Analyzing Emotions in Music Using Low-level Acoustic Features

Abigail Wiafe*
School of Computing,
University of Eastern Finland,
P.O. Box 111, FI-80101
Joensuu
wiafe@cs.uef.fi
*corresponding author

Sami Sieranoja
School of Computing,
University of Eastern Finland,
P.O. Box 111, FI-80101
Joensuu

Pasi Fränti
School of Computing,
University of Eastern Finland,
P.O. Box 111, FI-80101
Joensuu

2025 6th International Conference on Computers and Artificial Intelligence Technology (CAIT) | 979-8-3315-5882-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/CAIT68620.2025.11424831

***Abstract*—Music can evoke, influence and alter an individual's emotional state. However, the task involved with music emotion recognition (MER) is still challenging due to the subjective nature of emotional experiences and the lack of universally accepted emotion taxonomies in music. This study investigates how low-level acoustic features can be used to link to emotional responses to music. A correlation analysis was performed to investigate how low-level acoustic features correlate with perceived emotions in music. The findings revealed that spectral features positively correlate with energizing, but weakly with other positive emotions such as joyful, happy and amusing. Features such as spectral centroid, root mean square (RMS) and spectral flux features positively correlated with energizing and negatively to relaxing and sad. These findings illustrate the importance of understanding how low-level acoustic features influence emotional response in music.**

Keywords—Emotions, energizing, low-level features, spectral flux

I. Introduction

Music emotion recognition (MER) is an interdisciplinary field that integrates signal processing, machine learning, music theory, auditory perception, and psychological principles to analyze the emotional content in music. Recognizing the emotional content of music can offer novel approaches in music therapy, enhance user experience in music-streaming services by enabling more personalized and mood-congruent playlists, and create a more immersive environment in video games and virtual reality settings. Understanding the MER concept is challenging and varies depending on the methodology utilized, the emotional model adopted and individual differences among the listeners involved. MER is a field of music information retrieval (MIR) that focuses on identifying and categorizing emotional content in music. It involves the computational extraction of music information and makes it accessible to users for various purposes; thus, algorithms have been designed to analyze music in audio format [1] and interpret the emotions conveyed.

A key distinction among MER methods lies in the choice between dimensional and categorical emotional-response models. The categorical model uses qualitative descriptors such as sadness, happiness, and anger to identify emotions. However, a drawback of this method is the inability to capture the complexity and richness of music emotions perceived by humans because there is a limitation on the number of emotional classes. Using dimensional models, listeners report perceived emotions in real time using continuous values in vector space [2]. The mostly used model is represented by valence (pleasant/unpleasant) and arousal (activation/deactivation) in a two-dimensional space [3]. Different low-level acoustic features have been used for MER. These include mel-frequency cepstral coefficients (MFCCs), spectral descriptors, energy measures and zero-crossing rate, integrated with machine learning methods to map musical data with emotional dimensions. For instance, spectral and temporal features were shown to distinguish between emotional categories in [4], whereas the role of timbre-related features were considered in [5]. MFCC was shown to improve emotion recognition in [6]. All these methods aimed to recognize emotions using acoustic features.

Deep learning was employed in [7] to recognize emotions in musical instruments using MFCC features. The study in [8] revealed that acoustic cues play a role in identifying positive and negative emotions in music.

The relationship between emotions and low-level features in Classical music was studied in [9] using secondary emotions such as tension and energy. The study revealed that potency correlated with features such as spectral flux, loudness, and roughness. The relationship between emotions and musical features in Western music was studied in [10]. The study concluded that low-level spectral and temporal features, such as chroma and flatness are effective in modelling the arousal dimension of emotions, whereas high-level features, such as mode, pulse clarity, and articulation are used to measure the cognitive aspect of the valence dimension. The study [11] investigated how human perception of tempo correlates with audio descriptors when listening to classical orchestra music (Beethoven's 3rd Symphony "Eroica"). The study concluded that audio descriptors related to timbre, rhythm, loudness, and harmony exhibited high correlation ratings.

In this paper, we follow the previous studies by analyzing how well a given low-level acoustic features correlate to emotions in the Emotify+ dataset [12]. It includes 400 song excerpts from four genres (pop, classical, rock, and electric) which were annotated by a total of 181 listeners (mostly of African origin) with the perceived emotions. The following ten emotions were used: dreamy, neutral, relaxing, sad, annoying, amusing, happy, energizing, joyful, and anxious.

We study the following low-level acoustic features.

- Spectral centroid
- MFCCs
- Spectral flux
- Zero-crossing rate
- RMS energy

Spectral centroid indicates the brightness of a sound. It correlates with the perceived energy of music, which is a factor known to influence emotional perception. MFCCs are known to capture timbral characteristics [13]. Spectral flux captures dynamic changes and harmonic content. It is a strong indicator of arousal, as higher flux correlates with high arousal (energetic), while low flux aligns with low arousal (calm) states. Zero-crossing rate captures textural roughness or smoothness, which is connected to the arousal-related dimension of emotion. RMS energy captures the emotional intensity of a music signal and serves as a strong acoustic cue to distinguish between high-arousal and low-arousal emotional states. The combination of spectral centroid, MFFCs, spectral flux, zero-crossing rate, and RMS produces a complementary set of descriptors that help to define rhythm, dynamic and timbre with an emphasis on a single feature. The results of this study will increase understanding of emotions in music research.

II. Audio Features and Emotions

Music is composed and performed to express emotions and feelings. Matching musical features with the emotions expressed is challenging for humans, as emotions are subjective experiences. Therefore, the choice of the correct technique must be carefully selected to reveal the connections between the acoustic features and musical emotions. The results in [5] suggested that categorizing musical elements such as rhythm, dynamics, melody, harmony, and timbre into four to eight groups can aid in the expression of musical emotions through audio features. Timbre can be determined by both the spectral and temporal features of sound [14]. Rough timbre (sound quality) is associated with anger, whereas a pure tone -like sound quality is associated with happiness [15]. Rough harmony [16] is associated with anger, and loud harmony with high volume [17] is associated with energy [18]. Fast tempo is associated with positive emotions, and a slow tempo with negative emotions [19]. Musical features extracted from audio are often grouped into low-level or high-level features [20]. Low-level features are directly derived from audio signals, including spectral centroids and mel-frequency cepstral coefficients (MFCC). They capture characteristics of audio signals, such as frequency, time, and amplitude domain properties. Their advantage is simplicity [21]. High-level features, on the other hand, relate to the perceived attributes of music, such as harmony, rhythm, and timbre.

Among the low-level features, spectral flux, loudness, and roughness have been found to correlate with secondary emotions, such as tension and energy [9]. This highlights their usefulness in capturing nuanced emotional responses to music. Temporal features, such as chroma and flatness, were found to be effective in modelling the arousal dimension of emotions in [10], further validating the selection of these features for our study.

III. Dataset and Methodology

The Emotify music dataset [22] consists of 400 songs (44100 Hz, 128 kbps, one minute each) from four different genres: classical, rock, pop, and electronic music. It was created from 400 musical pieces selected from the Magnatune recording company. The reason to use this source was that they were less popular and the listeners less likely to know them beforehand. This reduces the potential bias for the precondition perceived emotions [23]. The genres to the songs were assigned by the recording company. The dataset contains music from 241 albums by 140 performers. The songs are in English, and the files in mp3 format.

Emotify was later extended to Emotify+ through emotion annotations collected using the EF Music tool developed by the University of Eastern Finland's Machine Learning [12]. In total, 181 participants listened to multiple songs and noted their emotions, resulting in an extended Emotify+ dataset. The overview of the approach employed is shown in Fig. 1.

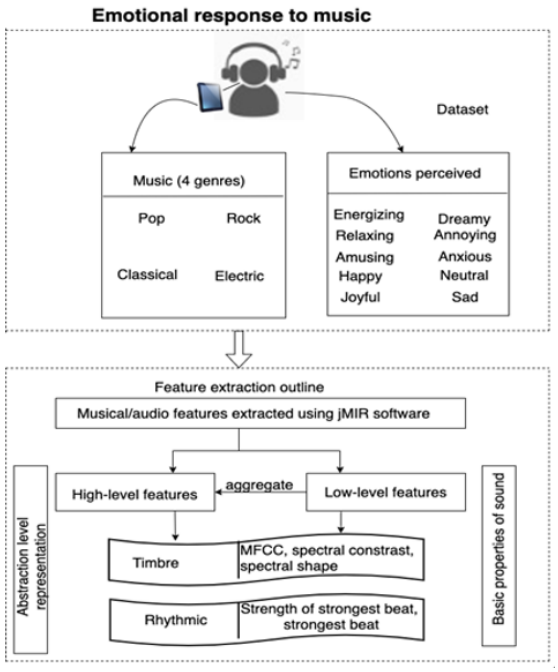

Fig. 1. Overview of our approach

A. Software Tool for Audio Extraction

To extract the acoustic features, we used the jMIR tool [24], which is an open-source Java software suite consisting of jAudio, jSymbolic, jWebMiner, and jLyrics. It performs tasks associated with automatic music classification. The jAudio component extracts low-level features from audio recordings in wav, mp3 and aiff format using autonomous classification engine (ACE) XML format, which are stored as separate files and can be used to prepare a dataset for training and testing classifiers. With a total of 26 distinctive features implemented in jAudio, 15 most promising features were selected, see Table I. Both the average (μ) and standard deviation (σ) of each feature were extracted.

To analyze the songs from the audio content, a set of audio features was extracted based on the temporal and spectral representations of the audio signal. Feature extraction was performed to reduce the large amount of data in audio files into analytically meaningful representations [25] that can be used for statistical analysis. It was suggested in [26] that no single feature emerging as the dominant feature to explain the classification of emotion, it is important to understand what each feature describes and how it affects the human listener's perception in general. Table II provides a summary of some acoustic features commonly used in emotion recognition tasks. These features form the foundation for analyzing the emotional content of audio signals and serve as essential inputs for machine learning models in music emotion research.

Due to the absence of standard criteria for selecting musical features [27], various audio features, including rhythmic, timbral and harmonic features (spectral features), have been considered [14]. Rhythmic features are often obtained by analyzing periodic variation through beat histograms [28]. Timbre, describing the sound quality of music, includes both spectral and temporal features that help differentiate between various instruments, vocalizations, and sound textures in a musical composition.

B. Correlation with Emotional Responses

A correlation-based analysis was conducted to examine the relationships between low-level audio features and perceived emotions. Pearson correlation coefficient (r) were computed between each feature and each emotion, and only associations with $p < 0.1$ were retained for interpretation. A correlation matrix was then constructed to summarize these relationships. Positive correlations indicate that higher feature value are associated with strong emotional ratings, whereas negative correlation indicate the opposite. Emotional ratings were treated as binary variables, each indicating whether a song was rated with a given emotion. Consequently, the dataset contains one row per song -rating pair, totalling 3031 ratings.

Table III shows a sample of the extracted features and their correlations to energizing and relaxing emotions.

TABLE I. A BRIEF DESCRIPTION OF THE SELECTED LOW-LEVEL FEATURES AND THEIR DIMENSIONS.

Acoustic features		Description	Dimensions
1.	Beat Sum	The sum of all entries in the beat histogram. This is a good measure of the importance of regular beats in a signal	1
2.	Compactness	A measure of the noisiness of a signal. Found by comparing the components of a window's magnitude spectrum with the magnitude spectrum of its neighboring windows.	1
3.	Fraction of Low Energy Windows	The fraction of the last 100 windows that has an RMS less than the mean RMS. This can indicate how much of a signal is quite relative to the rest of the signal.	1
4.	Linear Prediction Coefficients (LPC)	LPC is calculated using autocorrelation and Levinson-Durbin recursion.	10
5.	Method of Moments	A feature vector which consists of the first five statistical moments of the magnitude spectrum and provides a compact statistical representation of the magnitude spectrum.	5
6.	MFCC	It represents the Mel-frequency energy distribution of an audio signal and can identify the most important frequencies of the signal while being robust to changes in loudness and sound characteristics[29]. For the Mel-frequency scale, which is influenced by perceptions, a discrete cosine transformation is utilized. 13 coefficients are used.	13
7.	Peak-based Spectral Smoothness	It is calculated from partials, not frequency bins. It is implemented based on coefficient and spectral smoothness algorithm [30].	1
8.	Root Mean Square	It determines the average power of an audio signal. It shows the signal's average volume and can describe its loudness level [31].	1
9.	Spectral Centroid	Shows the average frequency at which the energy of a sound signal is centered. It can be used to estimate how bright or dark the sound is [32].	1
10.	Spectral Flux	A measure of the amount of spectral change in a signal. Found by calculating the change in the magnitude spectrum from frame to frame.	1
11.	Spectral Rolloff	It shows the frequency level at which 85% of the energy is contained in the signal. Further, it measures the skewness of the power spectrum and can identify the frequency ranges that are most strongly represented in the signal [33].	1
12.	Spectral Variability	The standard deviation of the bin values of the magnitude spectrum. This is a measure of the variance of a signal's magnitude spectrum.	1
13.	Strength of Strongest Beat	How strong the strongest beat in the beat histogram is compared to other potential beats.	1
14.	Strongest Beat	The strongest beat in a signal, in beats per minute, is found by finding the strongest bin in the beat histogram.	1
15.	Zero Crossing Rate	The number of times the waveform changed sign. An indication of frequency as well as noisiness.	1

TABLE II. A SUMMARY OF SOME SELECTED MUSICAL ELEMENTS AND THEIR RELATIVE ACOUSTIC FEATURES

Musical feature	Acoustic features associated
Rhythmic	Strongest beat, strength of strongest beat
Timbre	Spectral contrast, Spectral centroid, spectral Rolloff, MFCCs, Zero Crossings, Linear Prediction Coefficients (LPC), Peak Based Spectral Smoothness
Dynamic	Compactness, RMS, Fraction of Low Energy Windows

TABLE III. EXAMPLE OF SAMPLE DATA COLLECTED

User ID	Genre	Song	Energizing	Relaxing	Spectral centroid	RMS	Compactness
1	Pop	Beyond Late	-	Yes	14.6	0.131	1691
2	Pop	To climb	-	-	16.9	0.185	1633
4	Pop	Persephone	-	-	9.5	0.212	1599
5	Pop	Sanctuary	-	-	20.6	0.133	1654
102	Classic	Trio sonata in c major	Yes	-	25.0	0.892	1828
104	Classic	Donata for violoncello	Yes	-	11.7	0.069	1680
105	Classic	Demons	-	-	16.6	0.049	1735
304	Electric	The Spirit	-	-	28.0	0.073	1639
306	Electric	Beneath the skin	-	-	13.2	0.857	1709
307	Electric	Known Bugs	-	-	15.6	0.215	1493

IV. RESULTS

The correlation analysis revealed a consistent pattern across feature groups, although generally small in magnitude ($r \approx 0.03$–0.10). Detailed results are visualized in Fig. 2. The numbers in the heatmap represent correlation coefficients, and only values with $p < 0.1$ are shown. Correlations with $r > 0.045$ (approximately), corresponding to $p < 0.01$, indicate strong evidence and smaller values corresponding to $p < 0.1$ weak evidence.

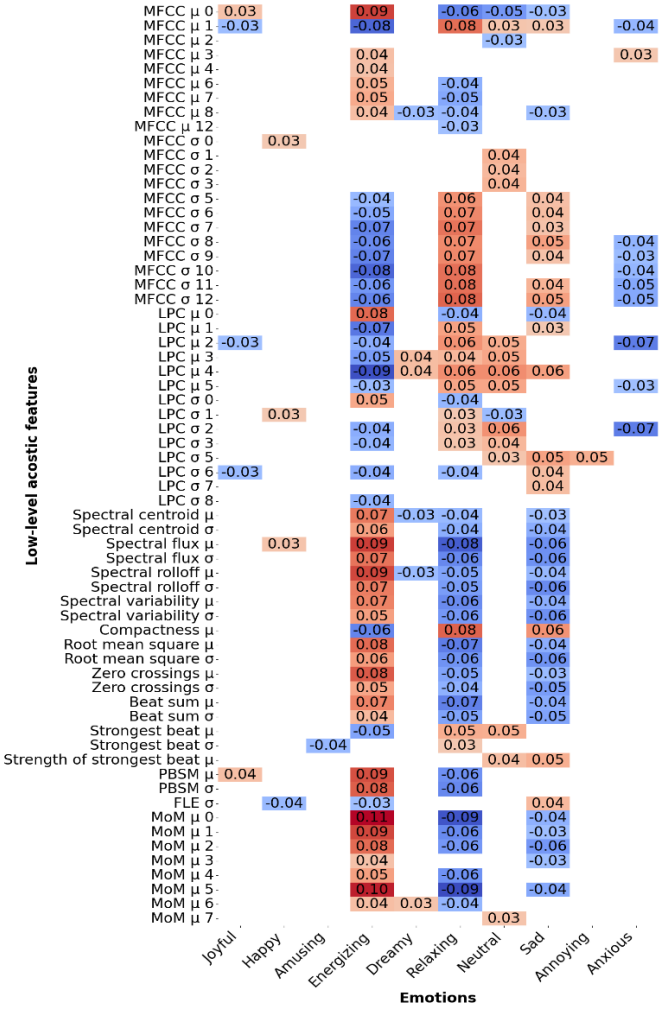

Fig. 2. Heatmap showing the correlation between low-level acoustic features and emotions. Only features with statistically significant correlations ($p<0.1$) are included. 'μ' denotes the average and denotes 'σ' the standard deviation, PBSM denotes Peak Based Spectral Smoothness, MoM denotes Method of Moments, LPC denotes Linear Prediction Coefficients, FLE denotes Fraction Of Low Energy Windows, MFCC denotes Mel-frequency cepstral coefficients.

Confidence intervals quantify the uncertainty of correlation estimates. We calculate 95% confidence intervals, meaning that if the sample procedure (n = 3031) were repeated many times, about 95% of the interval would contain the true correlation. Smaller observed correlation yield intervals that included the null value (r = 0), whereas stronger correlations are more likely to exclude it. Because the sample size is fixed, correlations of the same magnitude have identical confidence intervals. For example, The MFFC average (0) correlated with energizing emotion at r = 0.09 [0.06, 0.13], whereas MFCC standard deviation (0) correlated at r = 0.03 [-0.01, 0.07] with the happy emotion. For r values between 0.03 and 0.09, the confidence intervals would be somewhere between the examples mentioned.

Several MFCC averages and standard deviations show a correlation between energizing and relaxing. The MFCC 0 (μ) showed a positive correlation with energizing (r = 0.09) and a negative correlation with relaxing (r = -0.06). These suggest that songs with higher MFCC0 tend to be perceived as energizing and low as relaxing. While all correlations remain relatively small, the strength varies, with some coefficients showing notably stronger association than others. Thus, MFCC measures spectral envelope details, which generally relate positively to energizing, negatively to sad and relaxing. These findings indicate that emotional dimensions can be predicted by the MFCC characteristics, with at least $p < 0.1$, for many of its $p < 0.01$. The linear predictive coefficient (LPC) features also exhibited some distinct patterns. For instance, LPC(μ), including LPC0, showed positive correlation with energizing, but negative correlation with relaxing and sad. However, LPC1-5 were negatively correlated with energizing but positively correlated with relaxing, similarly to MFCC. Nonetheless, a few LPC (σ) such as LPC2, and LPC3 have a significant positive correlation with neutral and relaxing. Compared to MFCC, the LPC results show that energizing has higher energy variation features, while relaxing and sad have lower values.

Similar patterns were observed in dynamic features, including RMS energy, zero-crossing rate and beat-related features. RMS (μ) had correlation value r = 0.08 with energizing and r = -0.07 with relaxing. Zero-crossing (μ) had r = 0.08 with energizing, while sad had a negative correlation r = -0.03. These findings suggest that signals that are louder, percussive or have a rapid oscillating sound tend to be perceived as more energetic. Spectral features were prominently present and positively correlated with energizing, while both relaxing and sad were negatively correlated. This is substantiated in [34], [35], [36] reporting that spectral features negatively correlates with sadness. Songs that were rated more energetic and less relaxing exhibited rapid spectral changes (high spectral flux), higher spectral centroid or higher spectral roll-off. These findings are in line with the notion that energetic music tends to have higher frequency content and spectral variations. This is substantiated in [37] by reporting that spectral features are strongly associated with perceptions of energizing and arousal in music, and by [5], who emphasized that negative emotions such as sadness and depression negatively correlate with spectral characteristics.

Importantly, a distinct pattern emerged; correlation between sad and relaxing was often similar in the same direction and significance, while energizing consistently exhibits the opposite trend. Compared to the other emotions, energizing had the most significant correlations (cases of $p < 0.1$), followed by relaxing, while annoying exhibited only two significant correlations. The findings also revealed that compactness positively correlated with relaxing opposite to energizing. This indicates that relaxing emotional states are connected to spectral patterns that are more uniform and stable since sound is less noisy, whereas energetic states are associated with noisier and irregular spectral patterns.

Positive emotion, such as joyful, however, exhibited a weak correlation with peak-based spectral smoothness (avg), whereas happy exhibited low association with Spectral flux(μ), LPC(σ), and MFCC(σ). These findings substantiate previous study by [38] suggesting that positive emotions do not always associate with low-level acoustic variability, but rather often rely on prosodic and melodic cues. Contrarily, peak-based spectral smoothness and method of moments features showed significant positive correlations with energizing, whereas relaxing and sad were less significant. These results align with [39], who emphasized the role of spectral moments irregularities and higher moments in conveying perceived energy in music. Specific low-level audio features contribute to mid-level musical features, which are associated with distinct emotional responses. For example, spectral centroid, MFCCs, and LPCs contribute to timbre brightness and texture. These are key features for eliciting feelings of joy. Similarly, Spectral flux and beat sum, indicators of rhythmic activity, reinforce brightness. This influences happiness. Energizing or anxious emotions are linked to features that reflect dynamics (loudness), tempo, and sharpness (e.g., RMS and MFCC variability), while emotions such as sad or relaxing are associated with smoothness and low timbre, and they are derived from MFCCs, LPCs, and RMS energy. Table IV shows the relationship between acoustic features and human emotional perception in music.

TABLE IV. THE RELATIONSHIP BETWEEN LOW-LEVEL FEATURES, MID-LEVEL FEATURES AND EMOTIONS

Low-Level	Mid-Level Features	Emotion
Spectral centroid, MFCCs, LPCs, PBSM(μ)	Timbre (brightness) and texture	Joyful
Spectral flux, MFCCs, LPC(σ) 1	Timbre (brightness)	Happy
MFCCs (especially MFCC 3, MFCC 12)	Texture and timbre (variation)	Amusing
RMS, Spectral flux, MFCCs, Beat sum	Timbre (brightness), tempo (steady) and dynamics	Energizing
LPCs, Spectral centroid, Spectral rolloff, MFCCs, MoM	Timbre (variation and smoothness)	Dreamy
RMS, MFCCs, Spectral centroid, LPCs, Beat sum	Timbre (smoothness) and texture	Relaxing
MFCCs, LPCs, Strongest beat, MoM, Strength of strongest beat	Rhythm (balanced energy) and dynamics	Neutral
Spectral flux, RMS, MFCCs, LPCs	Timbre and texture	Sad
MFCC 7, LPCs	Timbre (sharpness and variation)	Annoying
MFCC(σ), LPCs	Timbre (sharpness, and high Variation)	Anxious

V. Discussions

The findings demonstrate that low-level acoustic features and emotional responses to music have a complex and context-sensitive relationship. It clearly shows the validity of how the features perform in a correlation analysis. The findings revealed that higher compactness positively correlates with relaxing but negatively with energizing. This indicates that the perception of calmness may be enhanced by uniform spectral features, whereas energetic experience has less compact and more dynamically varying spectral content.

Surprisingly, positive emotions such as joyful, happy and amusing did not have a significant correlation with spectral features. This observation highlights the importance of including high-level musical attributes in the modelling of positive emotional states. Conversely, energizing is strongly correlated to peak-based spectral features' smoothness and methods of moments. The findings suggest that the perception of energetic qualities in music often depends on spectral complexity and temporal fluctuation in acoustic features.

Additionally, features with higher frequency content, energy or increased spectral features were found to correlate positively with energizing and negatively with relaxing and sad. Similar patterns were also found in the analysis of MFCC, LPC and other spectral features. Thus, these findings indicate that spectral and cepstral features are reliable indicators of emotions related to arousal, while also highlighting their limitations in capturing low-arousal states such as relaxing and sadness. The spectral features showed a weak correlation between relaxing and sad, which indicates that smoothness and compactness are effective in conveying lower arousal (including sadness and relaxing) states than irregular spectral dynamics. The results demonstrate that emotional perception in music occurs due to a complex interaction between different levels of acoustic information. Low-level spectral features are effective in distinguishing arousal-related emotions, such as energizing and relaxing, whereas positive emotions, such as joyful and happy, appear to be influenced by prosodic and melodic cues beyond spectral features.

VI. Conclusion

We analyzed the relationship between low-level acoustic features extracted from song excerpts and the related emotional responses. The results indicate that the complexity of musical emotion cannot be fully captured by a single feature, hence there is a need to integrate multiple acoustic features, such as cepstral, spectral and prosodic features, to significantly enhance the recognition of emotion in music. Multimodal approaches can improve emotion recognition accuracy and further provide a comprehensive understanding of how these emotions are encoded in perceived in music. Future research should explore hybrid models that combine low-level acoustic descriptors with high-level musical structures to advance research in the field.

Authors' Contribution

Ideal of article was by PF and AW. AW did most of the writing. All authors contributed to the writing. SS did majority of the technical part, such as the correlation calculation. AW and SS did analysis. PF supervised the entire research.

References

[1] G. Tzanetakis, MARSYAS-0.2: A case study in implementing music information retrieval systems, no. May. 2007. doi: 10.4018/978-1-59904-663-1.ch002.

[2] J. C. Wang, Y. H. Yang, H. M. Wang, and S. K. Jeng, "Modeling the affective content of music with a Gaussian mixture model," IEEE Trans Affect Comput, vol. 6, no. 1, pp. 56–68, 2015, doi: 10.1109/TAFFC.2015.2397457.

[3] J. A. Russell, "A circumplex model of affect," J Pers Soc Psychol, vol. 39, no. 6, pp. 1161–1178, 1980, doi: 10.1037/h0077714.

[4] H. U. R. Siddiqui et al., "Emotion classification using temporal and spectral features from IR-UWB-based respiration data," Multimed Tools Appl, vol. 82, no. 12, pp. 18565–18583, May 2023, doi: 10.1007/S11042-022-14091-5.

[5] R. Panda, R. Malheiro, and R. P. Paiva, "Audio Features for Music Emotion Recognition: A Survey," IEEE Trans Affect Comput, vol. 14, no. 1, pp. 68–88, 2020, doi: 10.1109/TAFFC.2020.3032373.

[6] D. Bitouk, R. Verma, and A. Nenkova, "Class-Level Spectral Features for Emotion Recognition," Speech Commun, vol. 52, no. 7–8, p. 613, 2010, doi: 10.1016/J.SPECOM.2010.02.010.

[7] S. Rajesh and N. J. Nalini, "Musical instrument emotion recognition using deep recurrent neural network," Procedia Comput Sci, vol. 167, no. Iccids 2019, pp. 16–25, 2020, doi: 10.1016/j.procs.2020.03.178.

[8] H. Nordström and P. Laukka, "The time course of emotion recognition in speech and music," J Acoust Soc Am, vol. 145, no. 5, pp. 3058–3074, 2019, doi: 10.1121/1.5108601.

[9] A. Rodà, S. Canazza, and G. De Poli, "Clustering affective qualities of classical music: Beyond the valence-arousal plane," IEEE Trans Affect Comput, vol. 5, no. 4, pp. 364–376, 2014, doi: 10.1109/TAFFC.2014.2343222.

[10] K. Trochidis, C. Delbé, and E. Bigand, "Investigation of the relationships between audio features and induced emotions in Contemporary Western music," in Proceedings of the 8th Sound and Music Computing Conference, SMC 2011, 2011.

[11] M. Schedl, E. Gomez, E. S. Trent, M. Tkalcic, H. Eghbal-Zadeh, and A. Martorell, "On the interrelation between listener characteristics and the perception of emotions in classical orchestra music," IEEE Trans Affect Comput, vol. 9, no. 4, pp. 507–525, 2018, doi: 10.1109/TAFFC.2017.2663421.

[12] A. Wiafe, S. Sieranoja, A. Bhuiyan, and P. Fränti, "EURASIP Journal on Audio, Speech, and Music Processing Emotional response to music: the Emotify + dataset," EURASIP Journal on Audio, vol. 2025, p. 31, 2025, doi: 10.1186/s13636-025-00419-0.

[13] S. Rajesh and N. J. Nalini, "Musical instrument emotion recognition using deep recurrent neural network," Procedia Comput Sci, vol. 167, pp. 16–25, Jan. 2020, doi: 10.1016/J.PROCS.2020.03.178.

[14] M. Mokhsin, N. Rosli, N. A. Manaf, and H. A. Halim, "Music Emotion Classification (MEC): Exploiting Vocal and Instrumental Sound Features Mudiana," in International Conference on Soft Computing and Data Mining., 2014. doi: 10.1007/978-3-319-07692-8.

[15] X. Liu, Y. Xu, K. Alter, and J. Tuomainen, "Emotional connotations of musical instrument timbre in comparison with emotional speech prosody: Evidence from acoustics and event-related potentials," Front Psychol, vol. 9, no. MAY, pp. 1–10, 2018, doi: 10.3389/fpsyg.2018.00737.

[16] N. Di Stefano and C. Spence, "Roughness perception: A multisensory/crossmodal perspective," Atten Percept Psychophys, vol. 84, no. 7, p. 2087, Oct. 2022, doi: 10.3758/S13414-022-02550-Y.

[17] T. Eerola, R. Ferrer, and V. Alluri, "Timbre and affect dimensions: Evidence from affect and similarity ratings and acoustic correlates of isolated instrument sounds," Music Percept, vol. 30, no. 1, pp. 49–70, Sep. 2012, doi: 10.1525/MP.2012.30.1.49.

[18] G. Athanasopoulos, T. Eerola, I. Lahdelma, and M. Kaliakatsos-Papakostas, "Harmonic organisation conveys both universal and culture-specific cues for emotional expression in music," PLoS One, vol. 16, no. 1, pp. 1–17, 2021, doi: 10.1371/journal.pone.0244964.

[19] Y. S. Kwon, J. Lee, and S. Lee, "The impact of background music on film audience's attentional processes: Electroencephalography alpha-rhythm and event-related potential analyses," Front Psychol, vol. 13, p. 933497, Nov. 2022, doi: 10.3389/FPSYG.2022.933497/BIBTEX.

[20] E. Pampalk, "Computational Models of Music Similarity and their Application in Music Information Retrieval," 2006.

[21] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," in IEEE Transactions on Multimedia, 2011, pp. 303–319. doi: 10.1109/TMM.2010.2098858.

[22] A. Aljanaki, F. Wiering, and R. C. Veltkamp, "Studying emotion induced by music through a crowdsourcing game," Inf Process Manag, vol. 52, no. 1, pp. 115–128, 2016, doi: 10.1016/j.ipm.2015.03.004.

[23] E. Schubert, "The influence of emotion, locus of emotion and familiarity upon preference in music," Psychol Music, vol. 35, no. 3, pp. 499–515, 2007, doi: 10.1177/0305735607072657.

[24] C. McKay and I. Fujinaga, "JMIR: Tools for automatic music classification," in Proceedings of the 2009 International Computer Music Conference, ICMC 2009, 2009, pp. 65–68.

[25] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," Multimed Syst, vol. 24, no. 4, pp. 365–389, 2018, doi: 10.1007/s00530-017-0559-4.

[26] Y. E. Kim et al., "Music emotion recognition: A state of the art review," Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010, no. Ismir, pp. 255–266, 2010.

[27] X. Liu, Q. Chen, X. Wu, Y. Liu, and Y. Liu, "CNN based music emotion classification," 2017.

[28] A. Lykartsis and A. Lerch, "Beat histogram feature for rhythm -based musical genre classification using multiple novelty functions," in Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15), Trondheim, Norway, Nov 30- Dec 3, 2015, Norway, 2015.

[29] N. Sato and O. Yasunari, "Emotion recognition using mel-frequency cepstral coefficients," Information and Media Technologies, vol. 2, no. 3, pp. 835–848, 2007, Accessed: Feb. 06, 2025. [Online]. Available: https://www.jstage.jst.go.jp/article/imt/2/3/2_3_835/_article/-char/ja/

[30] S. McAdams, "Perspectives on the Contribution of Timbre to Musical Structure on JSTOR," Computer Music Journal, vol. 23, no. 3, pp. 85–102, 1999, Accessed: Feb. 06, 2025. [Online]. Available: https://www.jstor.org/stable/3681242

[31] M. Chourasia, S. Haral, S. Bhatkar, and S. Kulkarni, "Emotion Recognition from Speech Signal Using Deep Learning," in Lecture Notes on Data Engineering and Communications Technologies, Springer, Singapore, Feb. 2021, pp. 471–481. doi: 10.1007/978-981-15-9509-7_39.

[32] A. Klapuri and M. Davy, Signal processing methods for music transcription. 2007. Accessed: Feb. 06, 2025. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=AF30yR41GIAC&oi=fnd&pg=PR9&ots=OptccfW8Fz&sig=1DRxRWxS8TrG8HwqFcFX1mkmFp4

[33] P. Sandhya, V. Spoorthy, S. G. Koolagudi, and N. V. Sobhana, "Spectral features for emotional speaker recognition," in Third international conference on ad 2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC), Dec. 2020. doi: 10.1109/ICAECC50550.2020.9339502.

[34] B. Wu, A. Horner, and C. Lee, "Musical Timbre and Emotion: The Identification of Salient Timbral Features in Sustained Musical Instrument Tones Equalized in Attack Time and Spectral Centroid," in Proceedings ICMC|SMC|2014, Athens,Greece, Sep. 2014.

[35] E. Brattico et al., "A functional MRI study of happy and sad emotions in music with and without lyrics," Front Psychol, vol. 2, no. DEC, 2011, doi: 10.3389/fpsyg.2011.00308.

[36] Y. Song, M. Pearce, and S. Dixon, "Evaluation of Musical Features for Emotion Classification," 2012. [Online]. Available: http://www.music-ir.org/mirex/wiki/MIREX

[37] K. Siedenburg, I. Fujinaga, and S. McAdams, "A Comparison of Approaches to Timbre Descriptors in Music Information Retrieval and Music Psychology," J New Music Res, vol. 45, no. 1, pp. 27–41, Jan. 2016, doi: 10.1080/09298215.2015.1132737.

[38] P. N. Juslin and P. Laukka, "Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?," Psychol Bull, vol. 129, no. 5, pp. 770–814, Sep. 2003, doi: 10.1037/0033-2909.129.5.770.

[39] S. McAdams, C. Douglas, and N. N. Vempala, "Perception and modeling of affective qualities of musical instrument sounds across pitch registers," Front Psychol, vol. 8, no. FEB, p. 242203, Feb. 2017, doi: 10.3389/FPSYG.2017.00153/BIBTEX.