Using audio FX to alter music emotion

Stelios Katsis
Dept. of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
el20139@mail.ntua.gr

***Abstract*—In this paper, we argue that emotion is a fundamental aspect of music. It accompanies every song and gives a certain identity to it. This identity, however, is malleable and can be altered through the use of audio effects. That is exactly what we will focus on. After understanding what gives music a certain emotion through analyzing the fundamental aspects of music, with the use of certain tools and datasets, we will try to extract basic patterns that accompany audio effects. This is a still underdeveloped part of Music Information Retrieval and with this research we will try to set the basis for future research. The implementation of the code is located in GitHub: https://github.com/stelioskt/audio-fx**

***Index Terms*—music, emotion, FX, MER, valence/arousal, MERT**

I. Introduction

Music has long been recognized as a universal language, capable of evoking and expressing a wide range of emotions. Its unique ability to transcend cultural and linguistic boundaries has made it a central focus of musicology and psychology. With the rise of artificial intelligence and advanced computational methods, the field of Music Emotion Recognition (MER) has emerged, enabling researchers to classify and predict the emotional content of music with remarkable precision [1]. MER leverages machine learning models to analyze musical features such as melody, harmony, and rhythm, providing deeper insights into the emotional impact of music on listeners.

Despite these advancements, the role of audio effects (FX) in shaping the emotional perception of music remains underexplored. Audio effects like reverb, distortion, and tempo manipulation are widely used in music production to enhance aesthetic appeal or create specific moods. However, the degree to which these effects influence the listener's emotional experience—and how such changes can be quantified and modeled—poses an open question in the fields of Music Information Retrieval (MIR) and MER [4]. Existing work has focused mainly on analyzing music emotions based on static features [2], [4], leaving a significant gap in understanding the dynamic interplay between audio manipulation and perceived emotion.

This article addresses this gap by exploring how audio effects can alter the emotional perception of music. Attempts are made to analyze existing music datasets, employ state-of-the-art MER models, and experiment with various audio FX to study their impact on emotional classification [1]. The findings of this research have significant implications for music production, personalized recommendation systems, and therapeutic applications, where understanding and manipulating emotional perception can create deeper and more meaningful listener experiences.

By providing insights into the emotional effects of audio manipulation, this study contributes to the broader understanding of how music can be tailored to evoke specific emotional responses. As the interplay between audio effects and emotional perception remains an emerging area of inquiry, this work seeks to expand the existing literature and inspire further research into this compelling intersection of technology and human experience.

II. Audio, Emotion and FX: An Overview

This section provides a foundational overview of the core concepts relevant to the study. It begins with an analysis of audio and its principal features, including pitch, harmony, tempo, and timbre, as well as their influence on emotional perception. Subsequently, the section examines the fundamentals of emotions, exploring key emotion theories and various classification models. Finally, it focuses on audio effects, discussing their primary characteristics and their impact on the emotional quality of audio and music.

A. Introduction to Audio

Music is a powerful form of sound, deeply embedded in human culture and emotion. Its core elements—pitch, harmony, rhythm, timbre, and dynamics—shape emotional expression, allowing music to evoke a wide range of feelings, from joy to melancholy. These elements form the foundation for musical communication, enabling composers and performers to create diverse emotional landscapes.

1) Pitch: Pitch refers to the frequency of a sound, determining whether it is high or low. Higher pitches, such as those of a flute, often evoke excitement or brightness, while lower pitches, such as those of a cello, suggest calmness or melancholy. This emotional impact comes from the way humans perceive different frequencies [5].

2) Harmony: Harmony refers to the combination of notes that form chords, which is crucial as it sets the emotional tone of music. Major harmonies are associated with happiness, while minor harmonies often evoke sadness. Dissonance creates tension, while consonance offers stability [5].

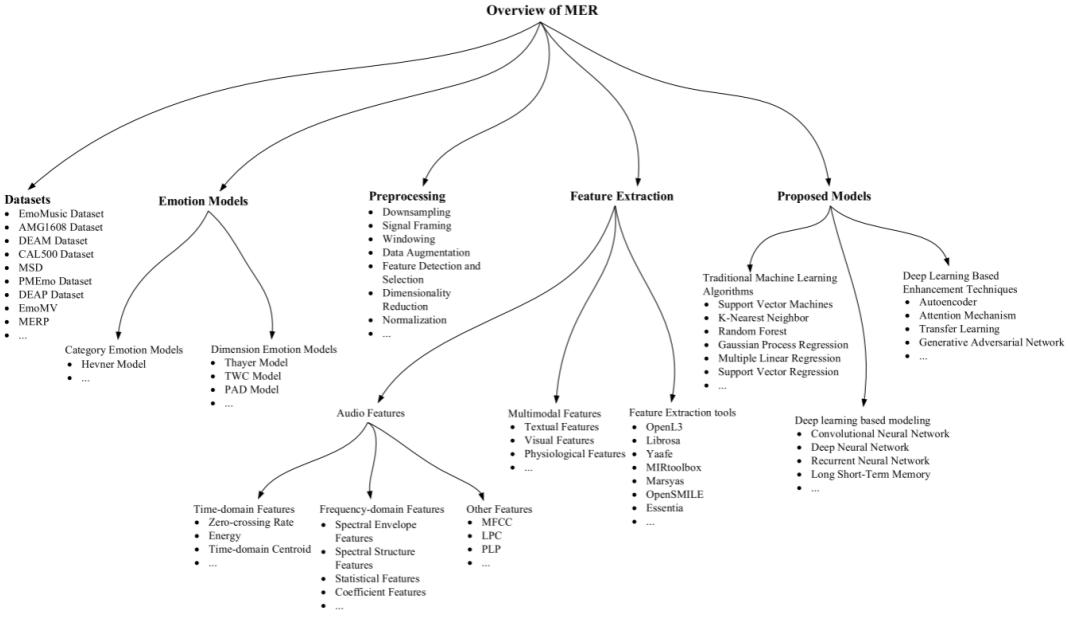

Fig. 1. Basic concepts of Music Emotion Recognition [7]

3) Rhythm: Rhythm structures music in time, with elements like tempo and syncopation influencing emotional tone. Faster tempos suggest energy and excitement, while slower ones convey calm or sadness. Syncopation adds unpredictability, while regular rhythms create stability. Rhythm's emotional impact comes from how it aligns with or disrupts listener expectations [5].

4) Timbre: Timbre is the unique quality of a sound that distinguishes different instruments, even at the same pitch. Smooth timbres, like those of strings, evoke warmth, while harsh timbres, like distorted guitars, suggest power or aggression. Timbre shapes the texture of music and enhances its emotional expression [5].

5) Dynamics: Dynamics refer to changes in loudness, influencing the emotional intensity of music. Crescendos build excitement, while decrescendos foster calmness. Loud music can convey power or anger, while softer dynamics evoke sadness or introspection. The emotional effects of dynamics are tied to the listener's physiological responses [5].

A brief summary of the previous audio features, along with the emotions they provoke are depicted in "TABLE I"

TABLE I
FEATURES AND THEIR EMOTIONAL EFFECTS

Feature	Emotion	Description
High Pitch	Happiness, Excitement	Bright and energetic tones induce joy and enthusiasm.
Low Pitch	Sadness, Calmness	Darker tones bring depth and a sense of gravity or relaxation.
Major Harmony	Joy, Warmth	Creates feelings of positivity and stability.
Minor Harmony	Sadness, Melancholy	Evokes reflective or sorrowful moods.
Fast Tempo	Energy, Happiness	Accelerated rhythms convey urgency, liveliness, or celebration.
Slow Tempo	Calmness, Sadness	Relaxed pacing fosters serenity or introspection.
Harsh Timbres	Anger, Power	Aggressive sounds induce feelings of tension or dominance.
Smooth Timbres	Tenderness, Nostalgia	Soft tones provide emotional intimacy or longing.
Loud Dynamics	Excitement, Anger	Amplified sound levels evoke power or intensity.
Soft Dynamics	Sadness, Calmness	Quiet levels induce reflection or peacefulness.

B. Understanding Emotions

Emotion is a central aspect of human culture, influencing decision-making, communication, and social interaction. Within the context of music, emotions play an equally vital role, shaping how listeners interpret and connect with compositions. Music's ability to evoke emotions is a defining feature of its universal and timeless appeal. For centuries, theorists and scientists have sought to understand the mechanisms by which music communicates emotions, whether through its intrinsic structural properties or its ability to elicit psychological responses in listeners. Modern advances in Music Emotion Recognition (MER) have enabled researchers to better understand and classify these emotional impacts using computational tools. To achieve this, MER is structured based on psychological theories of emotion, emotion classification models, and an understanding of the interaction between musical features and emotional responses.

1) Emotion Theories in Music: Through centuries of research and experimentation, two theories emerged as better structured and more appealing:

a) Expression Theory: : Expression theory suggests that music inherently communicates emotions through its structural components, such as melody, harmony, rhythm, and dynamics [7]. These structural elements imbue music with emotional meaning, which listeners perceive directly, independent of their personal emotional state. For instance, a slow, minor-key piece is often perceived as sad because of its musical structure rather than the listener's mood. This theory emphasizes the inherent and universal qualities of music's emotional language, making it an objective framework for understanding musical emotions. It is particularly relevant in explaining why similar emotional interpretations are observed across listeners from diverse cultural backgrounds.

b) Arousal Theory: In contrast, arousal theory supports that music acts as a stimulus, eliciting emotional responses within the listener [7]. According to this theory, emotions arise from physiological and psychological reactions triggered by musical features. For example, a fast, upbeat song with high energy may evoke happiness or excitement by stimulating the listener's arousal systems, while a slow, low-pitched melody might induce calmness or introspection. This theory highlights the subjective nature of musical emotions, as listeners' responses are influenced by individual experiences.

2) Emotion Tags/Labels: To analyze emotions in music systematically, researchers have developed several models that categorize or map emotions. These models provide structured frameworks for understanding how music conveys and elicits emotional responses:

a) Discrete Emotion Model: This model categorizes emotions into distinct classes, such as happiness, sadness, anger, and fear. It simplifies the complex spectrum of emotions into identifiable categories, enabling straightforward classification [3]. For instance, Hevner's Adjective Circle [7] (Fig. 2) organizes emotions into clusters of descriptive terms (e.g., joyful, solemn, sad) associated with specific musical characteristics like tempo and harmony. However, this approach may not capture the intensity or subtle variations within emotional experiences, as it confines emotions to fixed categories.

b) Dimensional Emotion Model: This model represents emotions along continuous dimensions, most commonly valence (positive to negative) and arousal (high energy to low energy) [2]. By mapping emotions in the valence-arousal space, the dimensional model allows for a more nuanced understanding of emotional content. For example, a high-valence, high-arousal piece might evoke joy or excitement, while a low-valence, low-arousal piece might convey sadness or melancholy [3]. For instance, Thayer's Two-Dimensional Model of Mood [7] (Fig. 3) combines arousal (activation) and valence (pleasantness) to describe mood states. For instance, a highly aroused and positively valenced piece might represent feelings of elation, while low arousal and negative valence might correspond to depression. This model is particularly valuable for capturing the fluidity and complexity of musical emotions, but it can introduce variability due to subjective interpretation.

c) Hybrid Models: Hybrid models combine the strengths of discrete and dimensional approaches to provide a more comprehensive analysis of emotions. These models aim to classify emotions into categories while also accounting for their intensity and subtle variations along continuous dimensions. As an example, The Geneva Emotion Wheel [7] (Fig. 4) integrates categorical emotions (e.g., happiness, sadness) with intensity ratings, offering a multidimensional representation of emotional experiences. In music, such a model might label a piece as "joyful" while specifying its valence-arousal coordinates to indicate its energetic or subdued nature.

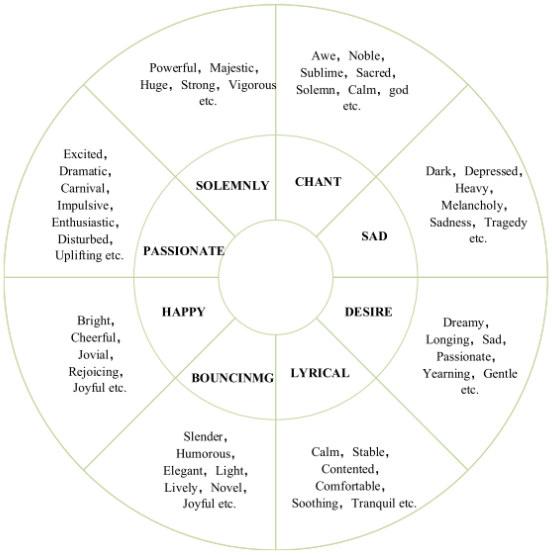

Fig. 2. Hevner's adjective circle

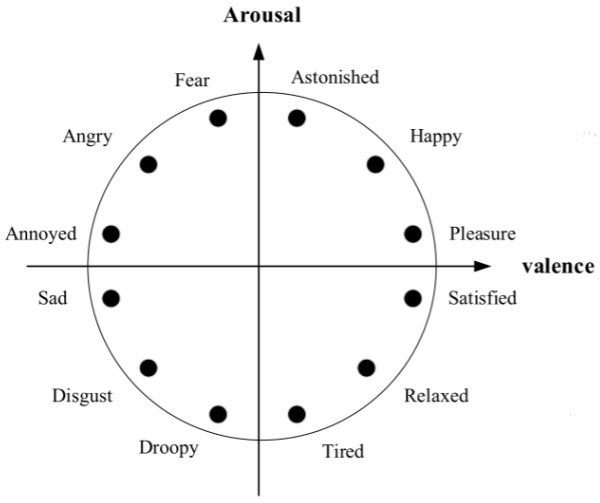

Fig. 3. Thayer's VA model

C. Audio Effects (FX)

Audio effects (FX) have become vital tools in music production, shaping the identity of compositions and amplifying their

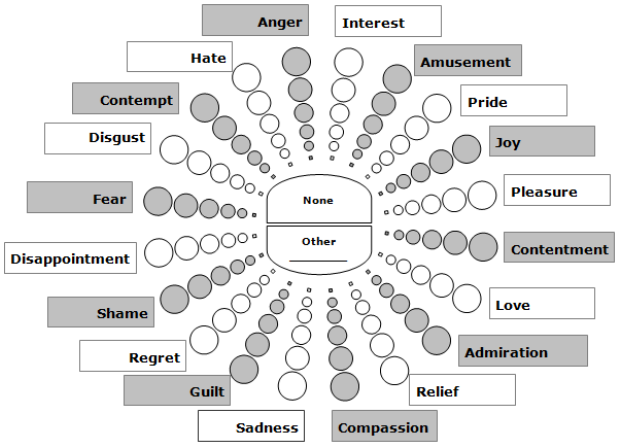

Fig. 4. Geneva Emotion Wheel

emotional resonance. By altering fundamental properties such as pitch, tempo, timbre, and dynamics, FX allow producers to craft unique soundscapes that enhance the emotional narrative of music. Historically, audio effects began with natural techniques, such as reverberation created in large halls, and evolved with technological advancements to include digital tools capable of intricate modifications. These effects are not merely aesthetic; they play a critical role in how listeners interpret and emotionally engage with music. Despite their widespread use, the systematic study of FX and their influence on emotional perception remains a relatively underexplored area in Music Emotion Recognition (MER).

1) Types of Audio Effects: Audio effects are categorized based on the specific musical features they modify. Below is an in-depth analysis of commonly used FX and their emotional implications:

a) Reverb: Reverb replicates the natural reflections of sound in a physical space, creating a sense of depth and ambiance. It can simulate environments ranging from small rooms to expansive cathedrals. Reverb adds dimensionality to music. A short, subtle reverb can make a track feel intimate and direct, while a long, lush reverb creates a sense of grandeur, mystery, or nostalgia. For instance, adding reverb to a solo piano piece can evoke feelings of melancholy or timelessness, as it mimics the acoustic properties of a concert hall.

b) Distortion: Distortion modifies the harmonic content of a sound by clipping waveforms, producing a rough, gritty texture. It is frequently used in genres such as rock, metal, and electronic music. Distortion amplifies intensity and aggression. It conveys tension, power, or rebellion, making it an effective tool for energizing a composition. For example, a distorted electric guitar solo in a rock anthem evokes raw emotion and dynamic energy. At higher intensities, distortion can border on chaotic, evoking unease or even anger.

c) Tempo and Pitch Shifting: Tempo changes adjust the speed of a track, influencing its rhythmic feel, while pitch shifting raises or lowers the perceived frequency of sound, altering its tonal height. Generally, a faster tempo enhances excitement, urgency, or joy, while a slower tempo induces calmness, introspection, or sadness. Pitch shifts have a similar effect: higher pitches brighten a track and add playfulness, while lower pitches imbue it with gravity or seriousness. For instance, a high-pitched, fast-tempo melody can evoke childlike excitement, while a low-pitched, slow-tempo bassline creates a somber or reflective mood.

d) Equalization (EQ): EQ adjusts the balance of frequency bands in a sound, enabling selective emphasis or suppression of specific frequencies. Boosting high frequencies enhances clarity and brightness, evoking cheerfulness or vitality, while emphasizing lower frequencies creates warmth, depth, or power. For example, accentuating bass frequencies in a dance track generates a sense of energy and physical movement, while reducing them in an acoustic ballad fosters intimacy and focus on midrange tonalities.

e) Delay and Echo: These effects produce repetitions of a sound at varying intervals, mimicking natural or artificial reflections to create rhythmic and spatial complexity. Delay effects often evoke playfulness, as the repeated sounds add layers to the rhythm, while echo can create a haunting, ethereal atmosphere. For instance, delay applied to a vocal line in an electronic track enhances its liveliness, while echo on a distant piano melody can evoke a sense of longing or mystery.

III. Methodology

This section will mainly focus on the basic tools that will be used in the later research, such as datasets, feature extraction tools and foundation models, along with their main features and advantages.

A. Datasets

This study utilizes carefully selected datasets to investigate the emotional impact of audio effects (FX) on music, acknowledging the limited availability of publicly accessible resources. The chosen datasets, EMOPIA, RAVDESS, and DEAM, each provide distinct advantages that enable comprehensive exploration of different aspects of Music Emotion Recognition (MER). Together, these datasets facilitate the analysis of both controlled and dynamic emotional annotations, advancing the understanding of how FX influences emotional perception.

1) EMOPIA Dataset: The EMOPIA dataset [8], [7] comprises over 1,000 monophonic piano pieces annotated with arousal and valence metrics, making it a significant resource for emotion analysis. Its monophonic focus ensures precision in isolating fundamental musical features such as melody, harmony, and tempo, allowing for a controlled investigation of the effects of FX. However, the dataset's simplicity limits its application to more complex musical styles, a gap this study addresses through comparative analysis with polyphonic datasets like RAVDESS and DEAM.

2) RAVDESS Dataset: The RAVDESS dataset [7] consists of 2,024 tracks of vocal and speech performances, annotated with discrete emotional labels such as happiness, sadness, anger, and calmness. Its inclusion of both sung and spoken

expressions offers a unique opportunity to analyze the interaction of FX with vocal emotional content. Unlike EMOPIA, RAVDESS includes polyphonic content with occasional instrumentation, adding complexity to the analysis. However, its lack of continuous temporal annotations restricts its utility in modeling dynamic emotional trajectories.

3) DEAM Dataset: The DEAM (Database for Emotional Analysis in Music) dataset [7] includes 1,802 excerpts of commercial music, each lasting 45 seconds, annotated with continuous arousal and valence scores over time. By spanning a wide range of genres such as pop, rock, and classical, DEAM enables the study of how emotions evolve dynamically within a musical piece and how FX influence these trajectories. This temporal dimension provides critical insights into the dynamic nature of emotional responses.

B. Feature Extraction

Feature extraction is a critical step in music emotion recognition, transforming raw audio signals into numerical representations that capture essential musical characteristics such as timbre, rhythm, and pitch. Tools like **MIRToolbox** [13], **Librosa** [12] and **OpenSMILE** [11] are widely used for this purpose, each offering distinct strengths. Librosa, a Python-based library, excels in extracting features such as Mel spectrograms, MFCCs, and chroma, making it ideal for capturing harmonic and rhythmic information. On the other hand, OpenSMILE is a robust toolkit designed for extracting a vast array of low-level descriptors, including prosodic, spectral, and energy-based features, particularly suited for dynamic and emotion-focused applications. These and many other tools contribute to the transformation of raw data into information suitable for an emotion classification model as input.

C. Emotion Classification Models

Emotion classification models underpin the prediction and analysis of emotional responses to music. These models utilize extracted features to classify or regress emotional attributes such as arousal, valence, or discrete emotion categories [3]. This study employs state-of-the-art (SOTA) models, each one with specific features and architecture.

1) MERT: The **Music Understanding with Large-Scale Self-Supervised Training** (MERT) model represents a pivotal innovation in Music Information Retrieval (MIR), offering a robust framework for understanding complex musical representations [6]. Leveraging self-supervised learning, MERT integrates acoustic and musical perspectives through its dual-teacher architecture. The acoustic teacher employs **Residual Vector Quantization-Variational Autoencoder** (RVQ-VAE) to capture low-level acoustic features, while the musical teacher utilizes **Constant-Q Transform** (CQT) representations to emphasize tonal and harmonic structures [9]. This dual approach enables MERT to bridge the gap between raw audio signals and musical semantics, making it highly adaptable across diverse MIR tasks [1].

MERT's training methodology is underpinned by a massive dataset of 160,000 hours of music, enabling it to capture a wide spectrum of musical diversity. Its two model scales—Base (95M parameters) and Large (330M parameters)—provide flexibility in balancing computational efficiency and performance. Despite the fact that MERT achieves state-of-the-art results in tasks like music tagging, transcription, and arousal-valence prediction, MERT's reliance on pretraining with pseudo-labels introduces challenges in fine-tuning for specialized tasks, particularly in contexts with limited labeled data. The model's open-source availability, however, facilitates further research and experimentation, allowing the MIR community to build upon its framework for applications in MER and beyond. The basic architecture of MERT's pre-training is depicted below:

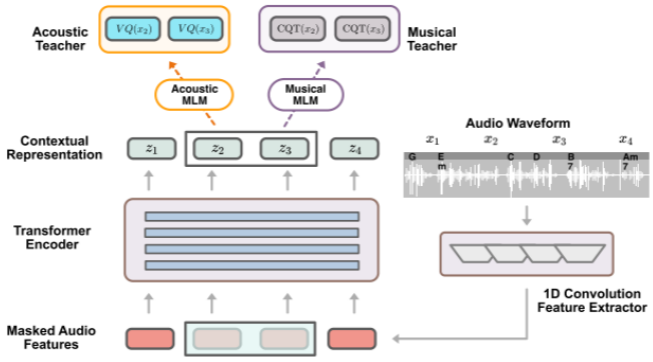

Fig. 5. Illustration of the MERT Pre-training Framework [6].

2) Other SOTA models: The **Multimodal Universal Language for Emotion** (MULE) model [9] stands out as a sophisticated deep learning architecture designed to handle the complexities of emotion classification in music. Leveraging its multimodal framework, MULE integrates diverse input modalities—such as audio, text, and symbolic data—into a unified representation space, enhancing its ability to discern subtle emotional nuances. Its use of contrastive learning techniques ensures robust alignment between different modalities, making it highly effective in tasks such as emotion recognition and multi-label tagging. The radar chart illustrates MULE's strong performance across key tasks like emotion tagging, vocal technology detection, and genre classification, demonstrating its adaptability to both content-specific and broad musical contexts.

Other state-of-the-art models shown in the radar chart also contribute significantly to music emotion classification. **Music2Vec**, for instance, focuses on generating efficient embeddings for musical data, achieving competitive performance in multi-label tagging and genre classification. Its lightweight architecture makes it well-suited for scalable applications without compromising accuracy. **CLMR** (Contrastive Learning for Music Representations) excels in capturing high-level musical attributes, as evidenced by its strong performance in source separation and key detection. Similarly, **Jukebox-based models**, inspired by OpenAI's generative framework,

demonstrate exceptional capabilities in tasks like vocal technology detection and pitch identification, showcasing their ability to process complex audio signals. These models, while distinct in their approaches, collectively highlight the diverse strategies employed by deep learning to address the challenges of emotion classification and broader music analysis tasks.

D. Research Implications

MERT's high capabilites, alongside those of other SOTA models, highlights the potential for advancing the field of MER through innovative computational approaches [1]. By leveraging the unique strengths of MERT and potentially exploring the performance of other models, there will be a significant effort correlating audio FX with emotional shifting.

The evaluation of the SOTA self-supervised models in 10 major aspects of MIR are depicted below:

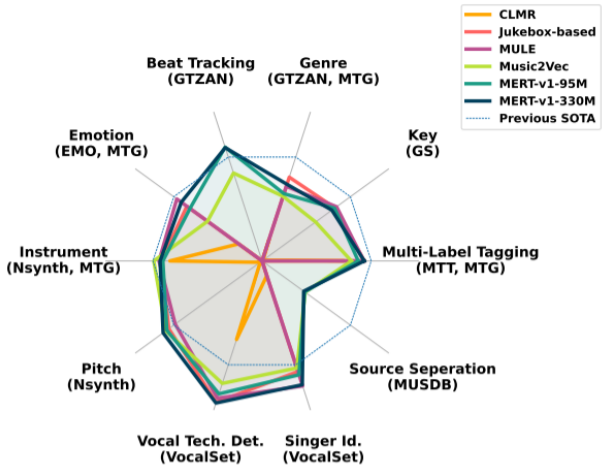

Fig. 6. Comparison of various music audio self-supervised models evaluated on a range of different tasks, as reported through the MARBLE [10] benchmark [1].

IV. Project Planning and Conclusion

A. Project Planning

The research project is structured to investigate the impact of audio effects (FX) on music emotion recognition (MER) through a systematic workflow. The initial phase involves preparing the selected datasets—EMOPIA, RAVDESS, and DEAM—by ensuring uniformity in audio formats, annotation schemas, and preprocessing techniques. This includes standardizing audio file formats (e.g., sampling rates and bit depths) and performing essential preprocessing such as trimming silence and normalizing audio levels, without applying any additional audio effects. The project will utilize **Low-Rank Adaptation** (LoRA) for fine-tuning the pretrained MERT model, enabling efficient learning of emotion-specific patterns directly from the datasets. MERT will be employed to predict continuous emotional dimensions such as arousal and valence. Comparative analyses will evaluate model performance across different datasets, providing insights into emotional patterns, which in turn will contribute in generating some general rules about the impact of audio FX in music.

B. Conclusion

This research aims to improve the understanding of how audio effects influence the emotional perception of music, using state-of-the-art tools, data sets, and modeling techniques. By systematically examining emotional dimensions across monophonic (EMOPIA), vocal (RAVDESS), and polyphonic (DEAM) contexts, the study will uncover patterns in how emotional responses are shaped by musical structures. The use of LoRA ensures efficient fine-tuning of the MERT model, facilitating accurate and scalable predictions. Findings from this study will have applications in fields such as music therapy, adaptive audio production, and personalized music recommendation systems. Future research will include the integration of additional datasets, detailed studies on specific FX, and perceptual validation with human listeners, contributing to innovative applications at the intersection of music and emotion analysis.

V. Dataset Preprocessing/Evaluation

We will start our research by analyzing the three datasets mentioned above in order to determine which of them are actually useful for our cause. Our target is to create a pandas DataFrame for each of the datasets, containing each track's name, path, and emotional category (whether it is a discrete label or a 2-D value).

A. EMOPIA

EMOPIA is a well-structured dataset with abundance of information. However, in contrast to what was previously thought, this dataset does not contain continuous VA values, rather than 4 discrete labels, depicting the four quadrants of the VA plane, meaning {Excitement, Anger, Sadness, Calmness}.

B. RAVDESS

RAVDESS is an also widely used dataset with countless applications. Nevertheless, we will not be using this dataset. This is because: a) it also contains discrete emotional labels, like the EMOPIA dataset but b) the audio tracks contain actors singing, which could be challenging for our model.

C. DEAM

The DEAM dataset is a highly informative resource for emotion analysis in music, providing continuous measurements of valence and arousal derived from evaluations by 10 distinct listeners. This dataset includes emotional annotations at two levels: (1) static values representing the overall emotional impression of an entire song, and (2) dynamic values recorded every 0.5 seconds throughout the song, beginning from the 15-second mark. The 15-second delay ensures that listeners have adequate time to familiarize themselves with the song before recording their emotional responses. For the purpose of our research, which requires a singular value for both valence and arousal, we utilize a weighted mean approach. This approach combines the dynamic mean (representing the

temporal evolution of emotion) and the static value (capturing the overall impression) to create a balanced and representative measure of the song's emotional profile.

VI. Model Evaluation

After analyzing the three datasets mentioned above, the goal is to create an accurate emotion detection model. To achieve this, we will explore two variations of the MERT model, each with unique characteristics to optimize performance.

A. MERT-v1-330M

The MERT-v1-330M model, with 330 million parameters, uses a 7-Conv, 12-Transformer network structure. It has a stride of 13.3ms and a context length of 5 seconds. Training required around 160k hours of data. This model excels at capturing subtle emotional shifts in audio.

B. MERT-v0-public

The MERT-v0-public model is a lightweight version with 95 million parameters. It uses the same network structure but operates with a stride of 20ms and a context length of 5 seconds. Training required only 0.9k hours of data, making it ideal for resource-constrained applications.

One of the challenges with the MERT model was its output, which consisted of a 1024-dimensional vector. For practical applications, this needed to be reduced to either two continuous values for regression (valence and arousal) or four discrete classes for classification (excitement, anger, sadness, and calmness). To address this, additional models were appended to the MERT output.

For the regression task, a three-layer neural network was implemented. After thorough hyperparameter tuning, separate regression models were trained for each MERT variant. Similarly, for the classification task, the XGBoost algorithm was employed. Following hyperparameter optimization, distinct classification models were created for both MERT variants.

Although the resulting performance metrics initially appeared modest, they closely align with the results reported in the original MERT study. This consistency indicates the robustness and validity of the approach, despite the inherent challenges of emotion detection in music.

The results that were extracted are presented in the following tables:

TABLE II
Performance metrics for regression predictions across MERT models.

Model	Valence			Arousal		
	MAE	**RMSE**	R^2	**MAE**	**RMSE**	R^2
MERT-v1-330M	0.1228	0.1588	0.4923	0.1009	0.1287	0.7270
MERT-v0-public	0.1270	0.1653	0.4497	0.1118	0.1429	0.6638

TABLE III
Performance metrics for classification predictions across MERT models.

Model	Precision	Recall	F1-score	Accuracy
MERT-v1-330M	0.61	0.61	0.61	0.6149
MERT-v0-public	0.68	0.68	0.68	0.6832

VII. Audio Effect Integration

Following the model selection process, audio effects were systematically incorporated into the sound to evaluate their impact on emotion detection. The `pedalboard` library was employed to implement five fundamental audio effects: Reverb, Distortion, Delay, Equalization (EQ), and Pitch Shift.

From each dataset, 40 songs were randomly selected, and five distinct levels of intensity were applied to each effect. These levels ranged from 0 (no effect applied) to 4 (maximum intensity). This structured approach allowed for a controlled assessment of the effects on the audio data.

The application of these effects resulted in a total of 1,000 modified songs per dataset, ensuring sufficient variability for thorough analysis. This setup was designed to examine how varying intensities of common audio effects influence the emotional perception of music, providing valuable insights for the development of emotion detection models.

As a subsequent step, the processed data was fed into the models to generate results for each experimental case. This procedure allowed for a detailed analysis of the model's performance across the predefined tasks of valence and arousal regression, as well as classification. The outcomes are discussed in the following sections and provide insights into the efficiency of the proposed methodologies.

VIII. Result Analysis

This section provides a detailed analysis of the findings from the application of audio effects on music, focusing on their impact on both regression (valence and arousal) and classification (emotional labels). The results reveal significant insights into how audio effects and their intensity levels influence the emotional perception of music.

A. Regression Analysis: Valence and Arousal

The regression analysis focused on evaluating how audio effects impact the continuous emotional dimensions of valence and arousal. Below, we summarize the findings for each audio effect:

1) Distortion: Distortion exhibited a significant and consistent increase in both valence and arousal as the intensity of the effect increased. At higher intensity levels, distortion strongly correlates with high-energy emotional states such as excitement or tension. This effect's ability to amplify emotional energy makes it an effective tool for evoking strong emotional responses in music.

2) Pitch Shift: Pitch shift demonstrated a consistent decrease in both valence and arousal across intensity levels. This effect shifts the emotional tone of music towards calmer, more introspective states. Its impact is subtle yet persistent, making

it suitable for applications that require subdued emotional expression.

3) Reverb: Reverb exhibited minimal impact on valence but showed a slight increase in arousal at higher intensity levels. This suggests that reverb's primary influence lies in enhancing the spatial and ambient qualities of music, creating subtle emotional shifts without significantly altering the emotional tone.

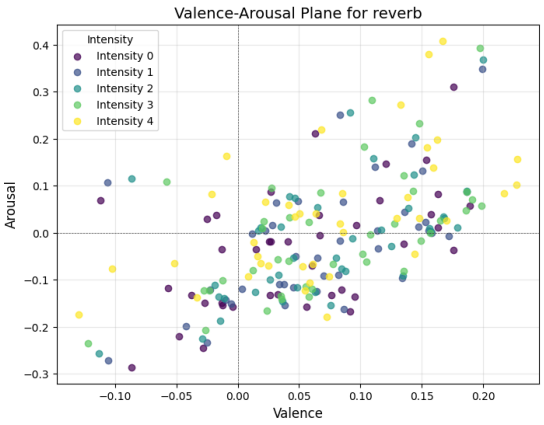

Fig. 7. Valence-Arousal Plane for Reverb

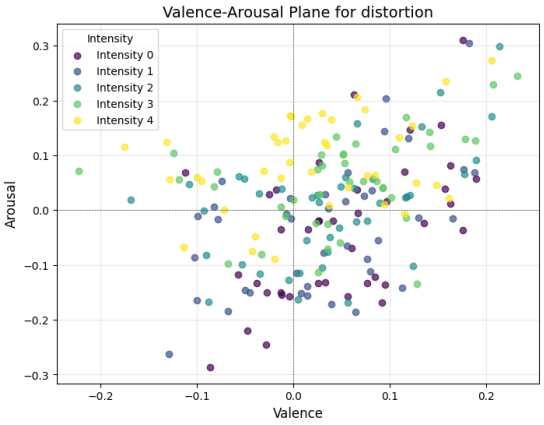

Fig. 8. Valence-Arousal Plane for Distortion

4) Delay and EQ: Both delay and EQ had negligible effects on valence and arousal. These effects maintained stable emotional dimensions across all intensity levels, suggesting their limited impact on directly influencing emotional energy.

B. Classification Analysis: Emotional Labels

The classification analysis investigated the relationship between audio effect intensity and discrete emotional labels (e.g., excitement, calmness, sadness, and anger). The findings for each audio effect are as follows:

1) Distortion: The Chi-square test revealed a statistically significant relationship between distortion intensity and emotional labels (p-*value* = 0.0000). Higher intensity levels of distortion strongly correlate with the "excitement" and "anger"

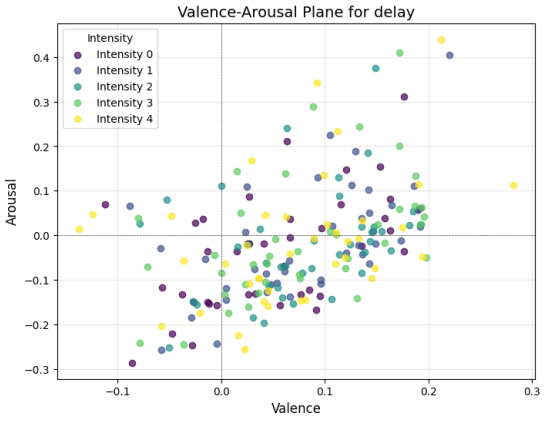

Fig. 9. Valence-Arousal Plane for Delay

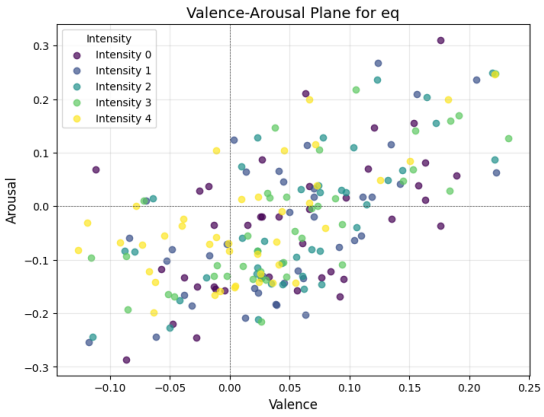

Fig. 10. Valence-Arousal Plane for EQ

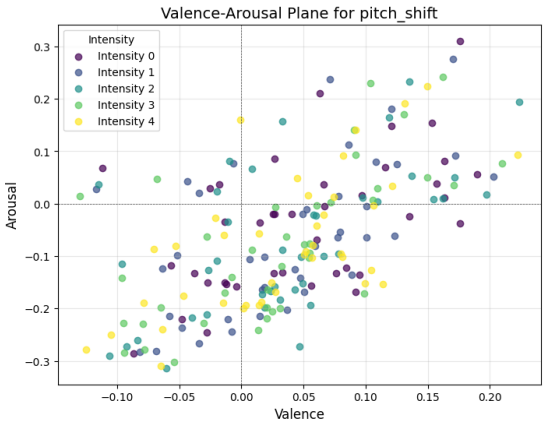

Fig. 11. Valence-Arousal Plane for Pitch Shift

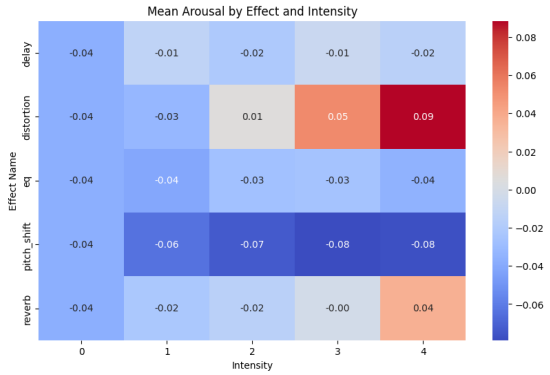

Fig. 12. Heatmap of Mean Arousal by Effect and Intensity

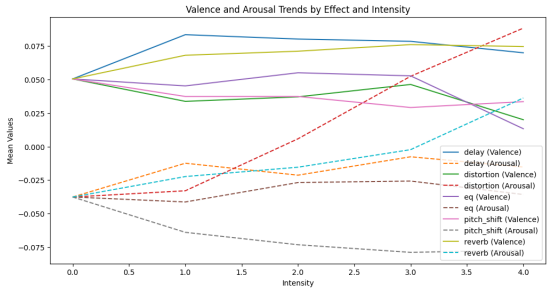

Fig. 13. Valence and Arousal Trends by Effect and Intensity

labels, reinforcing its role in generating high-energy emotional responses.

2) EQ: EQ also exhibited a statistically significant relationship with emotional labels (*p-value* = 0.0000). Changes in EQ intensity influenced emotional classifications, likely by emphasizing specific tonal qualities that align with certain emotions.

3) Reverb: Reverb's influence on emotional labels was not statistically significant (*p-value* = 0.9876). This indicates that reverb's effects are more subtle and context-dependent, without strong correlations to specific emotional classifications.

4) Delay: Delay did not demonstrate a significant relationship with emotional labels (*p-value* = 0.6535). This suggests that delay's primary role lies in spatial and rhythmic effects rather than emotional manipulation.

5) Pitch Shift: Pitch shift's impact on emotional labels was also not statistically significant (*p-value* = 0.7915). However, qualitative trends indicated an increase in the proportion of "calmness" at higher intensity levels, aligning with the findings from the regression analysis.

C. Correlation Analysis

The correlation analysis, as shown in Table IV, revealed the following:

- Intensity levels correlate more strongly with arousal than valence, indicating that intensity is a key factor in manipulating emotional energy.

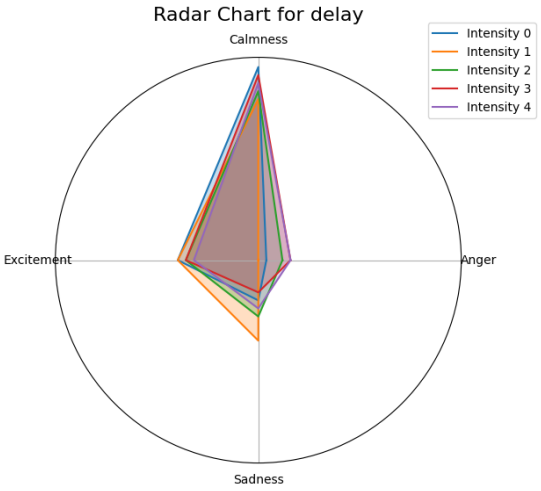

Fig. 14. Radar Chart for Delay

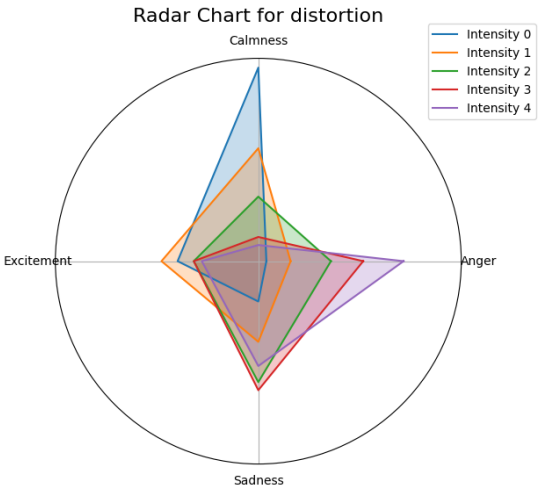

Fig. 15. Radar Chart for Distortion

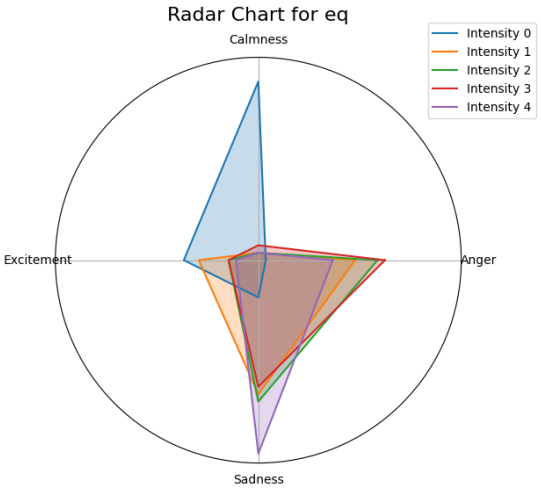

Fig. 16. Radar Chart for EQ

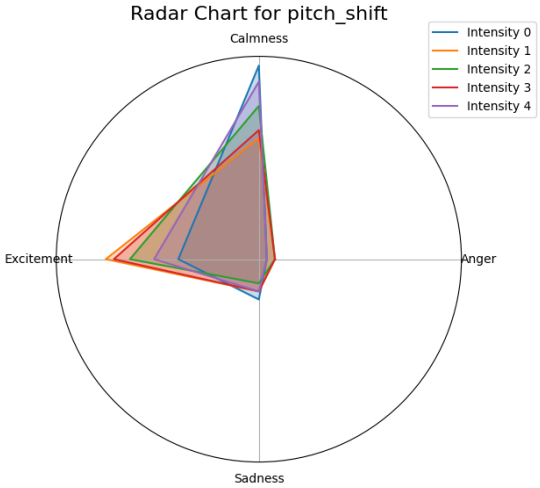

Fig. 17. Radar Chart for Pitch Shift

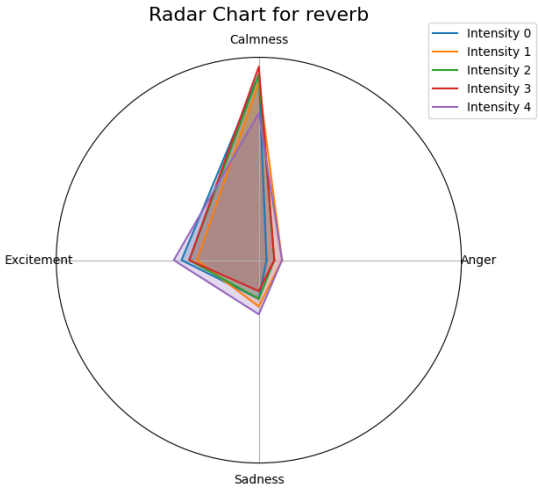

Fig. 18. Radar Chart for Reverb

- A moderate positive correlation (r = 0.51) exists between valence and arousal, suggesting that higher arousal often aligns with more positive emotional states.

These results emphasize that certain audio effects, such as distortion and EQ, are more effective in evoking noticeable emotional changes compared to others.

TABLE IV
CHI-SQUARE TEST RESULTS FOR AUDIO EFFECTS AND EMOTIONAL LABELS

Audio Effect	Chi-square Value	p-value
Reverb	3.75	0.9876
Distortion	61.36	0.0000
Delay	9.57	0.6535
EQ	105.88	0.0000
Pitch Shift	7.92	0.7915

D. Summary of Findings

The results highlight the distinct emotional effects of audio manipulation:

- Distortion and EQ are the most impactful effects, significantly influencing both continuous emotional dimensions and discrete classifications.
- Pitch shift, reverb, and delay have more nuanced or context-dependent effects, with pitch shift trending towards calmness and reverb enhancing ambiance.
- Intensity plays a crucial role in arousal modulation, with higher levels generally increasing emotional energy.

These findings offer valuable insights for music production, emotion-based music retrieval, and personalized music experiences.

IX. CONCLUSION

This study investigated the impact of audio effects on music emotions through regression (valence and arousal) and classification (emotional labels). The results highlight that distortion and EQ significantly influence both valence-arousal and emotional labels, effectively evoking high-energy states like excitement and tension. In contrast, reverb, delay, and pitch shift demonstrate subtler, more context-dependent effects.

Several challenges were encountered during this research. The most prominent issue was the lack of sufficient computational power, which limited the ability to process larger datasets and experiment with more complex models. Additionally, managing imbalanced emotional label distributions and ensuring consistent preprocessing across datasets posed difficulties. Furthermore, the subjective nature of emotional perception introduced variability in results, complicating the generalization of findings.

Despite these challenges, the study offers valuable insights and opportunities for future research. The findings could enhance emotion-based music recommendation systems, adaptive sound design, and therapeutic audio applications. Future work could leverage greater computational resources, integrate larger datasets, and explore real-time emotion tracking to refine these insights and broaden their practical applications.

REFERENCES

[1] Y. Ma, A. Øland, A. Ragni, B. Macsen, D. Sette, C. Saitis, C. Donahue, C. Lin, C. Plachouras, E. Benetos, E. Shatri, F. Morreale, G. Zhang, G. Fazekas, G. Xia, H. Zhang, I. Manco, J. Huang, J. Guinot, ... Z. Wang, "Foundation Models for Music: A Survey," [Online]. Available: https://suno.com/

[2] A. Huq, J. P. Bello, and R. Rowe, "Automated music emotion recognition: A systematic evaluation," Journal of New Music Research, vol. 39, no. 3, pp. 227–244, 2010. [Online]. Available: https://doi.org/10.1080/09298215.2010.513733

[3] M. Barthet, G. Fazekas, and M. Sandler, "Music Emotion Recognition: From Content- to Context-Based Models," LNCS 7900, n.d.

[4] D. Han, Y. Kong, J. Han, and G. Wang, "A survey of music emotion recognition," Frontiers of Computer Science, vol. 16, no. 6, 2022. [Online]. Available: https://doi.org/10.1007/s11704-021-0569-4

[5] 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, 2023. [Online]. Available: https://doi.org/10.1109/TAFFC.2020.3032373

[6] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos, N. Gyenge, R. Dannenberg, R. Liu, W. Chen, G. Xia, Y. Shi, W. Huang, Z. Wang, Y. Guo, and J. Fu, "MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training," 2023. [Online]. Available: http://arxiv.org/abs/2306.00107

[7] X. Jiang, Y. Zhang, G. Lin, and L. Yu, "Music Emotion Recognition Based on Deep Learning: A Review," IEEE Access, vol. 12, pp. 157716–157745, 2024. [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3484470

[8] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, "EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation," 2021. [Online]. Available: http://arxiv.org/abs/2108.01374

[9] M. Won, Y.-N. Hung, and D. Le, "A Foundation Model for Music Informatics," 2023. [Online]. Available: http://arxiv.org/abs/2311.03318

[10] Z. Wang, S. Li, T. Zhang, Q. Wang, P. Yu, J. Luo, Y. Liu, M. Xi, and K. Zhang, "MuChin: A Chinese Colloquial Description Benchmark for Evaluating Language Models in the Field of Music," 2024. [Online]. Available: http://arxiv.org/abs/2402.09871

[11] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE-The Munich Versatile and Fast Open-Source Audio Feature Extractor," n.d. [Online]. Available: http://www.speech.kth.se/snack/

[12] B. Mcfee, C. Raffel, D. Liang, D. P. W. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," in Proc. of the 14th Python in Science Conf. (SciPy), 2015. [Online]. Available: https://github.com/bmcfee/librosa

[13] O. Lartillot and P. Toiviainen, "MIR IN MATLAB (II): A TOOLBOX FOR MUSICAL FEATURE EXTRACTION FROM AUDIO," n.d.