Analysis of the Metrics and Evaluations Methods for Music

Ziteng Li
Kang Chiao International School Xi'an Qujiang Campus, Xi'an, China

Keywords: Music Evaluation, Deep Learning, Rhythm, Emotion Recognition, Hidden Markov Model (HMM).

Abstract: As a matter of fact, in recent decades, computer music has grown into a dominant force, revolutionizing both music creation and evaluation methods. This study explores the evolution of music evaluation from traditional, subjective approaches to more systematic, quantitative methods enabled by computational advancements. The research focuses on key evaluation metrics, including emotion, rhythm, and similarity, and how models like N-gram and Hidden Markov Models (HMM) capture melodic and rhythmic features. This research highlights recent progress in using deep learning algorithms for music assessment and their application in tasks like emotion recognition and music recommendation. Despite the successes, existing models often struggle with complex emotional expressions and cross-cultural diversity in music. The findings suggest that future improvements in music evaluation can be achieved through integrating advanced machine learning techniques and multi-modal analysis. These results contribute to the development of more objective and comprehensive evaluation methods, ultimately benefiting various applications in music classification, recommendation, and automated composition.

1 INTRODUCTION

Computer music has leaped into the mainstream of today's society after only a few decades of development. Not only has it brought new approaches to music creation, giving composers inspiration and the possibility to pursue the extreme, but it has also dramatically changed the way music is evaluated. The history of computer music can be traced back to the mid-20th century when advances in technology enabled the generation of sounds and music through algorithms. For example, The Silver Scale, composed by Max Mathews at Bell Labs in 1957, is considered the first piece of music generated through a computer. As computer performance has increased, computer music has expanded from simple audio generation to complex automated composition and real-time music processing. Nowadays, along with the disruptive development of AI, computers are not only used for music generation but also widely used for music classification, recommendation, and evaluation (Cope, 1989; Cone, 1981; Salamon et al, 1970).

With the development of technology, music assessment has also experienced a shift from the traditional subjective, authoritative individual-led assessment to a more scientific and systematic assessment method (Fink, 2014). Traditional music assessment methods rely on the subjective evaluation of listeners or experts, although this method has a certain degree of authority, the results are often characterized by strong personal bias. The masses tend to blur their true feelings because of the herd mentality and the judgment of the authorities. Therefore, with the popularization of computer music, more and more researchers are committed to developing assessment methods that can objectively quantify musical characteristics and reduce the interference of human factors.

In recent years, academic research on music assessment has gradually focused on how to quantize music features and how to apply them to different music analysis tasks. For example, some scholars proposed a data-driven deep learning tone level contour feature algorithm based on data and a hand-designed melodic feature extraction algorithm based on a priori knowledge (Yang & Chen, 2018), which provides theoretical algorithmic support for the automation of music recognition assessment. Some other scholars proposed an emotion recognition system based on music features (pitch, rhythm, timbre, etc.) from the analysis level. These researches are of extraordinary significance in application scenarios such as emotion recognition for music evaluation, music recommendation, and automated composition.

Li and Z.
Analysis of the Metrics and Evaluations Methods for Music.
DOI: 10.5220/0013512500004619
In *Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024)*, pages 198-202
ISBN: 978-989-758-754-2

Meanwhile, with the rise of machine learning and artificial intelligence technologies, deep learning-based models are widely used for music assessment. These models can automatically learn and extract high-dimensional features in music, providing powerful tools for music similarity analysis, sentiment classification, and style identification. For example, Convolutional Neural Networks (CNNs) have achieved remarkable results in music sentiment classification, audio signal processing, and music style classification.

Existing assessment models still have limitations in dealing with complex musical features, especially in terms of emotional expression, structural complexity, and cross-cultural diversity. The research motivation of this paper is to explore the limitations of existing music assessment methods and suggest directions for improvement. The research framework consists of the following parts: firstly, it introduces the main points of consideration in music assessment, such as emotion, tempo, and similarity; then it discusses the quantitative metrics used in recent years, which are quantitatively analysed by extracting features such as pitch, harmony, etc.; then it introduces the typical models used for music assessment and their applications in emotion recognition, recommender systems, and categorization; and finally, it analyses the limitations of the current methods and looks forward to possible future improvements, such as the introduction of machine learning and multimodal analysis.

2 DESCRIPTIONS OF MUSIC EVALUATION

Music evaluation involves analysing various musical elements to assess and categorize music, focusing on aspects such as emotion, rhythm, and similarity. These key considerations are vital in understanding how music affects listeners and how it can be quantitatively measured for various applications, including recommendation systems, automated composition, and emotional recognition.

One of the primary factors in music evaluation is emotion. Music has the power to evoke a wide range of emotions, from joy to sadness, and researchers have long focused on developing methods to quantify these emotional responses. Studies have shown that specific musical features such as tempo, key, and mode significantly influence emotional expression. Major keys and fast tempos are often associated with positive emotions, while minor keys and slower tempos may evoke sadness or melancholy. However, emotion recognition is not without its challenges. Human emotions are complex and multifaceted, and a single piece of music may evoke different emotions in different listeners depending on their personal experiences or cultural background. Moreover, the same musical features may be interpreted differently across genres. For instance, a minor key in classical music is often associated with sadness, while in jazz or blues, it may convey a sense of sophistication or reflection. Emotion recognition models need to account for such cross-cultural and genre-specific differences to make more accurate predictions. In particular, deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been employed to better capture the nuances of emotional expression in music (Lin & Qi, 2018).

Another crucial element in music evaluation is rhythm. Rhythm refers to the timing and arrangement of sounds and silences in a piece of music. It plays a critical role in defining the structure and flow of music, influencing how it is perceived by listeners. Researchers have explored various metrics to evaluate rhythm, such as beat alignment, tempo consistency, and syncopation. In addition to its role in music perception, rhythm is also a key indicator of technical skill. In genres like jazz or classical music, the ability to maintain complex polyrhythms or perform intricate syncopations is often associated with mastery. In contrast, genres like electronic dance music (EDM) emphasize steady, consistent rhythms, where tempo stability is paramount. Rhythm-based evaluation tools help in understanding both the aesthetic and technical aspects of rhythm across genres. In the realm of music evaluation, similarity refers to the degree of resemblance between different musical pieces. It plays a crucial role in various applications, such as music recommendation systems, automatic composition, and genre classification. Music similarity is often analysed based on features like melody, harmony, rhythm, timbre, and structure. This section focuses on the different methods used to quantify musical similarity and their applications.

The melodic similarity is one of the most fundamental aspects of music comparison. It involves analysing the sequence of pitches in a melody to determine how closely two musical pieces align. Traditional methods for measuring melodic similarity rely on calculating the Euclidean distance between pitch sequences. For instance, two melodies with similar pitch contours would exhibit a shorter Euclidean distance between their note sequences, indicating higher similarity. However, this method does not account for temporal variations or rhythmic complexities, which can significantly affect the perception of melodic similarity.

Advanced models such as N-gram and Hidden Markov Models (HMM) offer more sophisticated ways of capturing melodic similarity by considering not only pitch sequences but also the probability of transitions between pitches. These models are particularly useful in tasks like composer identification or genre classification. By analysing patterns in melodic transitions, these models can identify stylistic tendencies unique to certain composers or genres. For example, N-gram models can capture recurring melodic motifs that characterize a composer's style, while HMM can track how pitch changes unfold over time, offering a deeper analysis of melodic structure.

Harmonic similarity evaluates how closely the chord progressions or harmonic structures of two pieces resemble each other. Since harmony plays a vital role in defining the tonal character of music, harmonic similarity analysis can reveal relationships between pieces that might not be apparent through melodic or rhythmic analysis alone. Techniques for measuring harmonic similarity often involve analyzing the intervals between chords and the progression of these intervals over time. Chord-based models such as the Tonnetz (tonal network) have been employed to map harmonic relationships geometrically, allowing for the comparison of chord sequences based on their proximity in tonal space. For example, the Tonnetz model enables the identification of closely related chord progressions, such as those found in pieces from the same musical genre or period (Mor et al, 2021). Some application examples are shown in Fig. 1.

Harmonic similarity is particularly relevant in tasks such as genre classification, where certain harmonic progressions are characteristic of specific styles. For instance, classical music tends to employ complex, modulating harmonic structures, while pop music may rely on simpler, repetitive chord progressions. Analyzing these differences allows for a more nuanced understanding of genre distinctions.

Rhythm is another important factor in determining music similarity, as it defines the temporal structure of a piece. Rhythmic similarity can be evaluated by comparing the timing and duration of notes, the placement of accents, and the overall flow of a piece (Yang et al, 2019). One common approach to evaluating rhythmic similarity is through the use of beat-synchronous features, which analyse rhythm at regular time intervals. This allows for directly

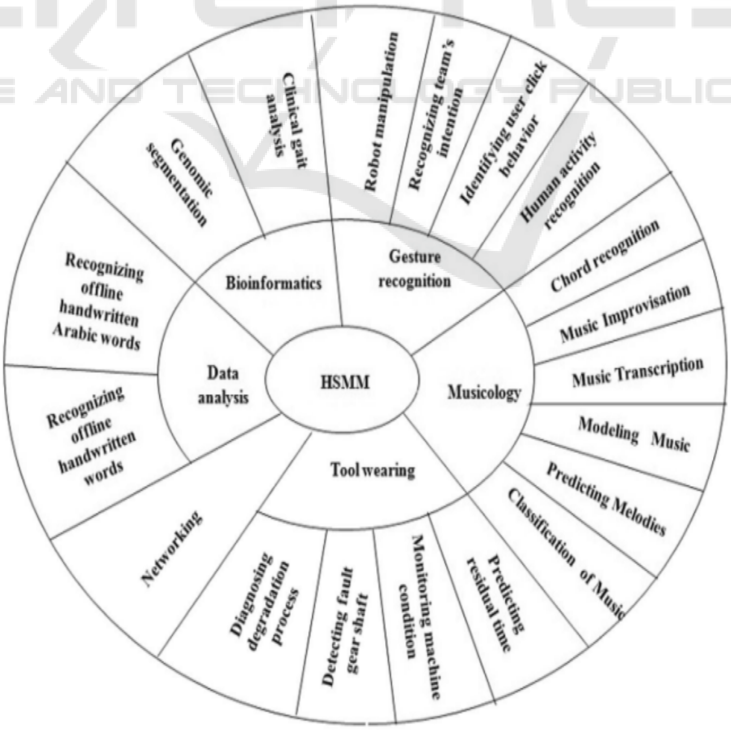

Figure 1: Applications of Hidden Semi-Markov model in music and other fields (Mor et al, 2021).

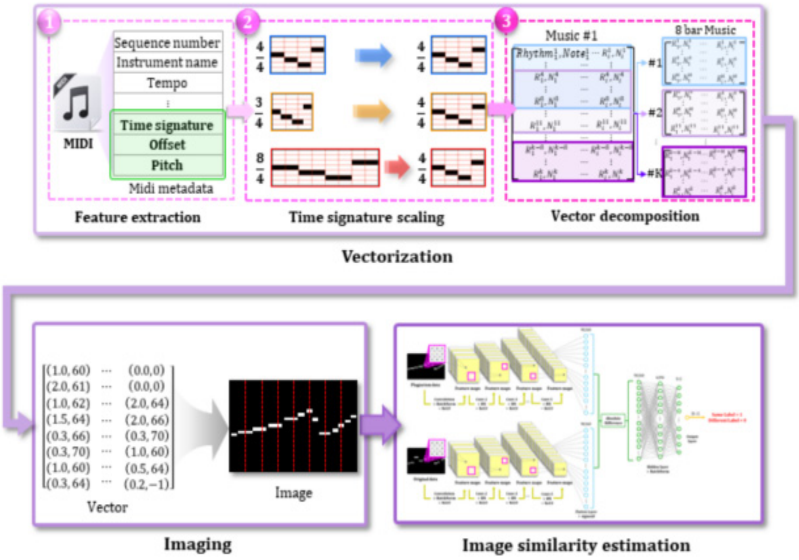

Figure 2: Scheme of music rhythm similarity detection based on Siamese CNN (ParK et al, 2022).

comparing rhythmic patterns across different musical pieces, even if their tempos vary. For example, two pieces with different tempos but similar rhythmic patterns may still exhibit high rhythmic similarity. Syncopation, the displacement of expected rhythmic accents, is another key element that influences rhythmic similarity. A typical sketch based on CNN is shown in Fig. 2 (ParK et al, 2022). Genres like jazz and funk are known for their heavy use of syncopation, and recognizing these patterns is essential for accurately assessing the similarity of pieces within these genres. Together, these elements—emotion, rhythm, and similarity—form the foundation of modern music evaluation.

3 MODELS AND APPLICATIONS

Previous researchers have used different mathematical models (e.g., N-gram model, HMM) to obtain temporal information of melodic features. It has been shown to be feasible to study state transfer methods of musical features for melodic classification or similarity computation through these time-series mathematical models. For example, Ruben Hillewaert et al. performed composer classification experiments by obtaining melodic timing features of string quartets composed by Haydn and Mozart through N-gram models, and showed that melody-specific N-gram models outperform global feature models in composer identification tasks.HMM models are also commonly used for melodic similarity computation and melodic categorization tasks, which are different from N-grams and others that produce only one state chain. gram and other temporal models that produce only a chain of states, HMM is able to study the dependency between two features (represented by hidden and observed sequences, respectively) while studying the temporal transfer of a feature, and thus achieves good results in the melody classification task. For example, some scholars used HMM to classify classical and popular music melodies by five different composers using the relative pitch and duration of the preceding and following melodic notes as features, and showed that note timing memories are reliable melodic categorization methods when there is no timbral texture or harmony involved, and that their categorization results can be comparable to the level of a music expert.Wei and Vercoe used HMM to categorize classification of folk songs composed in different countries and represented in different ways.The results show that using HMM to classify different folk songs based on length and pitch features has a significant effect on the classification accuracy ranging from 54% to 77% as compared to random guessing.Viterbi's algorithm, a widely used dynamic programming method, is proposed for the shortest path problem of directed graphs. Problems described using implicit Markov models can be decoded using the Viterbi algorithm, and the main application scenarios are digital communication, speech recognition, machine translation, pinyin to Chinese characters, and word splitting. The basic idea of the algorithm is that the path from the starting point to the end point must be the shortest path (Suriya & Kiran, 2022).

The basic idea of the algorithm is that the path from the starting point to the end point must be the shortest path, and if this path is not the shortest path, another shorter path will be chosen to replace the path between the starting point and the end point. If this path is not the shortest path, another shorter path will be chosen to replace the path between the starting point and the end point.

4 LIMITATIONS AND PROSPECTS

Quantitative evaluation and categorization of melodic similarity have been performed on many topics including the above studies with good results. However, most of the current studies consider melodic pitch and rhythm as relatively independent features for calculating melodic similarity and categorization. However, in recent years, some researchers have proved that melodic pitch and rhythm are interrelated unities from the perspective of music theory, and that pitch and rhythm have different synchronization relationships in different melodic styles. For example, it is proved through music theory analysis that the melodic pitch and rhythm of Chinese music and Western music have different synchronous development relationships. Therefore, it can be inferred that there may be significant differences in the quantitative computation results of the way different music styles depend on melody and rhythm (Hu, 2020). However, there is a lack of research on the differences in melodic pitch and rhythmic dependency patterns in different music styles. Verification of the differences in melodic pitch and rhythmic dependency modes in different styles of music requires a mathematical model that can reflect the dependency relationship between the two features. Future research may benefit from incorporating advanced machine learning techniques and multi-modal approaches to enhance accuracy and depth in music evaluation.

5 CONCLUSIONS

To sum up, this research investigated the metrics and evaluation methodologies used in music assessment, with a focus on the quantification of melody, rhythm, and harmony. This study investigated the development of computer-based music evaluation systems, emphasizing the efficacy of models such as N-gram and HMM in capturing melodic and rhythmic patterns. The research also examined recent deep learning algorithms to music evaluation, indicating their expanding importance in automated music analysis. Despite improvements, existing models still struggle to reflect the complexities of musical emotions and cross-cultural variety. To improve the accuracy and depth of music appraisal, future studies will most likely include more advanced machine learning approaches and multimodal analysis. This study contributes to the ongoing efforts to develop more objective and comprehensive techniques for evaluating musical works.

REFERENCES

Cope, D., 1989. *Experiments in musical intelligence (EMI): Non ‑ linear linguistic ‑ based composition.* Interface, 18(1‑2), 117‑139.

Cone, E. T., 1981. *The Authority of Music Criticism*. Journal of the American Musicological Society, 34(1), 1–18.

Fink, G. A., 2014. *n-Gram Models.* Markov Models for Pattern Recognition: From Theory to Applications, 107-127

Hu, X., 2020. *Recommendation algorithm based on sentiment analysis.* Southwest University of Finance and Economics

Lin, Q., Qi, Z., 2018. R*esearch on speech emotion recognition based on mixed HMM and ANN models*. Computer Technology and Development,10, 74-78.

Mor, B., Garhwal, S., Kumar, A., 2021. *A Systematic Review of Hidden Markov Models and Their Applications.* Arch Computat Methods Eng 28, 1429–1448.

Park, K., Baek, S., Jeon, J., Jeong, Y. S., 2022. *Music Plagiarism Detection Based on Siamese CNN.* Hum.-Cent. Comput. Inf. Sci, 12, 12-38.

Salamon, J., Rocha, B., Gómez, E. 1970. *Musical genre classification using melody features extracted from Polyphonic Music Signals.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 18.

Suriya, P. J., Kiran, S. 2022. *Obtain Better Accuracy Using Music Genre Classification Systemon GTZAN Dataset*. 2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon), 1-5.

Yang, M., Chen, N., 2018. *Cover song recognition model based on deep learning and manual design feature fusion.* Journal of East China University of Science and Technology, 5, 752-759.

Yang, Y., Jo, J., Lim, H., 2019. *Unifying user preference and item knowledge-based similarity models for top-N recommendation.* Personal and Ubiquitous Computing, 23(6), 901-912.