Dean&Francis
ISSN 2959-6157

Evolution, Evaluation, and Challenges of Automatic Music Style Clustering Techniques: A Review from Handcrafted Features to Self-Supervised Learning

Hongyu Hu[1, *]

[1]School of Computer Science, Zhejiang University of Science and Technology, Hangzhou, 310023, China
*Corresponding author: 1230204047@zust.edu.cn

Abstract:

Automatic discovery and clustering of music styles are fundamental to music information retrieval, recommendation systems, and computational musicology. Recent advances in deep learning and self-supervised representation learning have substantially improved audio embeddings, enabling more effective unsupervised clustering based on musical style and acoustic characteristics. This paper reviews representative approaches to music style clustering, including traditional feature-based methods, two-stage frameworks combining deep embeddings with clustering, and emerging self-supervised online clustering models that jointly optimize representation learning and clustering structure. Furthermore, this paper summarizes commonly used datasets and evaluation metrics in this field, reviews experimental findings from representative studies, and analyzes the advantages and limitations of various methods. Finally, this paper points out the current open challenges, including inconsistent evaluation standards, difficulties in cross-cultural and multi-label style recognition, insufficient interpretability of clustering results, and scalability issues in large-scale streaming scenarios. This paper aims to provide researchers with a clear and structured technical roadmap to facilitate the development and evaluation of future music style clustering systems.

Keywords: Music style clustering; Self-supervised learning; Unsupervised representation; Music information retrieval; Custer evaluation.

ISSN 2959-6157

1. Introduction

Musical style and genre are core dimensions for organizing, understanding, and retrieving music content. With the explosive growth of digital music libraries and the widespread adoption of streaming services, efficiently managing massive amounts of music based on style has become a critical task in the field of music information retrieval. Musical style clustering aims to automatically discover potential stylistic structures from music audio signals under unsupervised or weakly supervised conditions, grouping musical works of similar styles into one category.

Traditional music style identification often employs a supervised learning paradigm, which involves training classification models on datasets with style labels. This method can achieve high accuracy in closed environments with ample labels and well-defined styles. However, music style is inherently a subjective and dynamically evolving concept. The same piece of music may belong to multiple styles simultaneously, and style definitions vary depending on region and individual listening experience. This cultural dependence and subjectivity lead to high costs for manual annotation and difficulties in ensuring annotation consistency. Therefore, supervised learning methods that rely on large amounts of accurately labeled data face generalization bottlenecks in real-world large-scale, open-domain scenarios.

Against this backdrop, unsupervised or weakly supervised music style clustering methods have demonstrated their unique value. These methods do not rely on complete human labels but instead attempt to directly mine the inherent structure and patterns from the data. In recent years, two major technological waves have greatly propelled the development of this field: first, the rise of deep representation learning, particularly the successful application of self-supervised learning in the audio domain, enabling models to learn rich, high-level, and transferable audio representations from massive amounts of unlabeled audio data; and second, the shift in learning paradigms, from a separate pipeline of „representation first, clustering later“ to a collaborative paradigm that integrates representation learning and clustering objectives, such as online clustering and cluster-based self-distillation, aiming to learn a feature space more favorable to the clustering task itself.

Although existing studies have shown promising results on specific datasets or tasks, music style clustering as a research field has not yet been systematically sorted out and summarized in terms of its technical path, evaluation criteria and challenges. Existing reviews mostly focus on supervised music genre classification, while paying relatively little attention to the increasingly important direction of unsupervised clustering. For example, Green et al.‘s critical review pointed out the theoretical defects, label noise and evaluation inconsistencies in music genre identification research, and emphasized the urgency of developing more robust and universal unsupervised methods [1].

Current research has yielded a variety of technical approaches. For example, Spijkervet et al. used a contrastive learning framework to learn musical embeddings that can distinguish between melody and timbre features without explicit labels, effectively improving the quality of subsequent clustering tasks [2]. The deep spectrogram convolutional network designed by Li et al. enhances the separability of learned features in the style dimension by simultaneously modeling the local details and global structure of the spectrum [3]. More notably, inspired by speech self-supervised models, the research of Kang et al. and Chen et al. used pseudo-labels generated by the clustering process itself as supervision signals to iteratively optimize audio representations, proving that clustering and representation learning can form a virtuous cycle and jointly promote the understanding of the high-level structure of music [4,5].

However, the path to a practical and robust music style clustering system remains fraught with challenges. The main challenges lie in how to simultaneously capture low-level acoustic features and high-level stylistic semantics of music, and develop a universal audio representation robust to changes in recording conditions; how to design clustering models capable of handling style overlap, multi-label assignment, and hierarchical style structures; how to objectively, reliably, and perceptually align clustering results with human perception in the absence of „absolute truth“ labels; and finally, how to ensure the interpretability of clustering results and their efficient service to large-scale, dynamically updated music streaming platforms.

To address these challenges, this paper attempts a comprehensive review of the current state of research in music style clustering. It systematically introduces mainstream techniques, from classic methods based on handcrafted features to cutting-edge integrated self-supervised learning frameworks; provides detailed reviews of commonly used experimental datasets and evaluation metrics; comprehensively analyzes the performance and limitations of different methods in various scenarios; and delves into the core challenges currently faced and potential future research directions. This paper aims to provide researchers and practitioners in related fields with a clear reference framework, jointly promoting the further development of music style clustering technology.

2. Overview of mainstream technologies

2.1 Traditional features and clustering methods

Before the rise of deep learning, the field of music information retrieval relied heavily on handcrafted feature engineering that combined signal processing and musicological knowledge. These features aimed to quantify certain perceptual or acoustic properties of music, most notably the Mel frequency cepstral coefficients, which simulate human hearing and possess a powerful descriptive ability for timbre. Other features, such as beat intensity, rhythm histogram, and BPM estimates, were used to characterize the temporal structure of music. Chromaticity features and tonality features reflected the harmony and tonality of music. Spectral statistical features, such as spectral centroid, bandwidth, roll-off point, and spectral flux, described the macroscopic shape and dynamic changes of the spectrum. After extracting these features, researchers combined them into feature vectors and input them into classic clustering algorithms, such as K-means, hierarchical clustering, Gaussian mixture models, or spectral clustering. This paradigm dominated research in the early 21st century. Its greatest advantages were transparency and low computational cost. Features had clear physical or auditory meanings, and the clustering process was easy to understand and debug. It remained a viable starting point for small datasets or resource-constrained environments. However, its limitations were equally prominent. First, the feature representation capability is limited. Handcrafted features are usually statistical summaries of low-level acoustic properties, which are difficult to capture high-level, abstract semantics related to music style, such as "sadness", "the guitar wall of glam rock", "the synthesizer texture of electronic music", etc. Second, they are sensitive to data quality. Differences in recording conditions, compression formats, and background noise can significantly affect feature values, leading to a decrease in clustering performance. Finally, they have poor scalability. When the amount of data increases, the styles become more diverse and finer-grained, the discriminative power of handcrafted feature combinations often reaches its limit. Although benchmarks based on early datasets such as GTZAN are still cited [6], the problems of these datasets themselves make the universality of the results questionable. Research at this stage has demonstrated the feasibility of style clustering more, but has not yet shown the ability to solve complex real-world problems.

2.2 Deep embedding and clustering

The second common paradigm is "representation first, clustering later". It uses CNN/CRNN or more common pre-trained models to extract fixed vector representations of audio, and then uses algorithms such as k-means, HDBSCAN, GMM, and spectral clustering to perform clustering. The advantage of this paradigm is that deep embedding can capture richer musical semantics and temporal information, and is usually better than handmade features in downstream classification and retrieval tasks. The disadvantage is that the embedding quality is highly dependent on the pre-training task and data distribution, and there is domain bias; in addition, large-dimensional embedding requires additional design in terms of storage and clustering computation. Existing research shows that using pre-trained audio embeddings as "teachers" to guide lightweight models can reduce the computational burden while maintaining performance, which also provides a practical route for clustering scenarios [7].

2.3 Integrated/Online Clustering and Self-Supervised Learning

More advanced approaches attempt to jointly optimize clustering structure and representation learning during training. Self-supervised tasks are constructed by online clustering, vector quantization, or generating pseudo-labels from clustering results, thereby guiding representation learning to be guided by the clustering structure. The advantage of this paradigm is that it can learn clustering-friendly representations and improve the final clustering quality; the disadvantage is that training is more complex, hyperparameters are sensitive, and stability and convergence are still challenges on large-scale data. For example, the success of HuberT in the speech domain provides a theoretical and practical basis for the idea of mutually promoting "clustering ↔ representation learning"; similar strategies have also begun to be adopted in the music field [8].

2.4 Multimodal and User Behavior-Driven Clustering

To better reflect the perception of listeners, researchers have gradually combined "audio + lyrics + metadata + user behavior" for clustering or as a signal for clustering interpretation. Audio-text alignment models can project natural language descriptions and audio into the same embedding space, thus supporting clustering driven by text interpretation or direct text prompts. User behavior data can reflect the "perceptual proximity" of real listeners, which is very helpful for building clusters that are meaningful to users, but this also introduces privacy

ISSN 2959-6157

and cross-platform differences. Multimodal models such as MuLan and CLAP have shown good performance in multi-task applications, providing usable underlying representations for multimodal clustering [9].

3. Experimental Section

3.1 Dataset

The selection of datasets is crucial for evaluating clustering methods. Ideal datasets should be large in scale, diverse in style, relatively accurate in labeling, high in audio quality, and contain meta-information. However, real-world datasets often make trade-offs in various aspects. GTZAN Genre Collection is one of the oldest and most famous datasets in the field of music genre recognition. It contains 10 genres, 100 audio clips of 30 seconds each, with a total duration of about 8 hours. Its advantages are simplicity and ease of use, making it a benchmark for many early studies. Its disadvantages are also very prominent, including known duplicate files, label errors, inconsistent recording quality, and a small scale, making it difficult to support the training and robust evaluation of modern deep models [10]. MagnaTagATune contains about 25,000 30-second music clips from the MagnaTagATune music library. Its feature is that it provides "tags" collected by the Amazon Mechanical Turk crowdsourcing platform. These tags are not single genres, but descriptive keywords and are multi-labeled. It is suitable for evaluating clustering or prediction of fuzzy attributes and multi-label attributes, and is closer to the actual situation of music description [11]. MusDB18 is a dataset focused on music source separation, providing stereo mixes and separated dry tracks for 150 complete songs. Although primarily used for separation, its multi-track nature allows it to be used to study the impact of timbre and instrument composition on style clustering. For example, clustering can be performed based solely on drum tracks to study rhythmic styles. Researchers should choose carefully based on their task objectives [12]. GTZAN is still a valuable reference for method prototype validation, but conclusions should be extrapolated cautiously. For research on multi-label or user perception, MagnaTagATune and data containing user behavior are better choices. At the same time, when reporting results, the subset of the dataset used and the preprocessing method should be clearly stated to improve the comparability and reproducibility of the results.

3.2 Evaluation indicators

Evaluating unsupervised clustering results is a core challenge in this field because there is no absolute "correct answer." It typically requires a comprehensive assessment combining internal metrics, external metrics, downstream tasks, and human evaluation. Internal metrics evaluate only the clustering results and the data itself, without requiring real labels. They measure the compactness within clusters and the separation between clusters. External metrics, when the data has credible style labels, can be considered "benchmark truths" to evaluate the consistency between the clustering results and the labels. Downstream task metrics use the clustering results or learned embeddings as features for other supervised tasks to indirectly evaluate their quality. Human evaluation is the most direct but also the costlist method. Auditory tests are often designed where participants, unknowingly, listen to multiple songs within the same cluster and evaluate their stylistic consistency; or listen to songs from different clusters and evaluate their stylistic differences. Likert scales or A/B testing can be used. Human evaluation is the ultimate standard for verifying the "musical meaning" of clustering results, but the participants' background, experimental design, and other factors can influence the results, making standardization difficult.

3.3 Results Analysis

Based on representative studies in recent years, the paper can observe some general trends and important insights. On almost all benchmark datasets, the performance of embeddings extracted by deep pre-trained models and then clustering algorithms is significantly and consistently better than baseline methods based on handcrafted features such as MFCC. This demonstrates the powerful ability of deep networks to automatically learn high-level, style-related musical representations.

For example, using CNN embeddings pre-trained with AudioSet, the clustering NMI value on GTZAN can usually be 10-20 percentage points higher than the MFCC baseline [13].

In terms of clustering algorithm selection, for deep embeddings, due to their high dimensionality and possible non-convex distribution, spectral clustering and HDBSCAN often perform better than simple K-means. The advantage of HDBSCAN is that it does not require pre-specifying the number of clusters and can identify noise points, which is very useful for data in the real world where style boundaries are unclear and there are "alternative" songs. However, HDBSCAN may face the "curse of dimensionality" in high-dimensional space, so sometimes it is necessary to use UMAP or PCA for dimensionality reduction first [14].

The integrated paradigm, represented by online clustering

and pseudo-labeling methods, has reported impressive results in some studies, sometimes even surpassing the two-stage method of "pre-trained model + classic clustering". For example, some variants based on HuBERT iterative clustering have achieved higher NMI on music style discovery tasks [4,5]. However, the performance of such methods is highly volatile and very sensitive to hyperparameter settings, model initialization, and training strategies. Their reproducibility and stability are the main obstacles to practical application. But they point to an important direction for the future: learning representations in a customized way for clustering tasks.

At the same time, the evaluation results on different datasets may vary greatly. A model that performs well on FMA may perform poorly on MagnaTagATune, and vice versa [11]. This highlights the huge impact of data distribution on model performance and emphasizes the importance of evaluating on multiple heterogeneous datasets to test the model's generalization ability.

4. Challenges and Prospects

4.1 Core challenges

Musical pieces often exhibit characteristics of multiple styles simultaneously, leading to inherently fuzzy boundaries between genres. Traditional hard-assignment clustering oversimplifies this reality by forcing each song into a single category. Future research should focus on soft clustering, overlapping clustering, or probabilistic membership assignment, enabling models to better reflect the hybrid and fluid nature of musical styles.

Musical styles are not flat but inherently hierarchical, ranging from coarse-grained genres to fine-grained substyles. Existing flat clustering approaches fail to capture these rich hierarchical relationships. Developing hierarchical clustering methods or models capable of learning multi-granular representations is therefore an important research direction.

Musical styles continuously evolve, merge, and diverge over time. Static clustering models are unable to capture such temporal dynamics. Incremental or temporal clustering methods should be explored to adapt to long-term changes in music production, consumption patterns, and cultural trends.

There is no single objective ground truth for musical style classification. Labels from different datasets or platforms may contradict one another, making evaluations based on external indicators fundamentally uncertain. Metrics such as NMI or silhouette coefficient do not necessarily correspond to clusters perceived as meaningful by human listeners. A cluster may be acoustically compact while lacking emotional or cultural coherence. Designing objective proxy indicators that strongly correlate with human subjective perception remains a long-term challenge.

The decision-making processes of deep models and complex clustering algorithms are often opaque. Users and musicologists may find it difficult to understand why certain songs are grouped together. Improving interpretability is crucial for building user trust and supporting musicological analysis. This requires the development of techniques such as attribution analysis and concept activation vectors.

Beyond numerical evaluation, clustering results must be translatable into musically meaningful descriptions. Simply reporting that "cluster A and cluster B have an ARI of 0.8" is insufficient. Instead, clusters should be interpretable in semantic terms, such as identifying a cluster as "a variant of 1970s psychedelic rock."

Modern streaming platforms host hundreds of millions of tracks, posing significant scalability challenges for clustering algorithms. Efficient solutions must support distributed computation and incremental updates. Graph-based approximation algorithms and locality-sensitive hashing techniques should be more tightly integrated with clustering models to enable large-scale deployment.

Integrated self-supervised clustering models are computationally expensive to train. Research into more efficient network architectures, training strategies, and optimization algorithms is essential to make these methods practical for real-world applications.

Cultural, regional, and gender biases present in training data can be amplified by clustering models. For example, non-Western music may be overgeneralized or misclassified. Fairness auditing should be incorporated into data collection, model design, and evaluation processes to mitigate such risks.

When user behavior data is used for music clustering, strict compliance with data privacy regulations is essential. Privacy-preserving techniques such as federated learning and differential privacy offer promising solutions and may play an increasingly important role in this domain.

4.2 Future Outlook

Based on the above challenges, this paper anticipates the following promising research directions: First, designing pre-training tasks that are closer to the essence of music cognition. For example, designing prediction tasks by combining prior knowledge of music theory; utilizing multi-track information of music for cross-modal self-supervision; and constructing a large-scale, diverse music audio-text pair dataset to train a more powerful music-lan-

ISSN 2959-6157

guage joint model. Second, viewing style clustering as part of a broader discovery of music structure. Jointly modeling multiple dimensions such as style, emotion, instrument, and singing style to learn a multifaceted representation of music. Developing disentanglement representation learning techniques to separate style factors from other factors, thereby obtaining purer and more controllable style clusters. Furthermore, the limits of completely unsupervised learning may be difficult to overcome. Introducing a human feedback loop is crucial. Active learning strategies can be designed to allow the model to ask human experts questions about its most uncertain cluster boundaries; or interactive tools can be developed to allow music editors or users to manually adjust the clustering results and incorporate this feedback into model iteration. In addition, efforts should be made to promote the community to jointly establish a framework that includes multi-level evaluation: a legal layer, a task layer, and a human perception layer. And efforts should be made to construct a large-scale, multicultural, and multi-annotated benchmark dataset. Finally, through collaboration between industry, academia, and research, an end-to-end clustering system optimized for music streaming platforms was developed. This includes efficient embedded indexing, online clustering update algorithms, and product design that seamlessly integrates clustering results into the recommendation and browsing interfaces.

5. Conclusion

Music style clustering, as a crucial bridge connecting the underlying audio signals of music with high-level semantic understanding, has seen its research deepen alongside the evolution of artificial intelligence and music information retrieval. This paper systematically reviews the technological development in this field, from the early days of combining handcrafted features with classic algorithms relying on prior knowledge, to the golden age of the two-stage paradigm of "deep embedding + clustering" leveraging the powerful representational capabilities of deep networks, to the cutting-edge exploration of integrated self-supervised clustering pursuing collaborative learning of representation and structure, and the emerging trend of multimodal clustering that integrates multi-source information to approximate complex human cognition. Each paradigm shift aims to better address the core challenges inherent in music style itself, such as ambiguity, multidimensionality, cultural dependence, and subjectivity.

Regarding experimental evaluation, deep learning methods have established significant advantages on multiple benchmarks. However, issues such as inconsistent evaluation standards, dataset bias, and the gap between internal metrics and human perception still hinder the comparability of research and the measurement of progress. Currently, while state-of-the-art models can capture rich patterns from acoustic signals, they still face significant bottlenecks in handling style overlap, hierarchical structure, cross-cultural generalization, and providing musicological interpretability. Looking ahead, research on music style clustering is at a critical juncture, moving from "technically feasible" to "practically reliable." Future breakthroughs may depend on collaborative efforts in several areas: first, innovation in learning paradigms, including designing more musical self-supervised tasks and developing clustering models that handle soft assignments and hierarchical structures; second, refinement of evaluation systems, establishing multi-level evaluation benchmarks aligned with human perception; third, enhancement of interpretability and interactivity, opening up the model's black box and incorporating feedback from human experts; and finally, advancements in systems engineering to address efficiency, privacy, and fairness issues related to massive datasets.

In short, music style clustering is not only a challenging machine learning problem but also an interdisciplinary field rich in art and humanities. Its ultimate goal is not merely to enable machines to "calculate" style tags, but to enable machines to "understand" the organization of music, thereby better serving humanity's creation, appreciation, discovery, and inheritance of music. This article anticipates seeing more researchers from computer science, musicology, cognitive psychology, and the humanities and social sciences collaborate to write a new chapter in music artificial intelligence.This paper also has several limitations. As a review study, it mainly relies on previously published literature and reported experimental results, lacking large-scale original empirical validation. In addition, due to the rapid development of this field, some recent methods and practical systems may not be fully covered.

References

[1] Green O, Sturm B, Born G, et al. A critical survey of research in music genre recognition. International Society for Music Information Retrieval, 2024.

[2] Spijkervet J, Burgoyne J A. Contrastive learning of musical representations. arXiv preprint arXiv:2103.09410, 2021.

[3] Li J, Han L, Li X, et al. An evaluation of deep neural network models for music classification using spectrograms. Multimedia Tools and Applications, 2022, 81(4): 4621-4647.

[4] Kang W H, Alam J, Fathan A. An analytic study on clustering-based pseudo-labels for self-supervised deep speaker verification//International Conference on Speech and Computer.

Cham: Springer International Publishing, 2022: 338-348.
[5] Chen K, Wichern G, Germain F G, et al. PaC-HuBERT: Self-Supervised Music Source Separation Via Primitive Auditory Clustering And Hidden-Unit Bert//2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). IEEE, 2023: 1-5.
[6] Chen S, Wang C, Chen Z, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505-1518
[7] Ding Y, Lerch A. Audio embeddings as teachers for music classification. arXiv preprint arXiv:2306.17424, 2023.
[8] Niizumi D, Takeuchi D, Ohishi Y, et al. Byol for audio: Self-supervised learning for general-purpose audio representation//2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.
[9] Huang Q, Jansen A, Lee J, et al. Mulan: A joint embedding of music audio and natural language. arXiv preprint arXiv:2208.12415, 2022
[10] Sturm B L. The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. arXiv preprint arXiv:1306.1461, 2013.
[11] Wolff D, Weyde T. Adapting similarity on the magnatagatune database: effects of model and feature choices// Proceedings of the 21st international conference on world wide web. 2012: 931-936.
[12] Parvathi S S, Chandrasekar D. Feature separation of music across diverse dataset: a comparative perspective. Bulletin of Electrical Engineering and Informatics, 2025, 14(5): 3903-3912.
[13] Sun Y, Xu Q, Su Y, et al. AudioSet-R: A Refined AudioSet with Multi-Stage LLM Label Reannotation//Proceedings of the 33rd ACM International Conference on Multimedia. 2025: 13089-13096.
[14] Stewart G, Al-Khassaweneh M. An implementation of the HDBSCAN* clustering algorithm. Applied Sciences, 2022, 12(5): 2405.