STATISTICAL ANALYSIS OF MUSICAL FEATURES FOR EMOTIONAL SEMANTIC DIFFERENTIATION IN HUMAN AND AI DATABASES

Francisco BRAGA (0000-0002-9094-6320) [1*], **Jorge FORERO** (0000-0002-4662-8325) [2*], and **Gilberto BERNARDES** (0000-0003-3884-2687) [2]

[1] **NOVA University of Lisbon, School of Science and Technology & NOVA LINCS**, Lisbon, Portugal
[2] **University of Porto, Faculty of Engineering & INESC TEC**, Porto, Portugal
[*] These authors contributed equally.

ABSTRACT

Understanding the structural features of perceived musical emotions is crucial for various applications, including content generation and mood-driven playlists. This study performs a comparative statistical analysis to examine the association of a set of musical features with emotions, described using adjectives. The analysis uses two datasets containing rock and pop musical fragments, categorized as human-generated and AI-generated.

Focusing on four emotional adjectives (happy, sad, angry, tender-gentle) representing each valence-arousal plane's quadrant, we analyzed semantic differential meanings reported as symmetric pairs for all possible combinations of quadrants through diagonals, vertical, and horizontal axes.

The results obtained were discussed based on Livingstone's circular representation of emotional features in music.

Our findings demonstrate that the human and AI-generated datasets could be considered equivalent for diagonal symmetries, while horizontal and vertical symmetries show discrepancies. Furthermore, we assessed significant separability for both happy-sad and angry-tender pairs in the human dataset. In contrast, the AI-generated music exhibits a strong differentiation mainly in the angry-gentle pair.

1. INTRODUCTION

Music has long been associated with emotions and systematically studied across various disciplines since the beginning of the XX century.

The emotion felt and perceived in a musical fragment has been a topic of research since then, mainly using conceptual frameworks taken from psychology, such as categorical [1] and dimensional [2] models.

Diverse tests have been proposed to measure emotions from experiential, physiological, or behavioral perspectives [3]. Self-reports are commonly used for experimental evaluation. When using self-reports for categorical classification, participants label audio using adjectives. In contrast, a semantic differential approach is generally used in self-reports for dimensional analysis, where participants rank a pair of differential adjectives to build an emotional representation space, such as the valence-arousal plane. [1] Thus, language, or how we appraise emotions through these adjectives, is at the root of the experiential measurement methodology [5].

Furthermore, automatic music generation conditioned on textual description or melodic features has significantly progressed in recent years. In particular, some models, such as MusicLM [6] and MusicGen [7], have produced musical fragments within a short period of consistency, generated by prompting text with high-level music descriptions.

A crucial step in investigating the role of emotion in music involves identifying acoustic and structural elements reliably linked to emotional connotations.

Livingstone's circular representation of music and emotion [8] synthesizes an accumulation of structural features derived from studies reviewed by Schubert [9], Gabrielsson and Lindstrom [10], and Gabrielsson and Juslin [11] (see Figure 1). In this visualization, the placement of features along each axis of the valence-arousal plane corresponds to the frequency of studies indicating the association, with features positioned in quadrants if the association is confirmed across multiple independent studies.

To the best of our knowledge, no study has yet provided an empirical and systematic approach to statistical emotion differentiation using musical features within human-made music and semantically conditioned AI-generated music.

Although other research focuses primarily on human-generated datasets, exploring the relationship between musical features and perceived emotions [12], there is no evidence of a methodology incorporating statistical analysis to empirically compare and contrast AI-generated music.

This study aims to analyze emotion-related features in a dataset of music generated by a deep learning model and compare it with an existing human-made and annotated music-emotion dataset. We seek to investigate the confidence level over a set of musical features that differentiate emotional adjective pairs along the valence-arousal plane in both datasets. The comparison aims to highlight the differences and similarities in emotional expression between human and AI-generated music, contributing to the

[1] The adoption of semantically differential or polarized adjectives to infer meaning originated from Osgood's work [4].

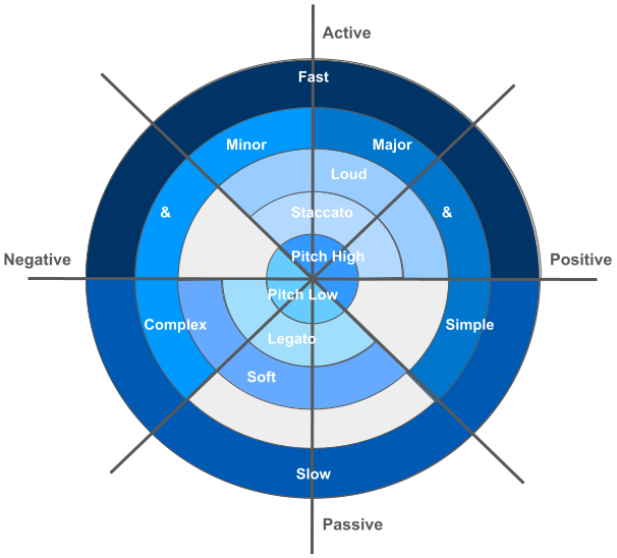


Figure 1. Livingstone Meta-analysis of emotional cues in music.

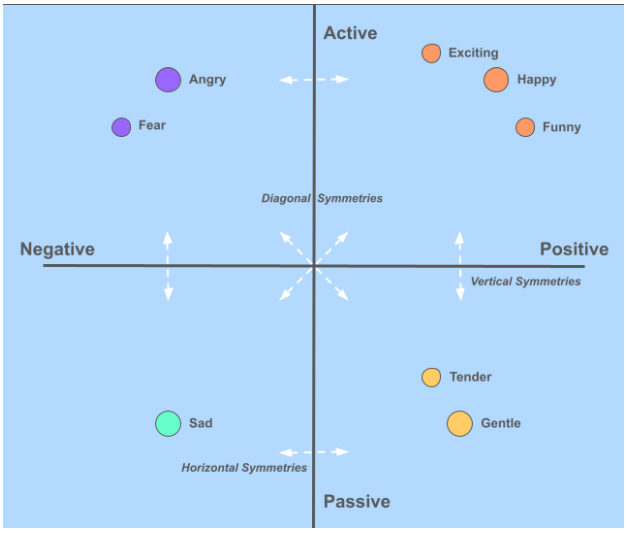


Figure 2. Emotional Adjectives and Symmetries in the Valence/Arousal space.

broader understanding of music's emotional conveyance for both scenarios.

The remainder of this paper is structured as follows. Section 2 describes the research methodology used, exposing the audio database used and the musical features selection. Section 3 provides the Results, proposing a radial plot to analyze the data. Finally, Section 4 offers a discussion and concludes in Section 5.

2. RESEARCH METHODOLOGY

We compared six pairs of adjectives distributed in the valence arousal plane and reported them as symmetric adjectives. The analysis uses two distinct datasets: one comprising music generated by humans (referred to as the "human-made" dataset) and the other containing music generated by a machine learning model conditioned by textual descriptions (referred to as the "AI-generated" dataset). The human dataset includes tracks annotated by human evaluators, while the AI-generated dataset's labels are based on semantics-driven input prompts used for generating the music.

We structured the input prompts based on Google's AudioSet ontology [13], which offers a comprehensive framework for categorizing audio. We adopted Russell's circumplex model and chose four emotional adjectives representing the four quadrants of the valence/arousal plane.

Our analysis aims to delineate relationships between these emotional states across differential adjectives conceptualized as three types of 2D symmetries: diagonal, horizontal, and vertical. For each symmetry, two pairs of emotions were analyzed: for the diagonal symmetries (different valence and arousal), the pairs Happy-Sad and Tender-Angry were analyzed; for the horizontal symmetry (different valence, same arousal), the pairs Happy-Angry and Tender-Sad, and for the vertical symmetry (same valence, different arousal) the pairs Happy-Tender and Sad-Angry (see Figure 2).

2.1 Databases

The AI-generated dataset was obtained using audios produced by MusicGen, an artificial intelligence model developed by Meta for generating music [7]. The model was used to create musical compositions based on text descriptions following Google's AudioSet ontology.

Prompts for the audio generation were created using the following phrase structure: - "A (mood) (genre) (role) (concept)" - with four variables "mood", "genre", "role", and "concept" (e.g., "A happy Rock soundtrack song").

In our study, "concept" and "role" remained constant as "soundtrack songs", while the chosen genres were Rock and Pop. Ten audios, 12 seconds long, were generated for each of the seven proposed emotions in the ontology (Happy, Funny, Sad, Tender, Exciting, Angry, Scary) for each selected genre. The AI-generated dataset comprises 80 musical fragments, 20 per emotion/quadrant, of which 10 were described within the genre of Rock music and the other ten were described within Pop music.

The human database, derived from the 4Q audio emotion dataset [2], includes tracks from AllMusic re-categorized in terms of Russell's quadrants [14].

Since not all adjectives proposed by the Google ontology were presented as tags in AllMusic musical fragments, we reduced our analysis to labels: Happy, Sad, Gentle, and Angry, ensuring one emotion per quadrant and assuming the adjectives gentle and tender as equivalent for the purposes of our analysis. From these, 20 samples per emotion, were randomly selected, leading to a human dataset comprised of 80 Rock and Pop music fragments, 20 per emotion/quadrant.

2.2 Features Selection

Based on Livingstone's music-emotion rules [8] and Panda's eight musical dimensions description [12], we selected seventeen musical features organized into a reduced

[2] Webpage: https://mir.dei.uc.pt/downloads.html

Table 1. Selected Features for each Musical Dimensions

Musical Dimension	Selected Features
Melody	Pitch Salience Mean
	Pitch Salience Stdev
Harmony	Chords Number Rate
	Chords Change Rate
	Dissonance Mean
	Dissonance Stdv
	HPCP Entropy Mean
	HPCP Entropy Stdev
	Scale (Krumhansl)
Rhythm	Tempo (BPM)
	Tempo Histogram Entropy
	Rhythm Onset Rate
Dynamic	Average Loudness
Tone Color/Timbre	Spectral Kurtosis Mean
	Spectral Kurtosis Stdv
	Spectral Centroid Mean
	Spectral Centroid Stdv

set of five categories (See Table1):

Melodic Features: A successful strategy for melody extraction derived from the salience-based methods [15]. We include Pitch Salience (mean and standard deviation) to capture the perceptual prominence of pitch.

Harmonic Features: Harmonic Pitch Class Profile Entropy (mean and standard deviation) quantifies the diversity of harmonic content, Chord Changes Rate assesses the dynamism in harmony, and Scale (major or minor) provides insights into the tonal context.

Additionally, the chord number Rate evaluates the harmonic variety, all of which might give insights into emotional nuances in music.

Rhythmic Features: Onset Rate Measures and Tempo (including Tempo Histogram Entropy) provide insights into the structure and pace of the music, which might reflect its energetic and, consequently, emotional states.

Dynamic Features: Average Loudness is considered to gauge the intensity level of the music, offering clues about its emotional impact.

Tone Color/Timbre Features: Spectral Centroid and Spectral Kurtosis (both mean and standard deviation) alongside Roughness/Sensory Dissonance explores the sound's texture and quality, which are crucial for emotional coloring.

Several algorithms have been proposed for audio analysis in music information retrieval. We used Essentia library [16] to compute the selected features.

2.3 Statistical Analysis

To assess the significance of the selected musical features across the possible emotional pairs, we employed various statistical tests tailored to the nature of the features under examination.

For continuous features, the Shapiro-Wilk test was initially conducted to assess normality, accompanied by Levene's test for evaluating equality of variances across groups. If both normality and equality of variances criteria were checked, the Student's t-test was utilized to determine the statistical significance of differences between groups. In cases where data passed the normality test but failed the equality of variances test, Welch's t-test was applied to account for variance discrepancies. For data not satisfying either normality or equality of variances, the non-parametric Mann–Whitney U test was selected to compare differences between groups.

The Chi-square test was deployed for categorical features to assess statistically significant differences among each pair of emotional states.

The significance value, or p-value, derived from the Student's t-test, Welch's t-test, Mann–Whitney U test, and Chi-square test, quantifies the probability that the observed differences between groups occurred under the null hypothesis of no actual difference. A lower p-value suggests a higher statistical significance, indicating strong evidence against the null hypothesis and implying that the observed differences are likely attributable to the variables under investigation rather than chance.

Across all tests, a significance level (α) of 0.05 was adopted, indicating that results with a p-value less than 0.05 were considered statistically significant, reflecting a less than 5% probability that observed differences occurred by chance.

3. RESULTS AND ANALYSIS

We present the findings of our study, analyzing the results derived from both human and AI-generated datasets [3] .

The significance values across the datasets are presented in Figure 3. These values are depicted as the confidence value, $1 - p$-value, to represent the significance more intuitively; higher values thus indicate greater statistical significance.

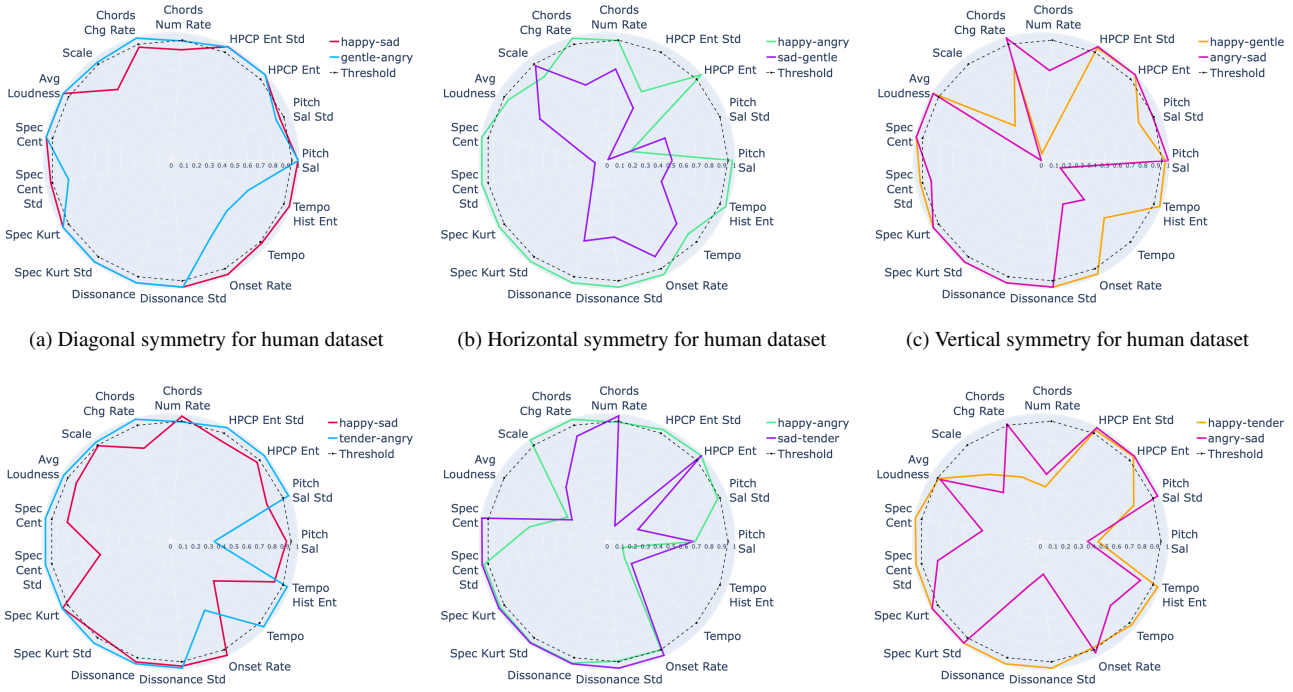
3.1 Diagonal symmetries

In analyzing the diagonal symmetries, we searched for contrasts in both valence and arousal dimensions. This approach entails comparing "happy" with "sad" and "tender/gentle" with "angry". Significance values derived from this analysis are illustrated in Figure 3a for the human dataset and Figure 3d for the AI-generated dataset.

3.1.1 Happy-Sad

The human dataset shows significant differences across features within the melodic, harmonic, rhythmic, and tim-

[3] Code for the data processing and feature extraction, statistical analysis, and further results can be found at `https://github.com/braga1376/musical-features-emotional`.



(a) Diagonal symmetry for human dataset (b) Horizontal symmetry for human dataset (c) Vertical symmetry for human dataset

(d) Diagonal symmetry for AI-generated dataset (e) Horizontal symmetry for AI-generated dataset (f) Vertical symmetry for AI-generated dataset

Figure 3. Comparative analysis of the confidence value ($1 - p$-value) across different movements for human and AI-generated datasets. From left to right: diagonal, horizontal, and vertical movements.

bral musical dimensions that effectively distinguish the emotion expressed by music. These differences highlight "happy" music's diverse chord usage, higher loudness levels, and faster tempo, as well as brighter, more complex, and rougher timbre compared to "sad" music.

For the AI-generated dataset, the analysis shows differences across features in the harmonic, rhythmic, and tone color/timbral categories. "Happy" music exhibits a greater diversity of chords and more frequent harmonic changes than "sad" music, indicating a richer harmonic complexity.

3.1.2 Tender/Gentle-Angry

Considering the human dataset, our results show statistical differences across all musical dimensions under analysis except rhythmical. The results indicate "angry" music's tendency towards greater harmonic complexity, the use of the minor scale, as well as brighter, more complex, and rough sound texture when compared to "gentle" music.

The AI-generated dataset has pronounced differences across all musical dimensions. These differences denote "angry" music's diverse chord usage, fast harmonic changes, and faster tempos, contrasting with "tender" music's less varied harmony and more constant timbre.

3.2 Horizontal Symmetry

Horizontal symmetry contains emotional adjectives with the same arousal but opposite valence within the Russell space. This symmetry compares high-arousal emotional descriptors as "happy" and "angry" and low-arousal descriptors as "sad" and "tender/gentle'. The findings from this comparative study are depicted in Figure 3b for the human dataset and Figure 3e for the AI-generated dataset.

3.2.1 Happy-Angry

Significant differences were found in the human dataset for the "happy-angry" pair across all musical dimensions except for dynamics. These differences suggest "angry" music's broader range of chords and more frequent chord transitions, brighter timbre, and greater dissonance, as well as less frequent tempo changes and greater tempo variability compared to "happy" music.

In the AI-generated dataset, the differentiation between "happy" and "angry" music primarily focuses on harmonic, rhythmic, and timbral features. At the same time, the melodic and dynamic dimensions show less significant features that differentiate both groups. The results indicate that "angry" music has a more dynamic harmony and greater chord diversity, often utilizing the minor scale, in contrast to the predominantly major scale of "happy" music, with more frequent changes, brighter spectral range, and increased auditory dissonance.

3.2.2 Sad-Tender/Gentle

Considering the human dataset for the "sad" and "tender" pair, no feature shows statistical significance, highlighting a possible similarity between the music that expresses the mentioned emotions.

The AI-generated dataset, in turn, denotes significant differences predominantly in features belonging to harmony, rhythm, and timbre dimensions. These differences indicate that "sad" music has a more diverse chord usage, less

frequent changes, and brighter timbre when compared to "tender" music.

3.3 Vertical Symmetry

In the vertical symmetry analysis, we explored emotional adjectives with similar valence but opposite arousal levels within the Russell space. This involves the comparative analysis between the positive valence adjectives "happy" and "tender/gentle" and the negative valence emotions "angry" and "sad". Results from this comparative analysis are presented in Figure 3c for the human dataset and Figure 3f for the AI-generated dataset.

3.3.1 Happy-Tender/Gentle

In the analysis of the human dataset for the "happy-gentle" pair, all categories of musical features show significant differences. The separability of the musical features highlights that "happy" music has a more diverse chord usage, brighter timbre, higher auditory dissonance, more frequent changes, and lower tempo variability.

In the AI-generated dataset, the distinction between "happy" and "tender" music is marked by significant differences in features in the timbral, harmonic, and rhythmic musical dimensions. These differences suggest "happy" music's more diverse chord usage, higher tempo, and brighter timbre.

3.3.2 Angry-Sad

Considering the human dataset, significant differences across melody, harmony, dynamics, and timbre features permit the distinction between "angry" and "sad" music. These results point to "angry" music's more diverse chord usage, brighter timbre, and increased auditory dissonance compared to "sad" music.

The AI-generated dataset also has significant differences across the melodic, harmonic, rhythmic, and timbral dimensions but has fewer features in each category. These suggest "angry" music's more diverse and frequently changing chords as well as more frequent changes.

4. DISCUSSION

We evaluated our findings according to Livingstone's framework on musical emotion featuring. Afterward, we compare human and AI-generated datasets regarding semantic differentiation through a set of musical features.

Our findings broadly agree with Livingstone's music-emotion rules, predicting differences across all musical characteristics for the diagonal symmetries analysis.

Specifically, for the "happy-sad" comparison, significant differences are observed in all musical feature categories, with more pronounced differences in the human-generated dataset. Similarly, the "tender/gentle-angry" comparison reveals significant differences across most musical dimensions. However, an exception arises with rhythm in the human dataset, where no significant differences are found, diverging from Livingstone's framework.

Our results intersect with and diverge from Livingstone's studies regarding horizontal symmetry, similar arousal levels, and opposite valence values.

For the "happy-angry" pair, our analysis aligns with Livingstone's hypothesis regarding dynamics and tempo consistency, as no significant differences in tempo were found. Some harmonic features significantly differed, supporting expected variances in harmony complexity. However, significant differences in timbre diverge from Livingstone's model.

The "sad-tender/gentle" pair further complicates the picture. The human dataset's tendency towards significant difference in scale aligns with expected mode consistency, yet harmony complexity remains unchanged. The AI-generated dataset, however, shows significant timbre differences and rhythmic distinctions through Onset Rate, not Tempo, diverging from Livingstone's model.

Results on the vertical symmetry partially align with Livingstone's findings.

For the "happy-tender/gentle" comparison of high valence emotions, significant differences are noted across all musical dimensions in the human dataset and into the timbral and rhythmic categories in the AI-generated dataset, differing from Livingstone's expectations by revealing distinctions in harmony complexity.

However, the expected consistency in scale is observed, with no significant differences in this feature. Similarly, the "angry-sad" comparison, reflecting low valence emotions, shows a broad feature differentiation in the human dataset and more constrained distinctions in the AI-generated dataset, corresponding to the expected scale consistency. Despite this, differentiation in harmony complexity and minimal rhythmic distinction contradicts Livingstone's model.

When comparing human and AI-generated datasets, the diagonal symmetries reveal a minimal divergence between these, although more differentiation is observed in the human dataset. This alignment suggests that although both sources adequately separate these emotions, the human database has a more robust distinction for the same musical features, especially in the happy-sad emotional pair.

The horizontal symmetry shows distinctions in the human dataset influenced by arousal intensity. High arousal conditions yield significant differentiation, whereas low arousal conditions do not. The AI-generated samples display a uniform response considering the arousal value. The number of significant features in the timbre categories increases notably for the "sad-tender" emotional pair compared to the human database.

In vertical symmetry comparison, the human dataset does not depend on the valence for distinguishing emotions, unlike the AI-generated dataset, which exhibits dependence on valence. Specifically, high valence in the AI-generated dataset shows greater differentiation than low valence, indicating the capability to generate music more distinguishable regarding valence.

5. CONCLUSIONS

Although the semantic representation of emotions is highly subjective, this study gives insights into the relationship between musical features and emotional states across human and AI-generated Rock/Pop music, focusing on diagonal, vertical, and horizontal differences within the valence/arousal space.

Our findings indicate significant separability for diagonal symmetry pairs in the human dataset, whereas the AI-generated music exhibits clear separability primarily in the angry vs. tender pair. Additionally, we inferred that the level of arousal influences the separability of horizontal symmetries for the human dataset and valence influences, although less, the separability of vertical symmetries for the AI-generated dataset.

We also conclude that our work aligns with Livingstone's regarding diagonal comparison and, to some extent, regarding horizontal and vertical. Still, it is essential to note the ability of the text-to-music model MusicGen [7] to generate notably different music when comparing the semantics from the prompt and mostly aligned regarding the valence/arousal comparisons.

Thus, this investigation contributes to the field by highlighting the differences in emotional perception between human and AI-generated music, offering a novel methodology for future research in music information retrieval and the analysis of emotion.

Further research could expand on two strands. First, the musical features used could be extended and further studied, in line with work on musical features for emotion recognition [12], to better understand the part played by the different musical categories and make a more robust analysis. Second, it could expand on the comparison between human and AI-generated music across various genres and emotional states, contributing to the evolving dialogue between musicology, psychology, and artificial intelligence.

Acknowledgments

The authors would like to thank Nuno Correia and Roger B. Dannenberg for their input on this work. This work was partially funded by Portuguese national funds through Fundação para a Ciência e Tecnologia (FCT) under CMU Portugal, by the Ph.D. scholarship with references PRT/BD/154690/2023 and 2022.11918.BD

6. REFERENCES

[1] P. Ekman, "Universal and cultural differences in facial expression of emotions," *Nebraska Symposium on Motivation*, pp. 207–283, 1972.

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

[3] I. B. Mauss and M. D. Robinson, "Measures of emotion: A review," *Cognition and emotion*, vol. 23, pp. 209–237, 03 2009.

[4] C. E. Osgood, G. J. Suci, and Percy H. Tannenbaum, *The Measurement of Meaning*. University of Illinois Press, 1957.

[5] J. Forero, G. Bernardes, and M. Mendes, "Are words enough?" *AIMC 2023*, aug 29 2023, https://aimc2023.pubpub.org/pub/9z68g7d2.

[6] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank, "Musiclm: Generating music from text," 2023.

[7] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[8] S. R. Livingstone, R. Muhlberger, A. R. Brown, and W. F. Thompson, "Changing musical emotion: A computational rule system for modifying score and performance," *Computer Music Journal*, vol. 34, no. 1, pp. 41–64, 2010.

[9] E. Schubert, "Measurement and time series analysis of emotion in music," 1999.

[10] A. Gabrielsson and E. Lindstrom, "The influence of musical structure on emotional expression." 2001.

[11] A. Gabrielsson and P. N. Juslin, "Emotional expression in music." pp. 503–534, 2001.

[12] R. Panda, R. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: a survey," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 68–88, 2020.

[13] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[14] R. Panda, R. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, vol. 11, pp. 614 – 626, 03 2018.

[15] J. Salamon and E. Gomez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[16] D. Bogdanov, N. Wack, E. Gómez Gutiérrez, S. Gulati, H. Boyer, O. Mayor, G. Roma Trepat, J. Salamon, J. R. Zapata González, X. Serra *et al.*, "Essentia: An audio analysis library for music information retrieval," in *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); p. 493-8.* International Society for Music Information Retrieval (ISMIR), 2013.