **Text Valence** at $t$ (coefficient $\approx 0.56$). This suggests that energetic lyrical segments tend to be followed by segments perceived as more positively valenced in their lyrical content.

Autoregressive effects (loops on the nodes) indicate the temporal stability or inertia of each stream. **Audio Valence** shows moderate positive autoregression (coeff. $\approx 0.26$), implying that its current valence level tends to carry over to the next segment. In contrast, **Text Arousal** exhibits a weak negative autoregression (coeff. $\approx -0.14$), suggesting that peaks in lyrical energy might be followed by a slight decrease.

Other cross-lagged influences are also evident. For example, higher **Audio Arousal** at $t-1$ moderately predicts increased **Text Valence** at $t$ (coeff. $\approx 0.31$), while higher **Audio Valence** at $t-1$ moderately predicts increased **Text Arousal** at $t$ (coeff. $\approx 0.30$). These relationships highlight a complex interplay where the affective state of one modality in one moment can set the stage for changes in both dimensions of the other modality in the near future. This temporal network analysis provides insights into the evolving emotional narrative constructed by the interplay of music and lyrics in 'Negro Drama'.

## 6. Discussion

The `song_sent_scores` toolkit demonstrates a promising approach for operationalizing Russell's circumplex model of affect for multimodal song analysis. Building upon established dimensional theories [3, 10] and leveraging recent advances in zero-shot learning for audio and text processing [4, 6, 12], the dynamic visualizations generated by our toolkit offer richer insights into musical affect than static emotion categories traditionally allow [9, 13].

The comparison with human ratings yielded varied re-

sults across the selected songs. Strong positive agreement for Rap ('Negro Drama', r = 0.79) and Thrash Metal ('Territory', r = 0.78) suggests the toolkit can effectively capture the perceived affect in these well-known and more 'global' genres. However, the poor agreement for Samba ('O Mundo é um Moinho', r = 0.20) and the striking inverse relationship for Jazz ('My Baby Just Cares for Me', r = -0.70) highlight significant challenges. These discrepancies likely point to genre-specific sensitivities of the current AI models, ASR limitations (especially for non-sung segments or vocally complex styles like Jazz and Samba), and the potential misinterpretation of culturally or aged specific expressive cues not well-represented in the models' training data. The negative correlation for the Jazz piece, for example, warrants deeper investigation into how its unique vocal delivery, improvisation, and harmonic complexity might be processed divergently by the AI compared to human listeners.

The dynamic analysis of 'Negro Drama' (Figures 4, 5, and **??**) showcased the toolkits' utility in mapping evolving emotional trajectories and modeling interrelations between affective streams. This moves towards a systemic understanding of how audio and lyrical valence and arousal interact and influence each other over time, offering a novel quantitative method for computational musicology and affective computing, complementing qualitative approaches. The ability to identify, for instance, how Text Arousal might predict subsequent Text Valence, or the strong negative interplay between Audio Arousal and Audio Valence within this specific song, provides a granular view of its emotional architecture and yield for a better understanding of how music affect is perceived.

While this pictorial demonstrates clear potential, certain limitations of the current study and toolkit should be noted. The human validation involved 13 judges and four stylistically diverse songs; larger, more culturally varied participant samples and a broader selection of songs, particularly multiple examples within each genre, are needed for more generalizable conclusions. The performance of the constituent AI models (CLAP, Whisper, NLI) is also a factor, as their inherent biases and limitations can propagate through the pipeline.

Future work should therefore focus on several key areas. Refining ASR performance for diverse vocal and styles and sung Portuguese, as well as Brazilian musical peculiarities is critical. Exploring alternative or supplementary audio features beyond CLAP embeddings, perhaps incorporating more psychoacoustically-grounded or music-theoretically informed features, could enhance the audio analysis module. Furthermore, the toolkit could be extended to practical applications: interactive media designers could use it to dynamically score experiences, music therapists to select or analyze music for interventions, and researchers to conduct large-scale comparative studies of emotional expression in music.

## Declaration on Generative AI

During the preparation of this work, the author(s) used a large language model from Google (based on the Gemini family of models) in order to assist with grammar and spelling checks, provide help in coding, and for brainstorming and discussing ideas. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

## References


[1] D. Han, Y. Kong, J. Han, G. Wang, A survey of music emotion recognition, Frontiers of Computer Science 16 (2022) 166335. doi:10.1007/s11704-021-0569-4.

[2] F. G. Pedrosa, song_sent_scores: Functions for multimodal song circumplex analysis in r and python, 2025. URL: https://github.com/FredPedrosa/song_sent_scores, [Software].

[3] J. A. Russell, A circumplex model of affect, Journal of Personality and Social Psychology 39 (1980) 1161–1178. doi:10.1037/h0077714.

[4] B. Elizalde, S. Deshmukh, M. A. Ismail, H. Wang, CLAP: Learning audio concepts from natural language supervision, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '23), Rhodes Island, Greece, 2023. URL: https://resourcecenter.ieee.org/conferences/icassp-2023/spsicassp23vid0039.

[5] OpenAI, Whisper: Robust speech recognition via large-scale weak supervision, https://openai.com/index/whisper/, 2022. Accessed: 18/05/2025].

[6] W. Yin, J. Hay, D. Roth, Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 3904–3913.

[7] A. Tomasevic, H. Golino, A. Christensen, Decoding emotion dynamics in videos using dynamic exploratory graph analysis and zero-shot image classification: A simulation and tutorial using the transforemotion R package, 2024. URL: https://osf.io/preprints/psyarxiv/hf3g7. doi:10.31234/osf.io/hf3g7, psyArXiv.

[8] A. Dash, K. Agres, AI-Based Affective Music Generation Systems: A Review of Methods and Challenges, ACM Computing Surveys 56 (2024).

[9] M. Nunes-Silva, P. S. da Conceição Moreira, P. Eyer, Modelos cognitivos de emoções musicais: Abordagens discreta e dimensional, Percepta: Revista de Cognição e Artes Musicais 11 (2023) 23–38. doi:10.34018/2318-891X.11(1)23-38.

[10] J. Posner, J. A. Russell, B. S. Peterson, The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology, Development and Psychopathology 17 (2005) 715–734. doi:10.1017/s0954579405050340.

[11] Y.-S. Seo, J.-H. Huh, Automatic emotion-based music classification for supporting intelligent IoT applications, Electronics 8 (2019) 164. doi:10.3390/electronics8020164.

[12] X. Xu, P. Zhang, M. Yan, J. Zhang, M. Wu, Enhancing zero-shot audio classification using sound attribute knowledge from large language models, arXiv preprint arXiv:2407.14355 (2024). URL: https://arxiv.org/abs/2407.14355. arXiv:2407.14355.

[13] R. R. McCrae, Music lessons for the study of affect, Frontiers in Psychology 12 (2021) 760167. doi:10.3389/fpsyg.2021.760167.

[14] H. Golino, A. Christensen, EGAnet: Exploratory Graph Analysis – A framework for estimating the number of dimensions in multivariate data using network psychometrics, 2025. URL: https://r-ega.net. doi:10.32614/CRAN.package.EGAnet, r package version 2.1.1.

[15] M. Civit, V. Drai-Zerbib, D. Lizcano, M. Escalona, SunoCaps: A novel dataset of text-prompt based AI-generated music with emotion annotations, Data in Brief 55 (2024) 110743. doi:10.1016/j.dib.2024.110743.

[16] S. Epskamp, psychonetrics: Structural Equation Modeling and Confirmatory Network Analysis, 2024. URL: https://CRAN.R-project.org/package=psychonetrics. doi:10.32614/CRAN.package.psychonetrics, r package version 0.13.

[17] S. Epskamp, A. O. J. Cramer, L. J. Waldorp, V. D. Schmittmann, D. Borsboom, qgraph: Network visualizations of relationships in psychometric data, Journal of Statistical Software 48 (2012) 1–18.